



A University of Sussex PhD thesis

Available online via Sussex Research Online:

<http://sro.sussex.ac.uk/>

This thesis is protected by copyright which belongs to the author.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Please visit Sussex Research Online for more information and further details

THE CORPUS EXPANSION TOOLKIT

JACK FREDERICK PAY

Finding what we want on the web

Submitted for the degree of Doctor of Philosophy
University of Sussex
September 2019

SUPERVISOR:
Prof. David Weir

DECLARATION

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Jack Frederick Pay

ABSTRACT

This thesis presents the Corpus Expansion Toolkit (CET), a generally applicable toolkit that allows researchers to build domain-specific corpora from the web. The main purpose of the work presented in this thesis and the development of the CET is to provide a solution to discovering desired content on the web from possibly unknown locations or a poorly defined domain. Using an iterative process, the CET is able to solve the problem of discovering domain-specific online content and expand a corpus using only a very small number of example documents or characteristic phrases taken from the target domain. Using a human-in-the-loop strategy and a chain of discrete software components the CET also allows the concept of a domain to be iteratively defined using the very online resources used to expand the original corpus. The CET combines feature extraction, search, web crawling and machine learning methods to collect, store, filter and perform information extraction on collected documents. Using a small number of example 'seed' documents the CET is able to expand the original corpus by finding more relevant documents from the web and provide a number of tools to support their analysis. This thesis presents a case study-based methodology that introduces the various contributions and components of the CET through the discussion of five case studies covering a wide variety of domains and requirements that the CET has been applied. These case studies hope to illustrate three main use cases, listed below, where the CET is applicable:

1. Domain known – source known
2. Domain known – source unknown
3. Domain unknown – source unknown

First, use cases where the sites for document collection are known and the topic of research is clearly defined. Second, instances where the topic of research is clearly defined but where to find relevant documents on the web is unknown. Third, the most extreme use case, where the domain is poorly defined or unknown to the researcher and the location of the information is also unknown. This thesis presents a solution that allows researchers to begin with very little information on a specific topic and iteratively build a clear conception of a domain and translate that to a computational system.

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and gratitude to David Weir, Jeremy Reffin and Julie Weeds for their help, guidance and advice during my time in the T.A.G laboratory and throughout my PhD.

I would also like to thank Simon Wibberley, Andrew Robertson, Matti Lyra, Alexander Savkov and Miro Batchkarov for your time, discussions, collaborations and fun times. With particular thanks to Simon Wibberley and Andrew Robertson for their development of the Method52 platform. Throughout my years in the lab I have learnt so much from all of you and truly appreciate what you have given me. I would also like to thank Carl Miller and Josh Smith for our collaborations on the case studies detailed in this thesis.

I would also like to thank my parents Graham and Karen Pay and also my three sisters Rachel, Charlotte and Holly. Without your love, advice, support and understanding I would not have been able to achieve this piece of work. Thank you all so much, I could not have done it without you. For the same reasons, I would also like to thank all of my friends, especially Tanya Kant who has helped me so much over the course of my PhD.

CONTENTS

1	INTRODUCTION	1
1.0.1	Defining a domain	2
1.1	Why might we want to build a corpus from the web? .	2
1.1.1	Informing professionals	2
1.1.2	Analysis of social media for discourse analysis and event detection	3
1.1.3	Monitoring and capturing illicit activity online	5
1.1.4	General corpora for research	6
1.2	Existing solutions	7
1.2.1	Web search	7
1.2.2	Web crawling	8
1.2.3	Multiagent systems	9
1.2.4	Summarising the issues of existing solutions . .	10
1.3	The Philosophy of the Corpus Expansion Toolkit . . .	10
1.3.1	The human-in-the-loop philosophy and the CET	11
1.4	Bridging the qualitative and the quantitative	12
1.5	Why not just use a search engine?	13
1.5.1	The issues	14
1.6	Multiagent systems for web research	14
1.6.1	The Corpus Expansion Toolkit	15
1.7	Research aims and contributions	18
1.8	Organisation of the thesis	20
2	BACKGROUND	23
2.1	Bootstrapping the web	23
2.1.1	Web for linguistic corpora: comparable and parallel corpora	26
2.1.2	Bootstrapping domain-specific corpora	27
2.2	Web search	28
2.2.1	Domain relevance and result filtering	28
2.2.2	Search engine effectiveness and ranking	30
2.2.3	Example-based search	33
2.2.4	Caveats of search engines	34
2.3	Web crawling	35
2.3.1	Rerawl strategies	36
2.3.2	Topical web crawling	37
2.3.3	Domain specific corpora using web crawling . .	39
2.3.4	Adaptive crawling	41
2.4	Multi-agent systems	42
2.4.1	Domain discovery	48
3	METHODOLOGY	51

3.1	Methodology: Case studies-based approach	51
3.1.1	Ethical considerations and arrangements	52
3.1.2	Case study: Filtering crawled content	53
3.2	Case study: Building corpora from known sources	54
3.3	Case study: Discovering domain content from unknown sources	56
3.4	Application-based evaluation method	57
3.5	Method52	59
3.5.1	Documents and annotations	60
3.5.2	Jobs	60
3.5.3	Work area	61
3.6	Components	64
4	FILTERING CRAWLED CONTENT	69
4.1	The external voting rights of diaspora	69
4.2	Tracking diaspora voting rights online	71
4.3	Methodology	73
4.3.1	Agile method	73
4.4	The Semi-automated web search and analytics tool	74
4.4.1	Active learning based classification	76
4.4.2	User interface	80
4.4.3	Crawl	81
4.4.4	Search	83
4.4.5	Document Analysis	85
4.4.6	Database	88
4.5	Discussion	90
5	BUILDING CORPORA FROM KNOWN SOURCES	93
5.0.1	Case study: Mental health discussions in public forums	94
5.0.2	Case study: Regularly collecting structured content from the Childline message boards	97
5.0.3	ACLED: Crawling and scraping news websites for reports of conflict	98
5.1	Methodology	100
5.2	Web scraping: Automatic and Configurable	103
5.2.1	Continuous web crawling	112
5.2.2	Named Entity Recognition and OpenNLP	117
5.3	Wellcome Trust: Results	118
5.3.1	Health forum case study	118
5.3.2	Ethical review	119
5.3.3	Data collection	119
5.3.4	Analysis	121
5.3.5	Cries for help	121
5.3.6	Cognitive behavioural therapy	123
5.3.7	NER	124
5.4	Discussion	125

6	DISCOVERING DOMAIN CONTENT FROM UNKNOWN SOURCES	128
6.1	Case study: Project Parrot	129
6.2	Background	130
6.2.1	Multiagent systems for identifying illegal wildlife trade	132
6.3	Methodology	133
6.3.1	The Corpus Expansion Toolkit	134
6.3.2	Human-in-the-loop strategy	136
6.3.3	Frequency Annotator	137
6.3.4	Custom Logic Annotator	137
6.3.5	Category Annotator	137
6.3.6	Keyword Annotator	137
6.3.7	Duplicate Annotator	138
6.3.8	Phase One: Bootstrapping	138
6.3.9	Phase Two: Web Crawling, scraping and machine learning	143
6.3.10	Clustering and topic modelling	145
6.3.11	Twitter and Reddit API	146
6.4	Results: Pangolins	146
6.5	Discussion	164
7	CONCLUSIONS	166
7.1	Future Work	168
7.1.1	Classifiers	168
7.1.2	Focused crawling	169
7.1.3	Web scraping	169
7.1.4	Feature extraction and seed generation	170
7.1.5	Search	170
7.1.6	User interface	171
	Appendices	i
	BIBLIOGRAPHY	viii

INTRODUCTION

The internet has increasingly become a dominant means of communication and information sharing across the globe (Daniels et al., 2017). This has precipitated an increasing migration away from traditional forms of communication, such as telephony and libraries to newer online methods, such as forums and social media (Prensky, 2001). These factors have led to various developments and the spawning of the Information Retrieval (IR) field of research. IR has a primary focus on the research and development of computational methods for discovering, organising, understanding and analysing corpora at scale. One of the key problems that IR and this thesis attempt to address is that building corpora and discovering relevant content on the web is extremely hard because the web is a large and continually growing information space, which at the time of writing consists of more than 5.47 billion pages (Kunder, 2019). In addition to its size, the web consists of heterogeneously structured web pages linked to each other in a largely unstructured network, which makes it hard to navigate in any meaningful way.

The problem of finding the things we want on the web has traditionally been solved by search engines, such as Google, Bing and DuckDuckGo. Typically these search engines crawl the web and index all content they discover. Users then query this index via keyword searches, with results ranked and presented via tools provided by the platform. This presents a number of issues, discussed below, as users are dependant on these tools and must implicitly trust that returned results best match their search criteria.

This thesis presents a case study based methodology that illustrates the development of a generally applicable method and toolkit, referred to as the Corpus Expansion Toolkit (CET), designed to address the above problems and allow users to build domain-specific corpora from online content. In addition, the CET provides a means to discover relevant content at scale, presenting a means to build potentially very large corpora from the web.

1.0.1 *Defining a domain*

Defining what we mean by domain is key to understanding what is meant by building domain-specific corpora from the perspective of this thesis. Much of the work presented and discussed refers to means of domain definition and the discovery of documents relevant to a target domain. These terms are often used to generally define methods that seek to capture a domain-specific vocabulary and documents written using that vocabulary. For instance, there are examples both in this work and the related literature, where the notion of domain relevance is built in to a proposed solution in order to identify documents originating from a target domain (Oliveira, 2014; Vicente et al., 2018; Pham et al., 2019). These classifiers are trained on a gold-standard corpus of documents known to be relevant and it is the vocabulary of these corpora that are used to build a probabilistic representation of a domain.

This introductory chapter is organised into six sections. First, a number of examples are presented to illustrate how and why one might choose to build corpora from the web. Second, a summary and discussion of current solutions to build corpora from the web are presented. Third, the Corpus Expansion Toolkit (CET) is briefly introduced to highlight the contributions of this thesis. Fourth, the main hypothesis of this thesis is discussed to highlight how the CET addresses the issues posed by current solutions. Fifth, the research questions and contributions of this thesis are presented. This chapter concludes with a summary of the remaining chapters.

1.1 WHY MIGHT WE WANT TO BUILD A CORPUS FROM THE WEB?

To answer the question of why we might want to build general or domain-specific corpora from the web five examples are given below. These examples are not exhaustive, but aim to illustrate the types and range of reasons for doing so.

1.1.1 *Informing professionals*

The use of online sources for advice and information has risen over time and has created a need for professionals to be aware of what and how people are seeking and using this information. For

example, The Office for National Statistics noted that approximately 43% of internet users have accessed the web to seek health advice or information. This increased use of the web to seek advice and information on diseases and drugs and their effects has prompted the investigation of how people convey and use this information. For example, Davey et al. (2012) collected forum posts from 8 English-language drug community forums to discover and explore the topics and discussion on newly emerging 'legal highs'. Threads found to be discussing topics surrounding legal highs were then picked for further analysis. In their analysis Davey et al. (2012) were able to identify key information such as methods and range of drug use, group dynamics, group membership and domain vocabulary.

In addition, the collection of domain-specific corpora from the web has also been used for the purposes of advising and aiding professional bodies in their own work regarding disease research and treatments. For example, Leaman et al. (2010) built corpora using posts taken from health related social networks that focus on the side effects of specific drugs. The creation and analysis of these corpora allowed for the discovery of new adverse side-effects from drugs not previously identified by drug manufacturers. To give another example, Santini et al. (2019) used the BootCaT method originally proposed by Baroni and Bernardini (2004) to build a large corpus on the topic of chronic diseases. The purpose of this work was to build a corpus to aid language technologies in the task of "layification" and summarisation of esoteric medical jargon. Methods such as these help professionals access information about both potential patients and treatments. These methods provide a new means of information discovery that might not have been possible through more traditional means, such as qualitative patient interviews. In addition to informing professionals, corpora generated from the web and social media has also be used for the analysis of cultural discourse and event detection.

1.1.2 *Analysis of social media for discourse analysis and event detection*

The emergence of social media has created a powerful new tool, as it provides a means to analyse mass public communications that are used by millions and produces up-to-date information every second. The very current nature of social media content and the persistence

of these communications provides a means for researchers to collect and analyse discussions and posts that appear in large quantities on a global scale. The tools to collect posts provided by social media platforms, such as Twitter and Facebook, provide a window to events and cultural phenomena as they occur. This window into public lives has generated a number of cases for the collection and analysis of topically focused posts from these social media platforms (Finkelstein et al., 2018; Törnberg and Törnberg, 2016). This provides a valuable resource to perform both qualitative and quantitative analysis but requires technologies to find and collect that content. For example, Törnberg and Törnberg (2016) performed digital discourse analysis on the portrayal and opinions of Muslims, by collecting a large 105 millions word corpus from of a large Swedish Internet Forum between 2008 and 2013. Törnberg and Törnberg (2016) binned the collected corpus across a number of time frames in which each post was authored. Each binned sub-corpora had a topic model trained from it and allowed the Törnberg and Törnberg (2016) to identify in what topics Muslims and Islam were discussed, and the general sentiment towards them. Their findings showed that these topics increasingly focused on terrorism and negative feelings towards those who follow Islam. At the time no other study of this kind had been conducted and signalled a progression away from traditional means of discourse analysis, such as interviews or questionnaires as tools for qualitative or quantitative research. Web based data collection and computational methods such as this allow for both quantitative and qualitative analysis of a dataset born from the web. In addition, these methods do not necessarily suffer the same biases inherent to traditional methods, such as interpretive bias of interviews or population variance in questionnaires.

In the previous example the researchers were required to establish their own tools to collect the information they required, but many social media platforms provide tools for the collection of their content. For example, Marres and Moats (2015) used the Twitter API to collect a dataset with the intention of studying debate surrounding divisive topics and events. In their work, Marres and Moats (2015) collected empirical data using the Twitter API by performing hashtag searches relating to privacy during the period when the Snowden scandal broke. Using a Twitter tool that plots content dynamics over time, Marres and Moats (2015) where able to show that over this

time period a large proportion of the discussion around privacy centred around topics that involved the release of information about the Snowden case, such as the collection of data by PRISM and the complicit cooperation and information provided by Facebook and Google.

As previously mentioned tools such as these can also be used for the purposes of event detection. For instance, [Abel et al. \(2012\)](#) who developed Twitcident and [Vicente et al. \(2018\)](#) who developed Talaia, developed methods for the purposes of discovering and monitoring events that require immediate attention as they occur on Twitter. The aim of these tools in both cases was to monitor and discover fires and disasters that demand the immediate attention of fire and disaster prevention services. In each case the methods collected candidate Tweets surrounding a particular issue, performed information extraction, such as people, places and times and presented a means for professionals to react to serious events from a more informed position.

There are also instances where time sensitive information has been intended for similar purposes but outside of social media, such as tackling online crime, an example of which is presented using the CET in Chapter 6 and some pre-existing work presented below.

1.1.3 *Monitoring and capturing illicit activity online*

The online global trade of illicit wildlife has been found to be a considerable driving force in the significant decline in endangered species, such as the African elephant ([Wittemyer et al., 2014](#)). Although this observed sale of illicit products has found to be common, the identification, monitoring and capture of cases has proved difficult as evidence of buyers, products and sellers is often hard to find. This can be due to reasons such as lack of knowledge about its location on the web or the transient nature of items for sale. Enabling the discovery of the markets in which these items are sold can prove extremely useful in combatting this illegal trade. Solutions to discovering these domains have included proposed¹ solutions for collection and filtering content to discover instances of illegal wildlife trade across the general web ([Oliveira, 2014](#)). The use of the

¹ but not implemented

internet to identify markets in illegal items is a complicated search and constantly evolving problem that requires technology that can evolve with it. Digital methods such as those presented in this thesis, propose one viable solution to this problem.

1.1.4 *General corpora for research*

The previous work discussed in this section has focused primarily on methods and corpora designed for specific tasks or research questions. However, corpora are often built for the purpose of servicing the wider research community. Typically these have been large, static knowledge bases, such as the British National Corpus (BNC) ([Libraries, 2007](#)). Though these are often useful for many machine learning tasks there are sometimes instances where general corpora need to have a more specific purpose, and where domain-specific corpora generation methods can be applied. For instance, in the field of machine translation and linguistics there is a need for large language specific corpora that are either general in nature or domain-specific. Typically these types of corpora fall in to the category of parallel or comparable corpora depending on their content. Parallel corpora are two corpora that each contain the same documents but each individual corpus is written in a single language different to the other. Comparable corpora serve as examples of documents from a target domain that are the same in nature to the second corpus and in a different language, but do not contain the exact same documents. Several methods for generating parallel or comparable corpora have been developed to find examples of the same or similar documents in some target language ([Rapp et al., 2016](#); [Jakubíček et al., 2013](#); [Rahimi et al., 2016](#); [Scannell, 2007](#); [Baroni et al., 2009](#)). For example, [Rahimi et al. \(2016\)](#) presented a Multigent System (MAS) for building domain-specific comparable corpora from the web. This was achieved through a process of extracting features from a source corpus in one language and using them to train a learning model, such as a classifier, and also construct search queries. Using a novel ranking method based on the similarity of search results with a source corpus, [Rahimi et al. \(2016\)](#) were able to create a method for building corpora semantically similar to the source corpus, but in different languages.

The previous four subsections have demonstrated five examples of why one might attempt to build domain-specific corpora using online sources. The next section discusses existing solutions, by organising them into three main categories: web search, crawling and MAS. The purpose of this section is to provide a brief overview of the field and to illustrate the main issues that are addressed in this thesis.

1.2 EXISTING SOLUTIONS

The increasing need and applications of building corpora from the web has generated a number of solutions that are summarised below. The purpose of this section is to provide examples of some of the methods that already exist and is used to precede a discussion highlighting where these methods are lacking and what contributions this thesis provides. The coming subsections briefly exemplify four categories of solution².

1.2.1 *Web search*

One of the most commonly known solutions to building domain-specific corpora from the web is web search-based solutions, known as bootstrapping. This method is discussed in more detail throughout this thesis and so is only summarised here. Bootstrapping is a method originally proposed by [Baroni and Bernardini \(2004\)](#) that seeks to expand an existing document or *seed* corpus with documents found on the web (([Baroni and Bernardini, 2004](#); [Baroni and Ueyama, 2006](#); [Baroni et al., 2009](#))). In their work [Baroni and Bernardini \(2004\)](#) proposed a bootstrapping method referred to as BootCaT. The BootCaT method has two key components: feature extraction and web search. BootCaT follows a cyclical process of extracting characteristic features from a target corpus and constructing queries that are then presented to a search engine. The results are then scrutinised by an analyst, who adds documents found to be relevant to the target corpus. This process is then repeated until a sufficient size and quality corpus has been constructed. This cyclical process achieves two key tasks. First, a corpus is iteratively expanded with documents known to exist in the target domain. Second, the target domain vocabulary is also expanded as the corpus increases. Increasing the target corpus has the potential to affect the

² That are covered in more detail in Chapter 2

characteristic features extracted and the subsequent search queries generated. Achieving this second task allows for a wider variety of search results and produces a feedback effect of increasing the variation and potential domain coverage of the final corpus. This is also a key characteristic of the CET and so it is worth noting at this point.

1.2.2 *Web crawling*

Web crawling is another common method for retrieving domain-specific content from the web as it provides an efficient automated means to capture content at scale. One key issue with a standard web crawler is its lack of any domain knowledge and what it should be collecting or searching for. A corpus generated from a general crawl of the web will typically consist of very few, or no documents from the target domain. To address this problem three solutions are typically used: seed choices, filtering using classifiers and focused crawlers. One of the simplest means to influence what a web crawler collects is through careful selection of seed URLs. The seed URLs dictate the starting point from which a web crawler searches and choosing sites taken from a target domain increases the likelihood of finding relevant web pages. One example of this solution was proposed by [Vieira et al. \(2016\)](#), who augmented the BootCaT methodology by using web pages and sites found to exist in the target domain as seeds to a web crawl. The assumption here is that there is a higher likelihood that pages written in the target domain vocabulary link to others also from the same domain. This assumption is flawed however, due to the nature of the internet there is often a large number of links to irrelevant content included on any site. For example, links which navigate to adverts for items for sale on e-commerce sites are often irrelevant, but common on any website.

Filtering crawled content using classifiers helps to address the above problem by automatically classifying documents returned by the crawler and filtering out those found to be unlikely of originating from the target domain ([Medrouk et al., 2016](#)). A common issue with using classifiers is a lack of efficiency as it does not control the direction of the crawl search path. As the crawl branches out across the web there is the potential for diminishing returns as it reaches beyond the original sites and pages deemed relevant.

The common solution to this problem is the use of focused crawlers. Focused crawlers are constructed to only return documents considered relevant to the target domain and in many cases prune the crawl search path when relevant content fails to be found (Remus and Biemann, 2016). Preventing search of irrelevant paths increases yield and addresses the problem of diminishing returns. The key problem to these methods is that it often requires a significant amount of a priori knowledge of the domain vocabulary, in the form of a training corpus, to properly inform the crawler. It also completely automates the process and removes a significant opportunity for the analyst to control or intervene in the web search. These methods are often task specific and do not provide a general solution to generating domain-specific corpora.

To combine the advantages of methods, such as web search and crawling, others have proposed combined solutions in the form of Multigent Systems, discussed below.

1.2.3 *Multiagent systems*

Multiagent Systems combine a number of components, referred to as agents, tasked with a specific sub-goal that collectively achieve an overall goal (Woolridge and Wooldridge, 2001). From the perspective of building domain-specific corpora many solutions combine web search, crawl and classification techniques to build corpora. For example, Krishnamurthy et al. (2016) presented the Domain Discovery Tool (DDT), designed to aid researchers discover and explore a domain through an interactive User-Interface (UI). The DDT provides a means to search the web, extract document features, summarise and explore results to construct domain-specific corpora. The main advantage to the method described above is that it brings the analyst to the core of the system and its development, whilst combining a number of web collection techniques and information extraction methods to collect content. This human-in-the-loop philosophy is a key component to this thesis and is discussed in the next section. The key drawback to the current solutions such as the Domain Discovery Tool (DDT) is a lack of features, such as large scale filtering and classification techniques with a focus on bringing the analyst to the center of the process. Although the above work

present reasonable solutions, methods such as these often lack the power of a complete toolkit, such as the CET presented in this thesis.

1.2.4 *Summarising the issues of existing solutions*

The key drawbacks of the existing solutions to building domain-specific corpora from the web can be summarised as follows.

- Lack of features
- Requires a-priori knowledge
- Removes the human analyst

First, a system lacking in tools which could aid an analyst can significantly impact the efficiency of the system and increase the likelihood of missing relevant content. For example, BootCaT does not make use of web crawlers which provide a means to collect documents from the web at scale. Second, requiring a clear definition of a target domain or a large training corpus to exemplify its vocabulary is an issue if no such corpus exists, or the domain definition is poorly conceptualised or unknown. Third, removing the analyst from the process presents an issue if the domain is yet to be defined or the system is required to inform the analyst of the target domain. This last issue is a key focus of this thesis which sits the analyst at the core as a defining component of the system. Using a human-in-the-loop strategy combines the quantitative advantages of automatic, computational document collection and analysis with the qualitative prowess of human analysis.

This section has summarised the existing solutions to building and bootstrapping corpora from the web and briefly discussed the drawbacks of each method. The next section introduces the main thesis of this work and the philosophy that drives the key hypothesis behind the CET. Section 1.3 is followed by a brief introduction to the CET, which illustrates how this thesis is manifested and how it addresses the issues described above.

1.3 THE PHILOSOPHY OF THE CORPUS EXPANSION TOOLKIT

This section will discuss both the key philosophy behind the CET and how it addresses those issues presented in the previous

sections. Namely, that this toolkit puts the power of defining and conceptualising a domain and building domain-specific corpora into the hands of the analyst by implementing a human-in-the-loop strategy. This section is broken down into four subsections. First, a discussion is given on the rationale behind the human-in-the-loop strategy of the CET. The second and third subsections address why we might choose this strategy. Namely, that we are able to bridge quantitative and qualitative methods and address the issues presented in the pre-existing solutions. This section concludes with a summary of the issues this thesis addresses and how it intends to solve them. The remainder of this chapter briefly presents the architecture of the CET to illustrate the general method, the main contributions of the research and a summary of the remaining chapters.

1.3.1 *The human-in-the-loop philosophy and the CET*

A methodology that follows a human-in-the-loop philosophy is intended to place the human analyst at the centre of the method. Keeping the analyst in the loop allows them to evolve an understanding of the domain and its vocabulary by directly experiencing the data. The human-in-the-loop strategy is realised in the CET in three distinct ways. First, through a set of methods that allow the user to interject in the cascade of information through the system, and influence the inputs and outputs of each intercommunicating software component. The filtering process within the CET presents the output of any one component and provides tools to codify and label this output to a number of classes, or remove it completely. Second, to have a direct influence on the components themselves. For instance, providing tools to train classifiers using data born from the method to both experience, interact and define a domain. Third, the CET follows a cyclical process of iterating back over previous stages to generate content from a more informed position, and improve the precision and recall of relevant content collected from the web.

These three factors keep the human-in-the-loop as more information is discovered over the course of many iterations, leading to an evolving system and analyst understanding. The exact details of these software tools are discussed throughout this thesis

and this section only serves to exemplify the MAS architecture that underlies the method of the CET. To summarise, the main focus of a human-in-the-loop strategy is to combine the qualitative analytical skills of a human with the quantitative, statistical analysis of computation methods. This thesis presents a solution that allows researchers to begin with very little information on a specific topic and iteratively build a clear conception of a domain and translate that to a computational system, that allows for the building of domain-specific corpora.

1.4 BRIDGING THE QUALITATIVE AND THE QUANTITATIVE

One issue that arises when doing large scale quantitative analysis of data using digital methods is one of interpretation. Quantitative analysis is often seen as relatively objective and is often considered to ignore the relative bias that can occur as researchers interpret their results. However, [Daniels et al. \(2017\)](#) argued that a common fallacy of large scale content analysis is to claim that one has captured objective realities, whilst ignoring data and researcher bias. Similarly, a key question when using digital methods and performing research online is whether it is appropriate to answer the original research questions. For instance, [Marres \(2017\)](#) argues that performing any substantive sociological or cultural research using only online content, such as social media platforms like Twitter, ignores the real, empirical variables and people that created that content.

One solution is to embrace and accept the inherent limitations and biases that can occur from digital methods, and large scale computational analyses. A second solution, proposed in this work, is to have a clear objective for the research and a clear understanding and interaction with the technologies that incorporates that bias. Developing or choosing methods as part of the project requirement, or to approach the project with dynamic learning objectives mitigates this issue by allowing the data to drive the findings of the analyst. This second solution is more reminiscent of qualitative approaches, which sees researchers perform a much deeper analysis of traditionally much smaller datasets. In this thesis the methods presented represent a close relationship between the qualitative analysis of corpora to drive the theme of its contents, whilst using

large scale quantitative, computational methods to assist in corpus generation.

1.5 WHY NOT JUST USE A SEARCH ENGINE?

Search engines are increasingly permeating our web technologies and have become almost exclusively the means by which users access the internet, and how businesses interact with its consumer base (Ferraresi, 2009). For instance, the phrase '*Google it*' originates from the company name Google, the most dominant search engine and advertising platform on the internet, and is part of common parlance meaning to search the internet for an answer (Ferraresi, 2009; Page et al., 1999). Google and other search engines have created a culture based on 'free' services that generate profit through the collection, profiling and sale of user information (Rogers, 2009). This culture has led to many of the large companies such as Facebook, Instagram and Twitter to commoditise their data and provide a service that is only financially free to the user. This culture has therefore generated much debate about privacy, data freedoms and security, but it also begs the question of the motivation of these platforms and what they present to us as search results to our queries (Rogers, 2009).

This creates a problem for internet research and finding relevant content on the web in two ways. First, a lack of tools to more comprehensively discover what we need on the web. Second, over-dependence on web technologies that have ulterior motives for us using their services. Namely, commercial interest in gathering our data and promoting products and services for profit (Lucas D. Introna, 2000). For instance, Marres (2017) argued that researchers too often...

".. go along with whatever ontology, epistemology or methodology is wired into the platforms, packages or tools they use to capture, analyse and visualize data, without querying whether and how they are appropriate to the research project at hand."

The Corpus Expansion Toolkit (CET) presented in this thesis aims to address this tendency by providing a new set of technologies that can be dynamically defined in two ways. First, the tools and their

inputs can be controlled explicitly for the task and domain. Second, by use existing computational methods in innovative ways to collect domain-specific online content. This is in keeping with the belief that technologies should be designed and used in new ways to create new methodologies for digital media, and not stand apart from the traditional methodologies (Rogers, 2013).

1.5.1 *The issues*

The above subsections can be summarised as four problems that are addressed in this thesis. First, the sheer scale of the web makes finding desired information or domain-specific content incredibly difficult. This is an issue that is traditionally solved by search engines, user queries and ranking algorithms. Second, that search engines present their own issues in the form of their method of recommendation, missed content, poorly formed queries by the user or the user simply not knowing for what to search. Third, that current solutions are highly dependant on the use of pre-existing technologies, such as search engines to perform the work, leading to potentially compromised results. Fourth, at the outset any work the researcher may have little knowledge or background information on the domain in question.

This thesis proposes a solution to the above four problems, that allows researchers to begin with very little information on a specific topic and iteratively build a clear conception of a domain and translate that to a computational system that allows for the building of domain-specific corpora. These forms of toolkit are commonly referred to as a Multigent System (MAS) and are defined in the next section, which provides a brief description of the CET as an example. Throughout the course of this thesis, five case studies are presented that introduce its specific software components and subsystems.

1.6 MULTIAGENT SYSTEMS FOR WEB RESEARCH

To find relevant information on the web, extract the desired content and perform analysis both at scale and at a qualitative level a number of cases studies were conducted which contributed to the development of a complete computational system. Systems of the kind proposed in this thesis combine a number of discrete software

components that each perform a specific task or achieve a sub-goal within the system. These collections of discrete software components that intercommunicate to achieve some ultimate goal are from here onward referred to as Multigent System (MAS). The purpose of a MAS is to divide a large or complex goal into a number of smaller sub-goals that simplify the task. Each software component within a MAS is referred to as an *agent* and each agent communicates the output of its work to one or more other agents. For example, an agent designed to perform the task of crawling the web may communicate its output to a scraping agent that extracts the text content from the html of pages discovered by the crawler. To modularise the process further, a collection of agents can be grouped by those that share a common goal. These collections are referred to as *experts*. For instance, the web crawling and web scraping agents described above could be combined into an expert that achieves the goal of web collection (Woolridge and Wooldridge, 2001). To exemplify a MAS further and briefly present the main contribution of this thesis, the CET is briefly described below and discussed in more detail throughout this thesis.

1.6.1 The Corpus Expansion Toolkit

The CET is an example of a MAS consisting of five distinct experts, listed below.

1. Domain characterisation
2. Human analyst
3. Document generator
4. Document parsing
5. Clustering and classification

Figure 1.1 is a diagram of the CET and a basic representation of all agents and how they interact. Each line of communication within the diagram can be considered a flow of information, or cascade. The cascade refers to the flow of the entire system, which sees data flowing from one end of the system to a final output, or back to previous steps in the chain. The final output being written as annotated documents and stored in a database table.

In Figure 1.1 the solid arrows represent the direction and flow of information through the system. The dotted arrows represent the cyclical nature of the cascade, that allows analysts to return to previous steps and affect the inputs, outputs and parameters of each agent. This cyclical nature is intended to leverage the knowledge learned from the cascade and better inform preceding steps, thus creating a positive feedback effect. The analyst *expert* represents the ability to inspect, filter, add or otherwise change the information at various points within the cascade, which is how the strategy is manifested. This requirement for human input, combined with the advantages provided by automated computational methods are what underpin the power of this system and are a significant contribution of this work.

This section has now completed the part of this introductory chapter that specifies this work's area of research, and provided a discussion on the problems this thesis addresses. The final sections of this introduction summarise the main research questions and contributions of this work and the organisation of the rest of the thesis.

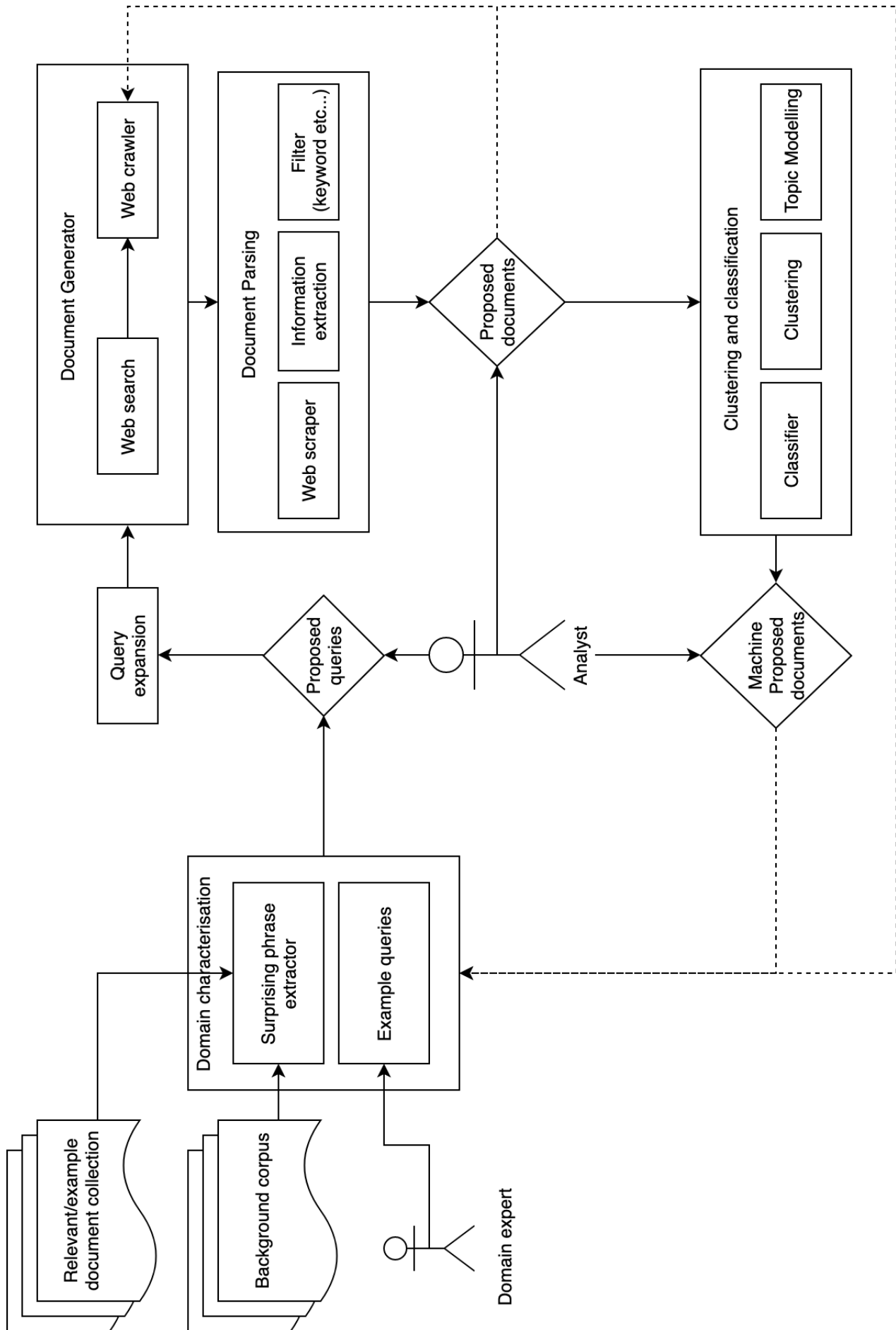


Figure 1.1: The Corpus Expansion Toolkit data flow diagram

1.7 RESEARCH AIMS AND CONTRIBUTIONS

The first part of this introduction provided the scope of this research as a piece of work that concerns using the web as a means to discover information on a topic and build web based domain-specific corpora. In addition, it achieves this using a method utilising a human-in-the-loop strategy to keep the analyst at the core, creating an evolving system and conceptualisation of the domain and its vocabulary that is inherent to the system. This work can be broken down into four main research aims described below.

First, to provide a complete set of tools to build domain-specific corpora from the internet that combines quantitative computational collection and analytics with qualitative, human analysis. One question that is raised when developing tools of this nature is whether it is generally applicable and comprehensive. To present a solution to this question the case study based methodology used in the thesis illustrates the development of the systems across a range of projects with a variety of research aims.

Second, to provide a general toolkit that utilises a human-in-the-loop strategy, that puts the power in the hands of the analyst. The aim of having a focus on keeping the human-in-the-loop is to combine the advantages that computational autonomy provide, namely processing documents at scale, with the qualitative analysis and reasoning that currently only a human can provide. The CET aims to do this by providing a system comprised of stages that have the option to manually filter the results of a given stage, such as information extraction, or during the configuring and training of a particular component, such as a classifier. The analyst then has the option to perform or forgo these stages as they see fit according to requirements or constraints, such as time scales. Putting this power in the hands of the analyst allows them to evolve and understanding of a domain and to translate that understanding to a computational method for document discovery.

Third, to provide sufficient evidence of the efficacy of the CET through three illustrative case studies that document its inception.

Fourth, to present a case for the value of this toolkit and the human-in-the-loop strategy by presented a number of real-world application of the method. Finally, considering the main aims of this thesis, the three main contributions it provides to the field can be summarised as follows.

- A complete and user-friendly software toolkit to perform research and build corpora on the web at scale that is suited for use by non-technical domain experts.
- To provide a research tool that is novel in two specific ways. First, it provides a more comprehensive solution than previous work in this area. Second, emphasis on a human-in-loop strategy to combine the qualitative advantages of human analysis and quantitative advantages of computational and approaches.
- To provide empirical evidence of the power of the CET, through the use of case-studies, that provide novel solutions to pre-existing problems in a wide variety of research areas. More details of these can be found in the organisation of this thesis given below in section 1.8. Specifically, summaries of the chapters 4 - 6.

The aims and contributions of this thesis have now been given and conclude the discussion of the scope of this thesis. The final section of this introduction describes the organisation of the rest of the thesis, that precedes the related literature.

1.8 ORGANISATION OF THE THESIS

This thesis presents a fully implemented multiagent system for quickly building domain-specific corpora from the web and discovering data relevant to a particular task. The organisation of the rest of this thesis is as follows.

Chapter 2: Background

Chapter 2 presents the background of this thesis. Listed below is a summary of each subsection, that discusses work related to building corpora from the web.

- Subsection 2.1 discusses bootstrapping corpora from the web through feature extraction and web search.
- Subsection 2.2 discusses literature focused specifically on the use of search engines to build corpora from the web.
- Subsection 2.3 presents work based on building domain-specific corpora using web crawlers.
- Subsection 2.4 presents work relating to multiagent systems designed to build web based corpora.

Specific attention is given to the work in Section 2.4 because it most closely relates to the work conducted in this thesis.

Chapter 3: Methodology

Chapter 3 begins with overview and discussion on the form of case studies undertaken and the kind of contributions they were designed to make. The methodology presented in Chapter 3 is broken down into 3 sections. Section 3.1 discusses the case study based methodology employed in this thesis and summarises each of the case studies that contributed to the development of the CET. Each summaries begin with the purpose and aims of the study, the technology developed to achieve the study outcomes, and a brief discussion of its contribution to the CET. Section 3.4 discusses the application-based method of evaluation this thesis employs. Finally, Section 3.5 introduces the software platform Method52, that was used to implement the CET.

The three chapters following Chapter 3 present each of the five case studies that introduce the individual components and sub-systems of the CET, that were used in a variety of applications.

Chapter 4: Filtering crawled content

Chapter 4 presents a case study that illustrates the original prototype system that was developed to enable domain experts to discover the voting rights of diaspora on the web. Chapter 4 begins with the problem definition and what originally inspired the work. Chapter 4 describes the original system, referred to as the Semi-automated Web-Search and Analytics Tool (SAWSAT), and how the system operated to help discover the voting rights of diaspora. Section 4.4 discusses the design process and each component within the system. Chapter 4 concludes with a discussion of the findings and what was learnt over the course of the study.

Chapter 5: Building corpora from known sources

Chapter 5 presents three case studies, each requiring corpora built from known sources, such as forums and new reporting web sites.

The purpose for each case study was to perform an analysis of the topics of discussions around the topics, mental health, child welfare and world conflict. At the time of conducting this research, the corpora constructed represented topics that had little or no attention paid, and was intended to provide an insight for professionals to better understand the discussions and articles that occurred in these topics. Chapter 3 provides a more detailed description of each of these case studies.

Chapter 6: Discovering domain content from unknown sources

Chapter 6 discusses a project that was done in collaboration with The Global Initiative to refine the CET and use it to discover instances of illegal wildlife trade online. More specifically, this project sought to discover instances of people selling ivory, pangolin scales and endangered orchid species. This project details the first use of the complete CET system, whereas previous case studies illustrate a number of potential subsystems. Chapter 6 presents the final system and discusses various refinements and findings of the method.

Chapter 7: Conclusions and future work

The final chapter of this thesis covers the overall conclusions drawn from this research and intentions for future work. Namely, that a significant amount of automation can be introduced to allow the analysis of web content at scale, but human intervention is still required to guide domain definition, training and analysis. Section 7.1 details a number of extensions and adaptations to the existing system, that has the potential to increase the precision and yield of relevant content discovered by the CET.

BACKGROUND

The CET is a system for bootstrapping and building corpora using content found on the web for the purposes of wider research, corpus expansion and directed search. The CET consists of four main research areas, listed below, that are organised into a semi-automated system of autonomous agents, known as a MAS. This literature review will therefore background of these four key areas.

1. Bootstrapping corpora and web search
2. Search
3. Topical web crawling
4. Multiagent systems

The background of this thesis is therefore broken down in to four sections, each covering one the four areas listed above.

2.1 BOOTSTRAPPING THE WEB

Using the web to generate a corpus or to perform research is a wide reaching area, as its scope is potentially unlimited depending on what research is being conducted or the ultimate purpose of the corpus being built. The web consists of billions of sites and pages covering a wide range of topics and languages that makes it a potentially highly valuable resource, but one that also can be very hard to navigate and find relevant content. One solution to this problem uses the paradigm generally referred to as bootstrapping. Bootstrapping utilises a small set of *seed* prototype documents or phrases that characterise the domain to discover similar documents on the web (Hearst, 1992). One of the most well known developments of this paradigm was originally presented by Baroni and Bernardini (2004) via their corpus bootstrapping methodology and toolkit BootCaT. BootCaT is an unsupervised bootstrapping methodology that can be broken down into four stages, that are illustrated below in Figure 2.1 (Baroni and Bernardini, 2004; Leturia and Vicente, 2009;

Baroni and Ueyama, 2006; Baroni et al., 2009; Baroni and Kilgarriff, 2006).

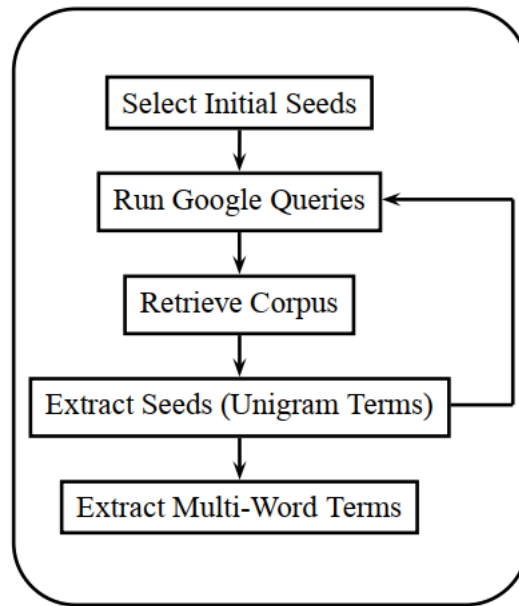


Figure 2.1: The BootCaT flow diagram

The process begins by extracting or selecting some seed words or phrases characteristic of the domain vocabulary that are then randomly combined to generate a list of search queries. These queries are then sent to a search engine and the top n results of each query returned and aggregated to create a new corpus. The phrase extraction process is then repeated by extracting further unigram seeds from the new corpus. Characteristic unigrams are found in a target corpus by comparing the frequency of each word it contains with the frequency of the same words in some reference corpus, in this case the Brown corpus, using the log odds ratio measure originally proposed by [Everitt \(1992\)](#) ([Francis and Kucera, 1964](#)). Words that appear with the highest likelihood, or surprisingly high frequency when compared with the reference corpus are assumed to contain a large amount of information about the target domain, and therefore used as seeds to the next iteration. This process allows one to quickly generate a corpus by repeating the process until the desired size or relevance of a corpus is obtained. To extract multi-word phrases, intended for the same purpose as unigram seeds, [Baroni and Bernardini \(2004\)](#) devised a method to expand terms starting from a previously extracted unigram. This process

can be summarised as follows. First, a number of *connectors* are identified within the corpus by looking for bigrams that frequently occur with a single word between them (e.g. *of*). These words that appear between bigrams are considered as common phrase *connectors* in the corpus. Once identified, all *stop-words* can be removed from the corpus, minus the connectors. Phrases are then searched for recursively, starting from bigrams the potential phrase is expanded left and right for a frequently occurring $n+1$ gram, which satisfies a number of constraints. For instance, the phrase must have a frequency above a certain threshold. These multi-word phrases can then be incorporated within the seed search terms from which to bootstrap the corpus.

To evaluate this method, Baroni and Bernardini (2004) envisaged a scenario of a translator attempting to translate a psychiatric article into English. The seeds in this case were 6 seed words selected from the target article's abstract in English and their Italian acronyms. Using the bootstrapping procedure, a set of unigram and multi-word phrases are extracted and sent as queries to the Google search engine. The corpus is bootstrapped by the analyst adding all relevant search results to the target corpus. To evaluate the method, 30 pages were randomly selected from each generated corpus and 100 unigrams and 100 multi-word phrases were also randomly selected. To evaluate the generated corpus' quality the randomly selected pages were assessed by human evaluators. Pages are given positive assessments if they were found to be in keeping with the domain and were seen as informative of the original seed document and topic. In their results 20 out of the 30 were found to match this criteria. To evaluate the precision of the word lists human evaluators assessed them for *well-formedness*, of which 73 were considered "good" by their assessment. To evaluate the recall of the same term and multi-word term lists Baroni and Bernardini (2004) sought to establish how many of these terms appeared in the source and target corpora. It was found that 88% of the unigram terms appeared in both the source and target corpus, and 38% of the multi-word phrases also appeared in both. As a result this methodology was proposed as a successful, fast and practical means to automatically generate corpora from online sources.

This methodology, pioneered by [Baroni and Bernardini \(2004\)](#), has since undergone a number of adaptations and uses that suit a specific task or corpus. For instance, [Kilgarriff \(2006\)](#) developed WebBootCat an online toolkit that allows others to generate a corpus using the BootCaT methodology. This toolkit was also then later integrated into the online linguistic analysis toolkit Sketch Engine, for the use of linguists to collect and analyse relevant online content ([Kilgarriff et al., 2019](#)). More recently this method of multi-word phrase detection and expansion has been developed upon by [Robertson \(2019\)](#) which adds a number of additional parameters to improve the method's capability to expand phrases and rank the results. This method is discussed in more detail in Chapter 6 as it is utilised in this thesis.

2.1.1 *Web for linguistic corpora: comparable and parallel corpora*

Some of the major proponents of bootstrapping corpora have been in the field of linguistics, machine translation and generating general linguistic, comparable and parallel corpora ([Baroni et al., 2009](#); [Jakubíček et al., 2013](#); [Rapp et al., 2016](#); [Scannell, 2007](#); [Benko, 2014](#); [Fiser et al., 2011](#); [Fantinuoli, 2005](#); [Gao and Vogel, 2011](#)). Comparable corpora consist of two distinct document collections each comprised of documents in a different language, but originating from the same domain. Comparable corpora are commonly used to exemplify domains in different languages for machine translation tasks. Parallel corpora are two corpora containing the same documents, but each corpus is written in a different language ([Kenning and McCarthy, 2010](#)). In many cases bootstrapping is used to take an existing corpora in one language and discover similar documents on the web in the target language, or in the case of parallel corpora, the exact same document in the source corpus but in the target language ([Coanca and Museanu, 2010](#); [Fantinuoli, 2005](#); [Santini et al., 2019](#)).

Aside from creating general, comparable or parallel corpora the field of linguistics has used corpus expansion and bootstrapping strapping to create word and sentence segments. [Qiu et al. \(2014\)](#) used a corpus expansion methodology to create a better training corpus for a Chinese language word segmenter. To achieve this [Qiu et al. \(2014\)](#) first trained a standard word segmenter using a manually annotated training corpus C_a and used it to segment

the sentences of another unannotated corpus C_t . To evaluate this method, a confidence value for each segmented sentence was first calculated. Using this confidence value Qiu et al. (2014) selected those sentences with a value below a given threshold and used them as seed queries to a web search. The pages taken from the results of this search had their raw text extracted, the segmenter used again and the confidence value for each new sentence calculated. Those with the highest confidence values above a given threshold were then added and used to train the original segmenter further. Using this method, Qiu et al. (2014) were able to produce an error reduction of approximately 26% over the baseline.

The above examples show how bootstrapping corpora from the web can be used to generate corpora for a wide variety of purposes and languages. The general aim of all methods described in this thesis are to generate corpora in a single language and often very specific domain, which is the focus of discussion in the next set of literature discussed.

2.1.2 *Bootstrapping domain-specific corpora*

At the beginning of Section 2.1 WebBootCat was briefly mentioned, which is an online implementation of BootCaT for the use of others to bootstrap corpora (Kilgariff, 2006). PVS et al. (2012) utilised this tool to produce a methodology for building domain-specific corpora from the web. In their work, PVS et al. (2012) augmented the BootCaT methodology (via WebBootCaT) originally proposed by Baroni and Bernardini (2004). To ensure an increase in the domain specificity of corpora returned from search, PVS et al. (2012) utilised DANTE a lexical database of 1.7 billion English corpus created by a group of lexicographers (McCarthy, 2010). Each word in DANTE has some lexical metadata associated with them, one of which is the subject field of a word as being in one or several of 156 domains. PVS et al. (2012) chose eight domains for their experimental work. Seed words were chosen for each domain by scoring and ranking each monosemous word in the corpus using a similar keyword extraction method as Baroni and Bernardini (2004), and the British National Corpus (BNC) as a background corpus (Libraries, 2007). In addition, PVS et al. (2012) leveraged their generated search queries by using human analysts to filter unrelated or ambiguous terms

and phrases. This methodology therefore sees a development over the fully unsupervised method proposed by [Baroni and Bernardini \(2004\)](#) through the introduction of the previously mentioned human-in-the-loop philosophy. Additional work has been conducted that performs bootstrapping through phrase extraction and search as part of a wider system for domain-specific corpora and are therefore discussed in the relevant section of this chapter.

This section discussed literature that presents some of the key methods and uses of bootstrapping, that utilise a set of seed documents to explore the web and generate a larger corpus. This section covered bootstrapping, which uses phrase extraction and search, that are two of the major components complete CET, discussed in Chapter 6. The remaining three sections of this background chapter discuss generating domain-specific corpora using other web search methods, topical web crawling and MAS that bring these components together as a complete system.

2.2 WEB SEARCH

One of the main components of bootstrapping corpora using the methodologies described above in section 2.1 is the use of search engines as a means of finding relevant documents on the web. In this section a number of relevant works concerning the outputs and uses of web search are discussed. It is worth noting that [Arora and Bhalla \(2014\)](#) identified three types of search engines, crawler based, human powered search engines, that require people to submit their content to the index, or a combination of the two. The search engines referred to in this thesis only concern the typical crawler based search engines, such as Google and Bing.

2.2.1 *Domain relevance and result filtering*

[Ferraresi \(2009\)](#) proposed that search engines provide a means to leverage the entire web to create corpora by using methodologies, such as bootstrapping. However, [Ferraresi \(2009\)](#) identified an issue with web search in that search results come in the form of an unstructured list of results with no linguistic information provided. Issues such as this mean that researchers, such as those building linguistic corpora, must undertake additional work creating systems

that go beyond these result lists, in order to organise and otherwise construct a corpus. In identifying the issue that search engines only provide partial answers, (i.e. web links to information it has algorithmically deemed relevant) others have sought to produce methods for directly analysing the quality of search results. For example, [Rahimi et al. \(2016\)](#) presented a bootstrapping methodology to construct comparable corpora by ranking search results based on each result's similarity with some source document that characterises the target domain. In their work [Rahimi et al. \(2016\)](#) first extract five 'driver' queries from the source document. These driver queries taken from a page include the content, title, originating website description and the parent directory names within its original URL. Queries were then searched using the Google Web API and the top 8 results for each query kept as possible candidates. RankingSVM, a support vector machine-based ranking algorithm was used to compare these results with the source document and re-rank them based on their similarity. To evaluate their work [Rahimi et al. \(2016\)](#) generated a comparable corpus in Persian and English to compare its quality with two comparable corpora, generated following the state of the art UTPECC corpus generation methodology ([Hashemi and Shakery, 2014](#)). Each generated corpus was evaluated on how comparable each monolingual corpus was to its counterpart using the standard 'binary Comparability Measure' (binCM) and 'cosine Comparability Measure-Vector Space Model' metrics introduced by [Lange et al. \(2010\)](#). Their results achieved similar comparability scores to the then state-of-the-art, and can be seen as an augmentation to the original bootstrapping methodology proposed by [Baroni and Bernardini \(2004\)](#).

Similarly, [Leturia and Vicente \(2009\)](#) wanted to bootstrap corpora by improving on the original BootCaT method and increasing the domain-precision of collected search results. In their work, [Leturia and Vicente \(2009\)](#) extended the methodology of [Baroni and Bernardini \(2004\)](#) by taking the results of BootCaT and filtering them by domain relevance. [Leturia and Vicente \(2009\)](#) achieved this by representing each document in the source and target documents as a feature vector of extracted nouns, proper nouns, adjectives and verbs. Each feature was then assigned a frequency based weight, such as log-likelihood ratio. [Leturia and Vicente \(2009\)](#) used cosine to measure each target vectors similarity with each of

the source documents. A newly collected document was accepted as being within domain if its maximum cosine similarity across all comparisons with the source documents was above a given threshold. Additional methods that go beyond search are discussed in the coming sections that look to expand on search results and post-process them, such as Multigent Systems discussed in section 2.4.

2.2.2 *Search engine effectiveness and ranking*

The relevance of search results is one of the most important factors for any search engine implementation, regardless of its intended purpose. This is especially true for web-as-corpus research, such as those above, that rely heavily on the results to construct a domain or language specific corpus. In addition to methodologies that attempt to create domain relevant queries or process the output of search queries, the way in which a query is presented to a search engine can also affect results.

Imani et al. (2019) identified that search queries are traditionally presented as a bag-of-words or n-grams, concatenated with some logic, such as AND and OR. However, Imani et al. (2019) hypothesised that if the ordering of individual terms affects the semantics of the query or sub-phrase it can have a significant impact on the results. To test this hypothesis Imani et al. (2019) employed axiomatic analysis, originally developed by Fang et al. (2004), a methodology for the construction of a systematic analysis of possible variables. Imani et al. (2019)'s methodology consisted of measuring the probability that a constructed query is relevant to some target document(s), based on the document frequency of the phrase constructed by the imposed ordering. The phrases probabilities were based on taking q_1q_2 , where q_1 and q_2 are terms within a constructed query, and calculating their document frequency of that ordering within a context window of 5 words within an evaluation set. The phrases were compared by reversing each phrase to q_2q_1 and establishing how this changed the relevance of the query within the evaluation corpus. In this work, Imani et al. (2019) provided evidence that when building domain-specific corpora using search engines, the word ordering of query terms can change the semantics of that query and can significantly impact results. The main contribution of

this work is that care should be taken when constructing queries if the domain relevance of results is crucial.

More recently [Pham et al. \(2019\)](#) presented work in search rankings and domain-discovery through an unsupervised method in the form of DISCO. This work is presented as the antithesis to human-in-the-loop strategies which [Pham et al. \(2019\)](#) argue is time consuming and difficult; as it requires domain experts to filter large amounts of information in order to build classifiers and otherwise supervise domain-discovery. The DISCO tool was developed to bootstrap corpora using a small number of examples provided by domain-experts and follows a ranking-based methodology designed to imitate the way in which users search for desired content on the web. This imitated method sees domain-experts entering queries to the web and then reviewing the results based on domain-relevance. [Pham et al. \(2019\)](#) use search engine APIs to automate this process and create an unsupervised method that takes a list of relevant sites as inputs and outputs that same list with newly discovered sites appended and re-ranked in order of relevance. The relevance-based ranking of discovered sites in this case was based on the results of six ranking metrics: binomial regression, positive and unlabelled example learning, novelty detection, similarity-based ranking, bayesian-sets-based ranking and ensemble ranking. Ensemble ranking proposed by [Pham et al. \(2019\)](#) was used as a means to combine the results of the five other ranking methods because each method bases its rankings on a different set of features. The calculated ensemble ranking score for some site w is computed as an average presented below in equation 2.1.

$$Score(w) = \frac{\sum_{f_i \in F}^{max} f_i(w)}{|F|} \quad (2.1)$$

Where in equation 2.1 $F = f_i$ is the list of the ranking functions and $f_i(w)$ demotes the position of w in its originating ranked list. This approach seeks to combine the different advantages of each ranking metrics and mitigate for significant differences in variance between them. To perform the search, a combination of web crawling, keyword search and related search using Google and Alexa APIs is used to collect potentially relevant content from the web. The web search method chosen on any given iteration of DISCO is decided using the Multi-Armed-Bandit (MAB) approach, which uses a reward

and punishment scoring system to evaluate the effectiveness of each search method based on the relevance of results it returns. The relevance score for a search method is based on where the collected results of a given search method rank in the current complete list of sites. A binary score of 1 is given to each site near the top of the list and 0 to those close to the tail. The score for each search method is periodically updated and the highest scoring search method used at the next iteration. On each iteration, the top k sites of the ranked list are used to seed the next iteration of the system. The algorithm for DISCO can be seen in Figure 2.2.

Algorithm 1 Website Discovery

```

1: procedure Discovery(seeds)
2:   results =  $\emptyset$ 
3:   ranked_results =  $\emptyset$ 
4:   topk = seeds
5:   while stop condition do
6:     op = select_discovery_operator()
7:     search_results = op.search(topk)
8:     results = results  $\cup$  search_results
9:     ranked_results = rank(results)
10:    topk = get_top_k(ranked_results)
11:  end while
12:  return ranked_results
13: end procedure

```

Figure 2.2: The DISCO algorithm

Here the *select_discovery_operator()* function uses the MAB to choose a search method, which is then used to discover new sites on the web. The search results are then combined with all results and ranked using one of the ranking methods or the ensemble method previously described. DISCO then continues until some stopping condition is reached and the ranked list of all results returned.

To evaluate the domain-discovery capabilities of DISCO Pham et al. (2019) compared their system with a number of state-of-the-art domain-discovery methods. To perform the comparison between DISCO and other domain-discovery tools, Support Vector Machines (SVM) based classifiers were trained using domain-expert labelled data as a gold-standard from one of four domains listed below.

- Human Trafficking(HT) - classified ads and and escort related adverts.
- Weapons Forums - forums discussing firearms.
- Weapon marketplaces - websites and classified ads for weapons sales.
- Stock promotions - stock clearance and promotion sites.

These domains were chosen due to the variance in vocabulary between each of them, their access to the domain-experts in these areas and access to potential training data. [Pham et al. \(2019\)](#) chose an extrinsic evaluation method, that used four SVM-based classifiers trained for each of the four domains chosen by the authors. DISCO and each of the baselines were tasked with retrieving 50,000 web pages for each domain, which were then classified for relevancy using the appropriate classifier. The *harvest rate* for any method was a proportion based on returned pages that are classified as relevant to the domain. Their results showed that in the worst case their method matched the state-of-the-art and in the best case increased harvest rate by more than 300%.

2.2.3 *Example-based search*

In the field of IR example-based search is something that has had a considerable amount of research conducted. Although this thesis should not be considered an implementation of example-based search it will be briefly touched upon below. The reason for this is that in some respects using the web, and specifically search engines, to bootstrapping or build a corpus based on some pre-existing concept or exemplar corpus could be considered as a form of example-based search. Traditionally, example-based search systems, similar in nature to domain-discovery, have been used in systems such as recommendation systems and search ranking methods, for a single domain or environment ([Pu and Kumar, 2004](#); [Shoham, 1998](#)). These serve users with specific search criteria or content desire informing a system through their search terms and selected choices. Recommendation systems often conceptualise a domain by building profiles of user's preferences, such as collaborative filtering and content filtering ([Jannach et al., 2010](#); [Faltings et al., 2006](#)). More recently, example-based search has expanded into entity search to

discover information on the web linked semantically through shared entity mentions, by leveraging structured entity relations found in common online knowledge bases, such as the Linking Open Data (LOD) cloud (Insight, 2019; Bron et al., 2013). Methods such as this are designed to create structural links between previously unlinked pages on the web, that share a similar vocabulary via their disambiguated entities.

In addition to traditional methods of example-based search Singh et al. (2018) presented *Expedition*, an example-based search system for domain and document discovery on the web. Expedition resides in Temporal Information Retrieval, which is a time sensitive means of judging document relevance. The assumption made in Temporal IR is that a document's date and time of creation contributes to its relevancy within a proposed corpus. Expedition presents users with a UI that allows for keyword search over a pre-existing corpus, which for their purposes was the The New York Times Annotated Corpus (Sandhaus, 2008). This search can be conducted using one of five choices of relevancy: temporal, textual, topical diversity, temporal diversity or historical diversity. Results are presented in news article form, with more real estate given to those with higher relevancy. The user is provided a number of temporal 'bursts' showing time periods with high relevancy to the original query. This allows the user to explore the corpus in relation to the temporal relevancy of documents. The user generates a corpus through a process of document exploration, whilst labelling those discovered to be of interest. In this sense the user generates an informed view of the dataset that drives the choice of search queries and iteratively generates a domain-specific sub-corpus from the original. The next subsection briefly highlights some caveats of corpus generation through search, that introduces the next section of this related literature, which discusses topical/focused crawlings.

2.2.4 Caveats of search engines

The use case of search engines for everyday users can be different from that of researchers, which can potentially present a number of caveats when using search engines to create or bootstrap corpora. For instance, Spink et al. (2001) found by studying over 1 million user queries that people tend to use very few search terms and

rarely use the advanced search features that are often available. More importantly, Spink et al. (2001) found that the highest frequency searches concerned topics surrounding recreation and entertainment. This presents a stark contrast between the average user and a researcher potentially attempting to find esoteric or domain-specific content across a potentially wide variety of styles (e.g. blogs, academic articles and new reports), topics and languages. This presents a problem as there is a potential gulf between the search engines design and purpose, which is guided by its most common uses, and the intended function of the researcher. To address this issue work has been conducted using web crawlers to perform a search of the web and generate corpora. These focused crawlers allow corpora to be generated from online sources, whilst removing the over reliance on search engines.

2.3 WEB CRAWLING

This next section of the background discusses work concerned with web crawling and crawling strategies. This section is broken down into two main areas of functional and topical/focused crawling. The former discusses general methods that have been developed to enable efficient control of crawler behaviour. The latter covers work that presents methodologies and case-studies designed to create domain-specific corpora.

Before discussing any specific research it is worth defining some of the most basic concepts around web crawling and web crawling strategies. The two main components of a crawler are the crawler itself, that takes some *seed* URLs from which the crawl begins. The crawler works by extracting the links contained in these seeds and continues to repeat the process — referred to as the *spider* — until some stopping condition is reached. The second component is the *indexer*, which concerns the structuring and storage of any extracted web content in a manner that makes it easy to use and analyse. The crawling component can be instructed to search either by a breadth first or depth first strategy (Olston and Najork, 2010). The former extracts all links within a page before crawling further, and the latter crawls outwards from the first extracted links it finds within a page to some stopping condition before moving to the next link that was found on the original page. In the majority of

cases the breadth-first search strategy is used, but there may be a task where depth-first is preferable (Raja and Akorli, 2011). A number of stopping conditions are typically given as parameters to prevent the crawler running indefinitely unless that is the desired behaviour (such as continuous crawling). The depth of a crawl is the most common of these parameters, and specifies how many increments outwards from the starting seed URL a crawler should go before terminating. The two subsections below first discuss some examples of functional methodologies for recrawl strategies and secondly information and boilerplate identification used to extract or scrape content from crawled web pages. The final subsection covers topical/focused crawling strategies.

2.3.1 *Recrawl strategies*

Recrawling strategies are those which attempt to improve the efficiency of a web crawler by controlling when or if a page should be revisited and its content extracted, based on a page's likelihood of containing *ephemeral* or persistent content (Olston and Pandey, 2008). Crawlers which implement a recrawl strategy often represent a quantifiable trade-off between two factors: *freshness* and *coverage*, which govern how often a page should be revisited. The coverage of a crawl database is a measure of how much of the desired source, such as an entire web site, has been collected by the crawler. Freshness is a measure of how closely the older static content in the database represents the current content of the source. Typically, work in this area attempts to improve the efficiency of the crawler by maximising freshness so that coverage is not significantly sacrificed, as any strategy chosen presents a trade-off between the two (Olston and Najork, 2010).

Recrawl strategies are commonly referred to as page-refresh policies and typically maximise freshness by estimating the rate of change a page has relative to the frequency of visitation by the crawler. An early strategy, proposed by Liu (1998), conceived of the refresh problem as a function that minimises the cost of recrawling sites in terms of the rate of change a page exhibits and how many times each page must be visited in order to maximise freshness. This work was later developed by Cho and Garcia-Molina (2003a) who posited and empirically proved that the rate change individual

pages have within a crawled database can be represented using a Poisson process of random events. Using this representation [Cho and Garcia-Molina \(2003b\)](#) presented a practically implementable method for calculating this rate of change by estimating the proportion of times each page changed relative to the number of times it was regularly checked for change by the crawler. In this same work [Cho and Garcia-Molina \(2003b\)](#) also presented a means of calculating rate of change using a maximum likelihood estimator for pages crawled irregularly. Similar to [Liu \(1998\)](#), [Olston and Pandey \(2008\)](#) and [Ford et al. \(2008\)](#) defined recrawl policies based upon information longevity, by calculating a cost function that attempts to maximise freshness for ephemeral articles, whilst minimising the overall cost to the system. Other methods have defined recrawl strategies based on other means, such as [Pandey and Olston \(2005\)](#) and [Wolf et al. \(2002\)](#) who present more user centric approaches to page refresh policies by implementing a quantifiable measure of web page *usefulness*, in relation to user experience. Usefulness in this case was based on a number of factors such as the similarity of a web page to a user query and the number of accesses a page has within a search engine. The recrawl schedule is then adapted to favour those pages that are deemed more useful to the user experience.

2.3.2 Topical web crawling

Section [2.1](#) discussed work related to building or bootstrapping corpora based on the content of documents returned by a search engine. The work presented in this subsection discusses research related to topical/focused crawling, that filters documents collected from a crawl based on their relevance to the desired domain, or controls the direction of a crawl based on whether collected documents match the desired domain ([Medelyan et al., 2006](#); [Pham et al., 2018](#)). In topical/focused crawling the purpose of a crawled corpus can be considered as the driving force for the path it has made through the web graph. For example, PageRank, the ranking algorithm developed by Google, recursively scores and ranks web pages based on the number of other pages that link to it, where each link is referred to as a vote and each vote is weighted by each voting link's own score. The original purpose of PageRank was to provide a general means to rank search results, but ranking algorithms such as this can also be used to inform crawlers on the importance of pages

and therefore how often it should be refreshed or crawled (Page et al., 1999; Biemann et al., 2013; Pham et al., 2019). Other work conducted on link based topical/focused crawling has used more linguistic information such as the context of words that surround a Uniform Resource Locator (URL) link to some target site. Liu et al. (2015b) used a tf-idf based measure to pick important words surrounding links so as to discover the potential topic of that link. Any contexts that matched the target domain had the corresponding link returned by the crawler.

Linguistic and foreign language corpora generation is an area of research that commonly uses topical/focused crawling. For instance, one might wish to build a large, general corpus for a specific language and simply scrape any and all pages found to be written in the target language (Bauer and Gaskell, 2000). Guevara (2010) presented NoWac, a large, general web corpus consisting of documents only written in the Norwegian language. This corpus was generated by extracting all mid-frequency words from a Norwegian language dump of Wikipedia and using them as queries to a search engine. The results were then used to seed a crawl, under the assumption that the search results would yield good examples of common Norwegian language content. The crawler was limited to only those pages with the Norwegian *.no* suffix, constraining the crawl to more likely cover pages only written in Norwegian. Boilerplate extraction and language detection were then used to filter and clean the resulting corpus.

However, when creating web-based corpora one may wish to collect documents representative of a very particular topic. For example, Baroni and Bernardini (2004) and Baroni and Ueyama (2006) achieves this by extracting words and phrases from an example domain-specific corpus. Developing this work, Guevara (2010) used BootCaT search results as a set of seed URLs to a crawl. However, the issue remained that links discovered whilst crawling and subsequent content may have no relevance to the original reference material. To address this, work has been conducted in the area of topical/focused crawling, whereby a system is put in place to filter or only follow links and collect documents that fit a specific theme. For example, Medrouk et al. (2016) proposed a methodology for creating a domain-specific web-corpus by using a combination of

HTML content extraction and keyword search to generate a domain specific corpus from web collected content. Medrouk et al. (2016) first performed a web search on a chosen set of keywords and scraped the results for their content. The HTML is then parsed and potentially relevant content (in the form of blogs, comments) extracted using a Document Object Model (DOM) parsing tool for HTML. The resulting raw text was then searched for relevant keywords and any matching texts returned to the user. Medrouk et al. (2016) did not control the crawl directly but simply provided a means to filter the output of a crawl for relevant content.

Similar to this work Baroni et al. (2009) created WaCky, a very large linguistic corpus for English, German and Italian by generating search queries from random pairs of words originating from the mid-range of a word frequency distribution of some source reference corpus. For example, to generate search queries for the English language, Baroni et al. (2009) used the British National Corpus as a reference corpus (Libraries, 2007). The top 10 seed URLs are extracted from each query and fed to a crawler that is restricted to only crawl the specific language domain (e.g. *.de* for the German language corpus). Schäfer and Bildhauer (2012) developed the work of Baroni et al. (2009) by adding further components to the system to create a tool chain, such as boilerplate removal, language detection and near duplicate removal to further enhance and clean the corpora return by a crawl.

2.3.3 Domain specific corpora using web crawling

Building domain-specific corpora using focused web crawlers can generate large corpora in short periods of time, but are often dependant on a clearly defined domain in the form of an example vocabulary (i.e. training corpus). One strategy is to find a method for picking the seeds given to a crawler to ensure it searches in a location on the web likely to contain relevant content (Vieira et al., 2016). Another common strategy is to crawl the web and use classifiers to make discriminatory decisions on whether to return that content or continue crawling a certain part of the web (Remus and Biemann, 2016; Qiu et al., 2015; Ester et al., 2004). For example, Remus and Biemann (2016) produced a means to expand domain-specific

corpora using statistical N-gram language models to focus a web crawler. Figure 2.3 illustrates the architecture of the system

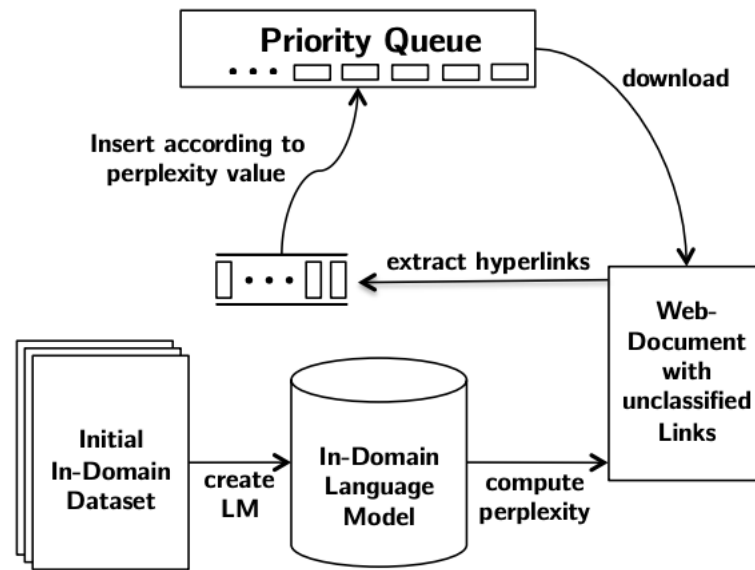


Figure 2.3: The focused crawler expansion model

In their approach [Remus and Biemann \(2016\)](#) present a 5-gram model built from the target domain and utilise this to order the priority queue of a focused web crawler. Pages discovered by the crawlers are scored for their similarity to the target domain according to their perplexity measure generated by the trained language model. Extracted links are then placed into a HIGH, MEDIUM or NORMAL queue for the crawler to follow, dictating their order of relevancy to the target domain. The result of this simple method is a crawler with a higher likelihood of navigating web pages generated from the target domain, operating under the assumption that web pages that are members of the target domain are likely to link to similar pages.

To evaluate their focused crawler, [Remus and Biemann \(2016\)](#) conducted two experiments, using a conventional un-focused crawl as a baseline. The first experiment was a focused crawl, limited to the sites found in the seed list and focused on the domain of *cats* and *technology*, using language models built from the respective categories found on Wikipedia. The seed URLs comprised of four seeds, two taken from each domain. Each crawler was limited to collecting just 100 pages. Each focused crawler was found to collect over 90 pages from their target domain, whereas the non-focused

crawler showed an equal distribution of collection from all sites. The second experiment used a focused crawl trained in the domain of education in the German language, using a target corpus from the German educational domain, to build the language model. The results showed that over 95% of the collected documents were in the German language and 92% related to education. [Remus and Biemann \(2016\)](#) proposed a simple and powerful means to build domain-specific corpora using language models, the caveat being that a sizeable corpus taken from the target domain is required as a seed.

Others have developed similar approaches, such as [Bel et al. \(2013\)](#) who developed PANACEA, a web based application for topical/focused crawling based on the Bixo web mining architecture ([Bixo, 2019](#)). PANACEA classifies documents based on the topic classification of their metadata, such as title or article. Similarly, [Medelyan et al. \(2006\)](#) developed a topic model representation of a target domain used to filter out irrelevant pages and links.

2.3.4 *Adaptive crawling*

Adaptive crawlers are those which change or adapt their crawling strategy and behaviour based on the domain relevance of links discovered. These domain-specific adaptive crawlers commonly come in the form of an augmented Form Focused Crawler (FFC), which focuses on finding online user forms ([Barbosa and Freire, 2007](#); [Hamid and Hassan, 2016](#); [Li et al., 2012](#); [Barbosa and Freire, 2005](#); [Barbosa et al., 2011](#)). For example, [Barbosa and Freire \(2007\)](#) presented Adaptive Crawler for Hidden Web Entries (ACHE), a fully automated system for finding hidden entry points to online databases. Hidden parts of the web are areas that exist and are neither indexed or present in search results. These hidden entry points are exposed via forms that are filled in by web users and has therefore precipitated the development of the FFC. [Barbosa and Freire \(2007\)](#) adapted their original FFC to discover and filter forms based on their domain relevance, allowing for the collection of hidden web content that is relevant to some target domain.

The components in [Figure 2.4](#) show the additions made to the original FFC. The domain-specific elements of ACHE are the

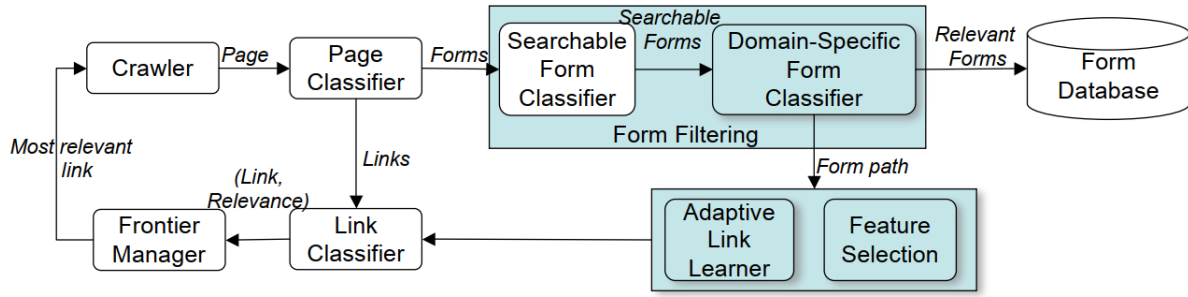


Figure 2.4: The ACHE crawler

link classifier and domain-specific form classifier. The form classifier makes discriminatory decisions based on the text content of an online form. The adaptive link learner is an unsupervised classifier that is periodically retrained on all previous links found to link to a relevant form. The training document within the adaptive link learner is the URL broken down into features along with a surround context window where the link was found. Once trained, links matching the domain are sent for form discovery and extraction. To evaluate the ACHE, domain-specific form classifiers were built for 8 domains and both fixed domain link learners used as a baseline. Barbosa and Freire (2007) showed significant improvements in harvest rate and precision of collected corpora between 34% and 584% across each domain.

The majority of the work presented in this related literature has so far focused on individual components of web-based domain-specific corpora generation, such as web search and crawling strategies. The final section of this related work presents system that combines a variety of these approaches to both discover and build domain-specific corpora.

2.4 MULTI-AGENT SYSTEMS

A Multigent System (MAS) is a software development paradigm that arose from the need of a systems owner, designer or users not being able to perform a specific task for themselves and therefore implementing a computer system to act on their behalf. For example, the field of aviation has grown from a purely mechanical system operated by one or very few human operatives to the modern approach that comprises pilots supervising and operating large

complex machinery comprising of many autonomous sub-systems acting without the need for human interaction unless an error occurs. A MAS comprises of agents, where each agent is responsible and capable of performing independent actions on behalf of the user (Spanoudakis and Moraitis, 2007). For instance, a *classifier* agent may analyse a document and make a discriminatory decision of that documents class and send that document to a number of different agents depending on the decision made, without the need for a human to supervise the process. A MAS comprises of any number of these agents that communicate with each other via messages containing information, such as post-processed documents. A MAS is designed to collectively achieve an ultimate goal through a divide and conquer strategy of sub-goals that would have otherwise been too difficult or time consuming to achieve by a human analyst. The goal of a MAS developer is to build agents that efficiently perform these tasks and achieve the intended goal (Spanoudakis and Moraitis, 2007).

The concept of MAS was born from the object-oriented software development paradigm, which conceives of a software system as a collection of discrete "objects" that individually represent a single abstract or concrete entity that holds and manipulates data that is relevant to what it represents (Weisfeld, 2013). The complete system operates through these objects, communicating and processing information between other objects to achieve sub-goals that ultimately achieve the goal of the entire system.

In the previous sections of this related literature the focus was on specific aspects of web-based information retrieval, such as feature extraction, search and web crawling that could potentially be used as single agents of experts in a wider system. One key difference in these methods is that they do not necessarily represent a continuous information processing environment that consist of autonomous agents, but more a single methodology for discovering relevant documents. The works discussed below describe systems that fit the MAS paradigm as they consist of agents that communicate and iteratively evolve into domain discovery systems.

Oliveira (2014) proposed the concept of WATES (Wild Animal Trafficking Evidence Seeker), a MAS designed to break down the

complex task of analysing posts within a social network and finding evidence of illicit trade in protected wildlife. The diagram for WATES can be seen in figure 2.5. The main contribution of this proposed system is the three expert agents: *External mediation agent*, *Fetching agent*, *word or phrase*. Each expert in WATES is a collection of smaller single purpose agents that autonomously achieve a single sub-goal within the system. WATES is a system comprised of many agents which communicate between each other and ultimately relay the results through experts to discover instances of those trafficking illegal wildlife. The example Oliveira (2014) use for an expert is The Fetching Expert, a sub-system designed to retrieve more content from the web for analysis (i.e. a web crawler). The main agents within this system comprise of a spider or web-crawler, a page-filter for filtering useful information and a page-exporter for conveying that information in a structured manner for further analysis (such as storage in a database). This would encapsulate not just web-content, but structured metadata and fields that may exist in the source content that could be of use.

Although conceptual in form, WATES has a lot of parallels with existing MAS. For instance, those tracking, discovering and analysing specific topics or events discovered on the web and social media as they occur (Aiello et al., 2013; Abel et al., 2012).

For example, Abel et al. (2012) presented Twitcident, a Twitter-based application that provides real-time tracking and definition of incidents as they occur on the platform. Twitcident operates under 3 *agents* that collectively curate information about a specific event on Twitter, in their case incidences of people provided statements about fires as they occur. The first agent profiles events and adds semantic information to the event by aggregating over Tweets and collecting attribute/value pairs such as, location, incident and persons. The second agent is a social media aggregation component that collects all related information regarding the event, such as related pictures or videos. The third agent performs named entity recognition of the Tweets for use in the system, such as the incident profiler.

Similarly, Vicente et al. (2018) developed a system for the real-time analysis of trending topics on social media consisting

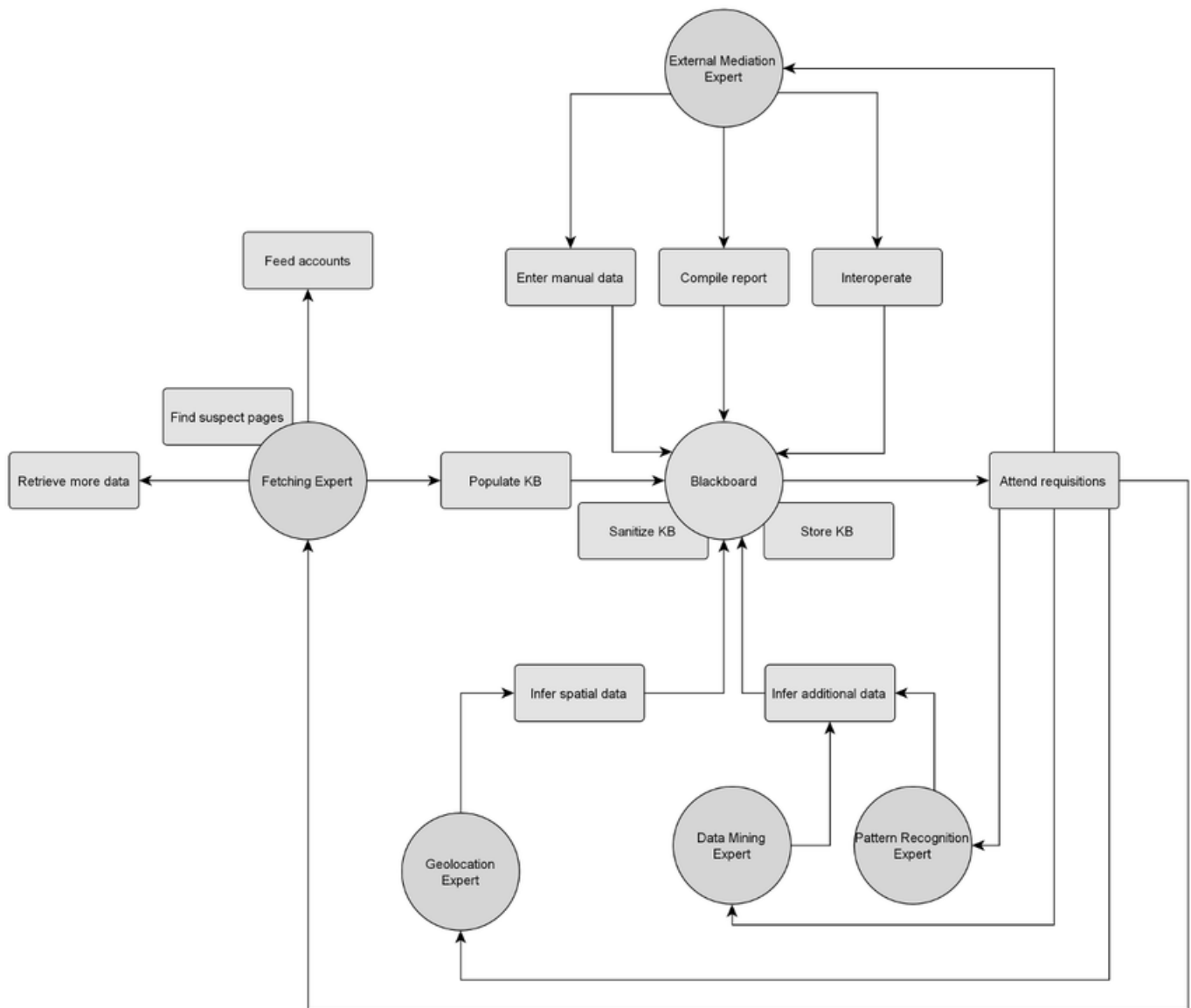


Figure 2.5: The WATES multiagent system

of 6 discrete experts: crawler, keyword discovery, topic extraction, sentiment analysis, user-profiling, IXA-pipes and a GUI. IXA-pipes is a simple system used to perform typically needed tasks in Natural Language Processing (NLP) such as, tokenisation, lemmatisation and pos-tagging (Agerri et al., 2014). In Talaia information flows from the crawler, which collects new information, to keyword discovery, topic extraction and sentiment analysis, depending on the task. The intermediary steps between the crawler and GUI acted as means to identify key information within the texts and to discern relevant documents to be presented to user.

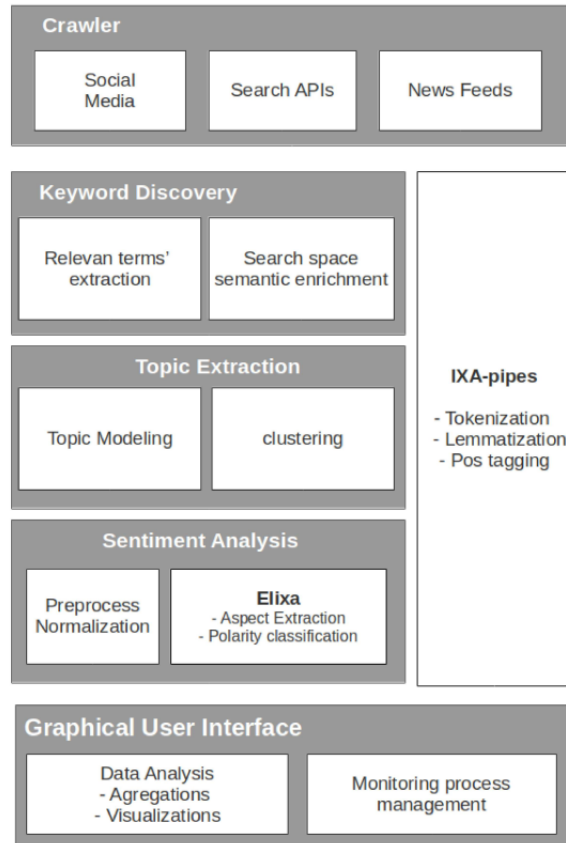


Figure 2.6: The Talaia component view

Figure 2.6 shows the component view for the Talaia system, which sees information flowing from the Crawler to Keyword Extractor, Topic Discovery and finally Sentiment Analysis agents. The final results are presented to an analyst via a bespoke UI using the Django Web Application framework, which supports data analysis and visualisation. This UI allows the user to communicate to the crawler which topics and users to follow, and evolve a system focused on a specific set of users and topics of interest.

To evaluate Talaia [Vicente et al. \(2018\)](#) focused on two case studies: tracking cultural events and citizen opinions of political parties. The first project focused on monitoring cultural shifts that occurred as result of the Donostia European Capital of Culture framework (DSS2016). This was chosen due to a number of pre-existing datasets on DSS2016 that are manually annotated in cultural polarity as either positive, negative or neutral in effect. Talaia was run on the Amazon

AWS system and collected a total of 166k relevant Tweets and press mentions.

For the Political domain, the Tweet crawler was set to collect Tweets during the 2016 Basque electoral campaign. The crawler was configured to collect Spanish or Basque language Tweets mentioning the main political parties and their respective candidates. In this instance a gold-standard was compiled by tasking human annotators with hand annotating an evaluation set of the collected dataset with their positive, neutral and negative sentiment.

To evaluate the Talaia the SVM implementation within the LIBLINEAR package was trained on the gold-standard datasets mentioned above. Across each language and case study the classifier was able to achieve between 65-70% accuracy for each sentiment classification. The conclusion of this work was that using this method of collection and classification the tracking of trending events on social media could be relatively well monitored in real-time, in spite of the authors observation of a need for a more accurate classifier.

In addition to MAS for bootstrapping domain-specific corpora, work has been conducted in building domain-specific networks across large corpora. Shang et al. (2018b) presented AutoNet a system for the mining and unsupervised construction of heterogeneous information networks (HIN) used for the further extraction of domain knowledge. AutoNet achieves this by mining entities and entity relations and constructing a network from large bodies of unstructured text, informed by pre-existing knowledge bases. AutoNet then expands these networks using new corpora by extracting and connecting pre-existing entities and constructing new HIN from the newly discovered entities. Figure 2.7 shows the process AutoNet follows.

AutoNet constructs networks in three phases. First, phrase extraction using SegPhrase and AutoPhrase which use light or distant supervision to discover common phrases in domain-specific corpora (Shang et al., 2018a; Liu et al., 2015a). Second, corpus level entity recognition is performed. Third, relation extraction and attribute discovery is performed and the network augmented with the new knowledge. Using this method, knowledge bases are

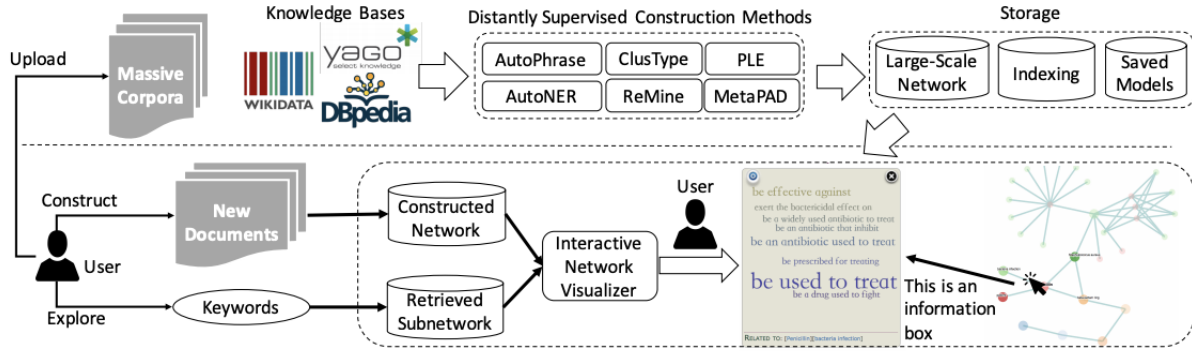


Figure 2.7: The AutoNet process

constructed and visualised creating a practical means for users to contextualise, discover and construct domain-specific corpora based on the inputs to the system. Integrating and constructing networks from corpora exposes knowledge and informs users of the domain and its connection with others, allowing for the construction and expansion of larger domain-specific corpora.

2.4.1 Domain discovery

In the majority of cases presented above, the focus has been on methods for the automatic generation of corpora, however there are some instances where the discovery of a domain and the generation of corpora is partly dependent on an analyst. For instance, in cases when the domain is still poorly conceptualised or unknown. This use case necessitates a human-in-the-loop strategy, such as that presented in this thesis. Prior to this, others have developed similar methods. [Krishnamurthy et al. \(2016\)](#) presented the Domain Discovery Tool (DDT), a tool designed to aid researchers in discovering and explore a domain by translating the needs and conceptual view of a domain to computation methods. The contributions of this work were to provide a system that supports exploratory data analysis and a means to translate the analysts actions over the data to build a computational model of the domain vocabulary. The DDT consists of six components listed below.

1. Web search
2. Focused crawling
3. Visual summary

4. Descriptive statistics - total relevant, irrelevant and neutral pages
5. Domain modelling
6. Evaluation

The key component of the DDT is its UI, which allows users to interact with and analyse the data and build a computational model of domain relevance. Figure 2.8 presents a view of the DDT taken from the original paper.

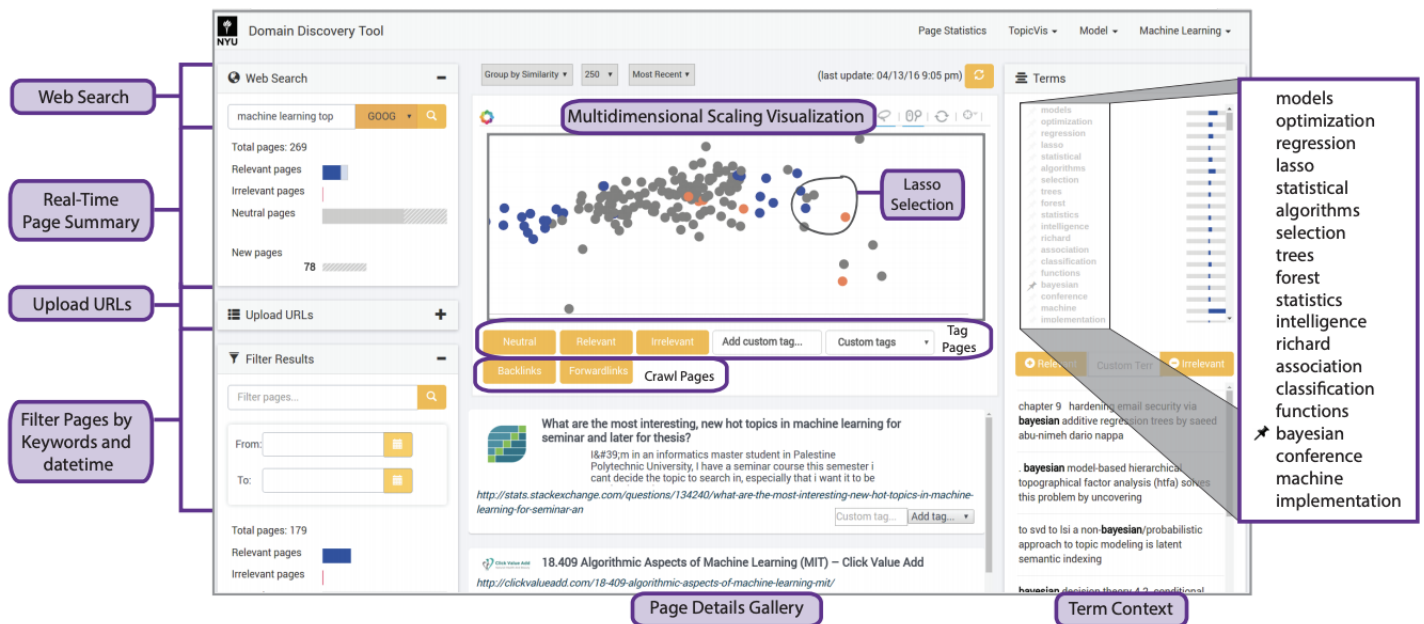


Figure 2.8: The UI for the DDT

The UI presents users with a visual summary of the results using multi-dimensional scaling to cluster documents in a two dimension plane, based on the similarity of their feature vectors. The visual summary of documents, keyword search and their context within the collected documents allow researchers to label documents as either relevant, irrelevant or neutral based on their content. These labelled documents are then used to train a computational model of relevancy which is used to guide the ACHE focused crawler, that filters links and pages based on their relevancy classification (Barbosa and Freire, 2007). As users experience the data, the notion of relevance is communicated to the crawler, creating an increasingly

specific concept of the target domain and its vocabulary. To evaluate their work, Krishnamurthy et al. (2016) used an empirical, human evaluation method which asked users to report on their experiences, using basic web search as a baseline. Users reported a generally positive experience of using the tool, specifically with regards to the document clustering component which provided a summary of the domain. The persistence of documents and domain vocabulary was also crucial, as using just web search required the use of methods such as bookmarking to save relevant documents. Using the focused crawler also showed a significantly higher yield in relevant results when compared with standard search methods.

The system proposed in this thesis attempts to develop on work such as this, through a more dynamic relationship between the analyst and system, by utilising a similar human-in-the-loop strategy. Using this methodology allows the analyst to influence and evolve the definition of relevance as they experience the discovered information. The next chapter describes the case study-based methodology this thesis employs.

METHODOLOGY

In this chapter the methodology of this research is presented over three sections. Section 3.1 discusses the various case studies that contributed to the development of the CET. Section 3.4 discusses the application-based evaluation method that was used. Section 3.5 presents Method52, the software platform that the majority of the CET was developed within.

3.1 METHODOLOGY: CASE STUDIES-BASED APPROACH

The development of the CET and associated technologies followed a case studies-based methodology. This means that the technological requirements of each case study generated a need for further development of the system's capabilities. This thesis presents five case studies over the course of three chapters that focus on one or a number of aspects that introduce the most significant contributions to the CET.

The main advantage of this methodology is that the analysis of each project's requirements and findings helped to develop a robust system. Taking inspiration from the technological requirements of each case study, and findings of the previous, which highlighted potential issues needing to be addressed.

The five case studies presented in this thesis were chosen as they best exemplify the evolution of the CET and the various types of use case which presented themselves and guided the development of its components. The CET has been utilised in a number of works that are omitted as it is not believed they help clarity or add more detail to the discussion of this thesis. For example, the CET has been successfully utilised in aiding researchers discover relevant literature in a field that is new to them. One such piece of work was conducted in collaboration with Sophie Valeix, Paul Roberts and Laura Chapmen from The University of Sussex. Here, the CET was utilised to discover grey literature found on the web relating to

the mental health and well being of students in higher education. To generate this corpus, phrase extraction was performed on a small set of related documents to generate an initial set of web search queries. Over several repeated iterations of phrase extraction and web search the original set of seed documents was expanded upon. A taxonomy was then used to divide the expanded corpus into a number of manageable sub-collections. This was achieved by representing each class in the taxonomy as a set of keywords, and placing any document containing at least one keyword for a class into that classes sub-collection. Although the above example presents a use case that is very illustrative of the CETs prime purpose; it does not provide anything new to the arguments for its efficacy or development. However, the opposite can be said for case studies, discussed below, that make up the next three chapters of this work.

The remaining chapters of this thesis present five case studies that illustrate the contributions to the final method that is the CET. These five case studies centre around developing methods for discovering relevant information online and the building and analysis of domain specific corpora. One of the key contributions each of these case studies make is that in many cases the work was breaking new ground in the domain of focus, with little or no previous attempts made to discover information from the web on that topic. The remaining subsections summarise each of these case studies and what contributions to the CET are introduced as a result.

3.1.1 *Ethical considerations and arrangements*

The ethics behind the collection, retention and analysis of potentially sensitive information when performing research must always be considered. In the various case studies presented in this thesis these ethical considerations concern the rights of individuals and the companies or organisations that own or host the content we collected. When considering the rights of individuals, thought should be given to how that person intended any content they posted to be used. For example, an individual posting sensitive information to a forum thread regarding their mental health may only considered to be viewed or relevant to those also discussing the topic on that thread. Although this example may have been posted to a public online forum, careful consideration should be given to an

individuals intentions for the information they provide. To adhere to this consideration and protect the rights of individuals, ethical approval was sought when necessary and all information collected regarding names and usernames were anonymised. It should also be noted that no individuals were profiled, and analysis took the form of large-scale thematic characterisation of a domain.

When considering the rights of organisations, care should be taken that collecting content from their sites was done so lawfully and in a considerate manner. When crawling content for this work, four precautions were taken. First, collecting content from a site lawfully involves adhering to that sites robots.txt rules and sitemap.xml regarding where and what content can be collected, and by whom¹. Second, it is also necessary to provide crawlers a meaningful user-agent when collecting from a site so that owners can identify and potentially contact you. Third, ensuring that a sufficient delay is applied to a crawler between download requests for pages on a single site prevents an inadvertent Distributed Denial of Service (DDOS) attack. Fourth, no sites were crawled that require a login to enter the site as these commonly have more restrictive, proprietary access to content.

3.1.2 Case study: Filtering crawled content

The case study presented in Chapter 4 represents a collaboration with Michael Collyer, a domain expert in the cultural and political lives of diaspora. The problem being addressed was that information relating to the external voting rights of diaspora is commonly transient due to factors such as election type and country of origin. This transient nature of information means there is often little information directly provided by governments online, which leaves diaspora and domain experts with the task of searching the web for information and discussing it in online spaces places, such as forums.

The two key problems raised by this study were first how to discover these web pages that gave evidence of diaspora voting rights, and second providing a means for analysts to aggregate the information, add metadata and provide a public database for diaspora to access the information. A method was developed that

¹ In many cases sites only allow Google to crawl their site.

used a web crawler to collect web pages, starting from a select set of seed URLs to some specified depth, that were then filtered for relevance using a classifier trained by a domain expert.

The contributions to the CET were the development of web crawling capability and the application of an active learning-based classifier to filter content. In addition, the development of a bespoke UI enabling the visualisation, search, annotation and recording of relevant information by analysts.

The resulting system came to be named the Semi-automated Web-Search and Analytics Tool (SAWSAT), but was found to be lacking in two areas. First, the ability to discover new seeds was not available and still relied on the work conducted by the analysts to discover new areas of the web to collect information. Second, a significant amount of noise was present in the text content of the crawled web pages, due primarily to the general structure most web content adopts, which includes adverts, out-links and other superfluous content, known as boilerplate. For this case study, a simple rule-based approach was implemented to extract text from the html, using just the Hyper Text Markup Language (HTML) tags commonly used by developers to encapsulate text within web pages. However, this still left a large amount of boilerplate text which limited the ability to train accurate fine-grained classifiers, resulting in a two class relevancy classifier. To address both these issues the next study developed a method to extract only desired text content from a web page, along with any associated metadata to augment this pre-existing system.

3.2 CASE STUDY: BUILDING CORPORA FROM KNOWN SOURCES

Chapter 5 presents three case studies that represent research in computational methods to discover and analyse discussions surrounding mental health, war and conflict and current issues surrounding children that occurs on online health forums. Each case study represented a collaboration between one of three organisations and members of The University of Sussex, DEMOS and the Centre for the Analysis of Social Media (CASM). The National Society for the Prevention of Cruelty to Children (NSPCC), a British organisation focused on the protection of children. The Wellcome Trust, a funding

body with a primary focus on research aiming to solve large health challenges, and The Armed Conflict Location and Event Data Project (ACLED), an organisation aimed at aggregating, analysing and mapping crisis across the globe. The main purpose for each of the three case studies was to provide a method to collect information at scale and to give an insight on these online sources for interested parties, such as health care professionals and policymakers.

In each case the location of the desired information is known, but requires the ability to continually crawl these sources for new content and extract specific content and metadata. The two main contributions to the CET introduced in this chapter are technologies designed to collect, scrape and structure web content, and the ability to continually crawl a specific site for new content. A key conceptual contribution of this work is the capability to capture metadata, which allows for the contextualisation of documents within the rest of a corpus. For instance, it was necessary to capture individual forum posts that occurred on a single web page, the position within a thread a post occurred, the date of a post and who it was authored by, for the purposes of understanding a thread and topic of conversation. This capability was required so that data could be organised and structured in a manner that allows analysts to later reconstitute conversations, understand the context of messages, visualise and explore the output.

A key functional contribution of this work is the ability to continually crawl a domain. This capability provides a method to monitor current topics as they are discussed.

Using content extraction methods such as those described in this chapter allowed for the discovery of other forums that discussed similar topics but nothing was implemented to discover more forums on the wider web. This again highlighted the need for a system that goes beyond the initial seed URLs or forums of interest.

To summarise, the SAWSAT contributed a means to crawl and discover relevant documents in order to build and annotate a corpus, whilst the forum scraping technologies, developed for analysing online forums, added a computational means to better extract desired content from discovered web pages. This web scraping capability

makes a contribution in two ways. First, the removal of boilerplate and irrelevant information from the text content of interest. Second, the development of the capability to programmatically extract and structure the text content of individual forum posts and associated metadata. The next case study summary outlines the further development of this web scraping technology by creating a UI for any user of the method52 platform to configure and deploy web scrapers more efficiently to collect only desired information.

3.3 CASE STUDY: DISCOVERING DOMAIN CONTENT FROM UNKNOWN SOURCES

Chapter 6 presents a collaboration with the Global Initiative, an organisation tasked with the monitoring and reporting of global organised crime (Initiative, 2019). The focus of this case study was to develop a method for discovering places on the web involved in the trade of illicit wildlife products. This work used three case studies to develop and refine the method of discovery and evaluate its efficacy. Each case study focused on a single product that is commonly traded within the illegal wildlife community. The three case studies investigated the illicit trade of ivory, pangolin scales and endangered orchid species respectively. Each was chosen as they represent some of the most widely trafficked animal products across the globe. The aim of this research was to provide a scalable method that allows law enforcement agencies to discover instances of illegal wildlife trade across the web. Previous work had proven to be time consuming and hard due to the sheer quantity of content on the web and the methods traders often use to obfuscate their intentions, such as code words and misinformation.

The work conducted in this project utilised the most complete iteration of the CET. The method for this project provided tools for discovering relevant words and phrases, web search to discover seed URLs and relevant content, web crawlers and active learning-based classifiers. Through the use of these tools a significant amount of information was gathered and much learned on the methods and language used by illegal traders online. In addition, this project helped refine the methodology behind the best use of the CET, with a significant focus being placed on reiterating over the phrase discovery and search stages at the start of a new case study. As this

yields a small but high quality set of documents and URLs from which to expand outwards on the web.

The next section of this methodology discusses the application based approach to evaluation this thesis takes.

3.4 APPLICATION-BASED EVALUATION METHOD

Using a case study-based methodology has meant that an application-based method of evaluation was chosen. This application-based method bases evaluations on the observed results from the CET's various applications. This application-based evaluation provides an accompanying discussion for each case study that illustrates the experience of the CET and any results drawn from the work. This method of evaluation is similar to that proposed by [Krishnamurthy et al. \(2016\)](#), when evaluating their own domain discovery and corpus generation tool, the Domain Discovery Tool (DDT). The DDT provides a bespoke UI that allows users to search the internet, extract features and view words in context via document summaries. The DDT is similar to the CET as it also follows a human-in-the-loop strategy that places the analyst at the core of the system. This allows the analyst to utilise the DDT, to explore and define a domain, whilst simultaneously building a corpus of relevant documents. To evaluate the DDT [Krishnamurthy et al. \(2016\)](#) used a human evaluation method, where users were asked to describe their experiences of using the system. As a baseline, subjects were asked to describe their experiences in comparison with constructing a corpus using just Google search.

This approach to evaluation taken by [Krishnamurthy et al. \(2016\)](#) is similar to that of this thesis, which discusses the tools, methods and results with domain experts and project collaborators as a means of evaluating performance. For example, Chapter 4 uses an agile methodology to develop a bespoke system, in collaboration with a domain expert, to help capture, annotate and present instances of users discussing the external voting rights of diaspora. As part of the agile development and evaluation process the domain expert iteratively used prototypes of the system and frequently reported on its functionality and appropriateness to the task. The resulting system, named the Semi-automated Web-Search and Analytics Tool

(SAWSAT), was proposed as an efficient and appropriate means to discover diaspora voting information.

One drawback of this application-based evaluation method is the absence of a clearly defined extrinsic or commonly applied intrinsic evaluation method. In the field of IR and more specifically the area of building corpora using online sources, the lack of any intrinsic evaluation is not uncommon and none of the examples provided in Chapter 2 propose one. This is often due to the esoteric nature of this form of research, that often proposes solutions to very specific problems. This esoteric nature results in many solutions proposing an extrinsic evaluation method, based on evaluating the performance of a proposed solution on the task for which it was designed. For example, [Pham et al. \(2019\)](#) proposed DISCO, a search and result ranking technique to discover sites containing domain relevant content. To evaluate DISCO, [Pham et al. \(2019\)](#) proposed an extrinsic evaluation method and compared DISCO's performance with a number of similar state-of-the-art solutions. The evaluation saw an SVM classifier being trained for each of the four domains chosen by the authors. DISCO and each of the baselines were tasked with retrieving 50,000 web pages for each domain, which were then classified for relevancy using the appropriate classifier. The *harvest rate* for any method was a proportion based on returned pages that are classified as relevant to the domain. Although this work bears some similarities to the work in this thesis it would be hard to replicate or compare the evaluation method [Pham et al. \(2019\)](#) chose, as it does not match closely with the method or the human-in-the-loop strategy this thesis adopts.

A solution to address any possible short-comings of this application-based means of evaluation is discussed in the future work proposed in Chapter 7. Chapter 7 discusses the implementation and integration of a number of currently existing solutions into the CET, including the search and ranking method proposed by [Pham et al. \(2019\)](#), discussed above. In addition, the project discussed in Chapter 6, which sees the CET being used to identify instances of people selling illegal wildlife products online, has secured a further three years funding. As part of the work for the second phase of this project, several of the proposed additions to the CET are to be developed. This includes but is not limited to: the

implementation of a focused crawler, improved crawl seed generation and improvements to the crawl and search ranking methods. These further developments of the CET are designed to improve the precision and yield of relevant documents and sites returned by the CET. During this period it would be prudent to implement a more rigorous evaluation framework as there now exists a set of previous results and experiences to use as a baseline. The implementation of a number of methods drawn from the literature also provides a better means of comparison.

The final section of this methodology 3.5 introduces Method52, the software environment the CET was predominantly developed within.

3.5 METHOD52

In this thesis a significant number of the technologies used in the case studies were developed on a web application based software platform named Method52, developed by a number of researchers at the University of Sussex (Wibberley et al. (2013), Wibberley et al. (2014)). Method52 is a software platform, predominantly developed in Java, designed to modularise different elements of data science and machine learning into a simple UI driven architecture and is intended as a tool for social scientists to simplify the creation and use of these methods. Through its UI, Method52 allows analysts and social scientists to quickly and easily connect individual document processing units together to form a pipeline of components through which information flows. These pipelines represent the ability of these components to extract information and perform computational analysis over large document collections. The modular architecture of Method52 allows for complex information processing systems to be built quickly and easily according to the requirements of the work. This modular MAS approach allows researchers a greater degree of freedom over the required tools, and a means to explore datasets. One significant advantage to the work presented in this thesis is that it has contributed to the continued development and refinement of the CET and is the environment in which it resides. This section discusses how this is achieved and introduces all parts of Method52 relevant to this thesis.

3.5.1 *Documents and annotations*

Data in Method52 is represented as a collection of individual documents referred to as a datum. Each datum of information is stored in a database, where each row in the database stores a single datum and each table within the database represents a single document collection. The database implementation used by Method52 is PostGres 10.X as it provides a powerful and reliable feature set, with full support for Java based applications. A datum contains a single document and any number of associated annotations, with each column within a table representing a single annotation type. Each table can be thought of as a sparse data table as it is not necessary to have all annotation fields populated for a datum. An example of a single table within a database in Method52 can be seen in Figure 3.1, which is an image of the data view available within the Method52 platform.

Figure 3.1 shows the available databases and tables on the left hand side. A selected table can be configured to show whichever annotations the user chooses. The majority of the data view in Figure 3.1 shows the data itself. This view is organised in a typical fashion to any table or spreadsheet. At the top, all configured column headings indicate the annotation contained in that field and each datum listed as rows. To generate annotations on documents, information processing pipelines are constructed inside of Jobs, which are described below.

3.5.2 *Jobs*

A Job within Method52 represents a single information processing pipeline. Once a job is started, information begins to flow from one component to another and ends when all information has been processed by each component or it is manually stopped by the user. Jobs are organised in directories and can be selected, deleted and named on the left hand side, as shown in Figure 3.2. Each Job is instantiated and configured in a work area described below.

3.5.3 *Work area*

Figure 3.2 shows that the majority of the visible space within Method52 is taken up by the work area. In this area components (described below) can be searched for and instantiated within the work space via the component menu on the right hand side. Components appear in the main work area in the centre of view. Each job has its own work area and cannot intercommunicate. In this work area, components are connected in order to communicate their inputs and outputs to other components within the pipeline, a method that generates a MAS.

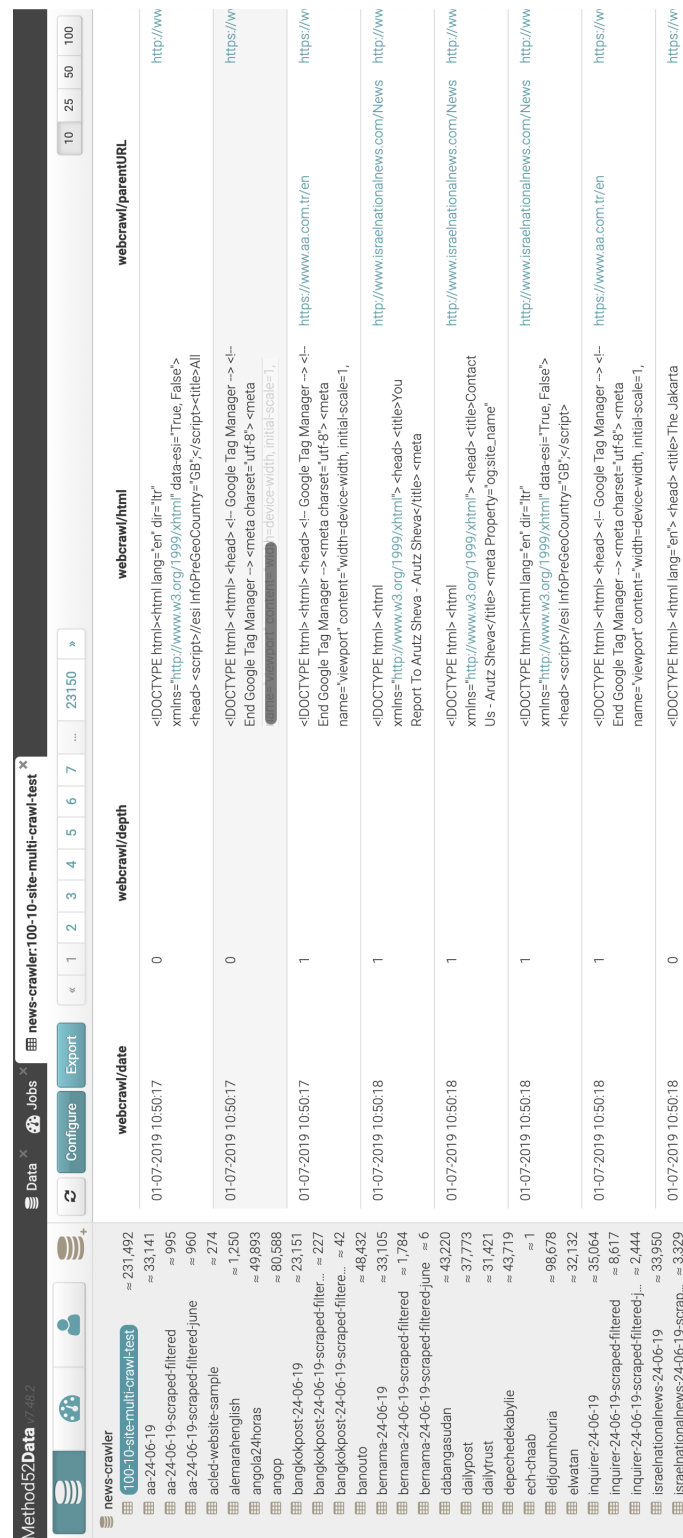


Figure 3.1: The data view within Method52

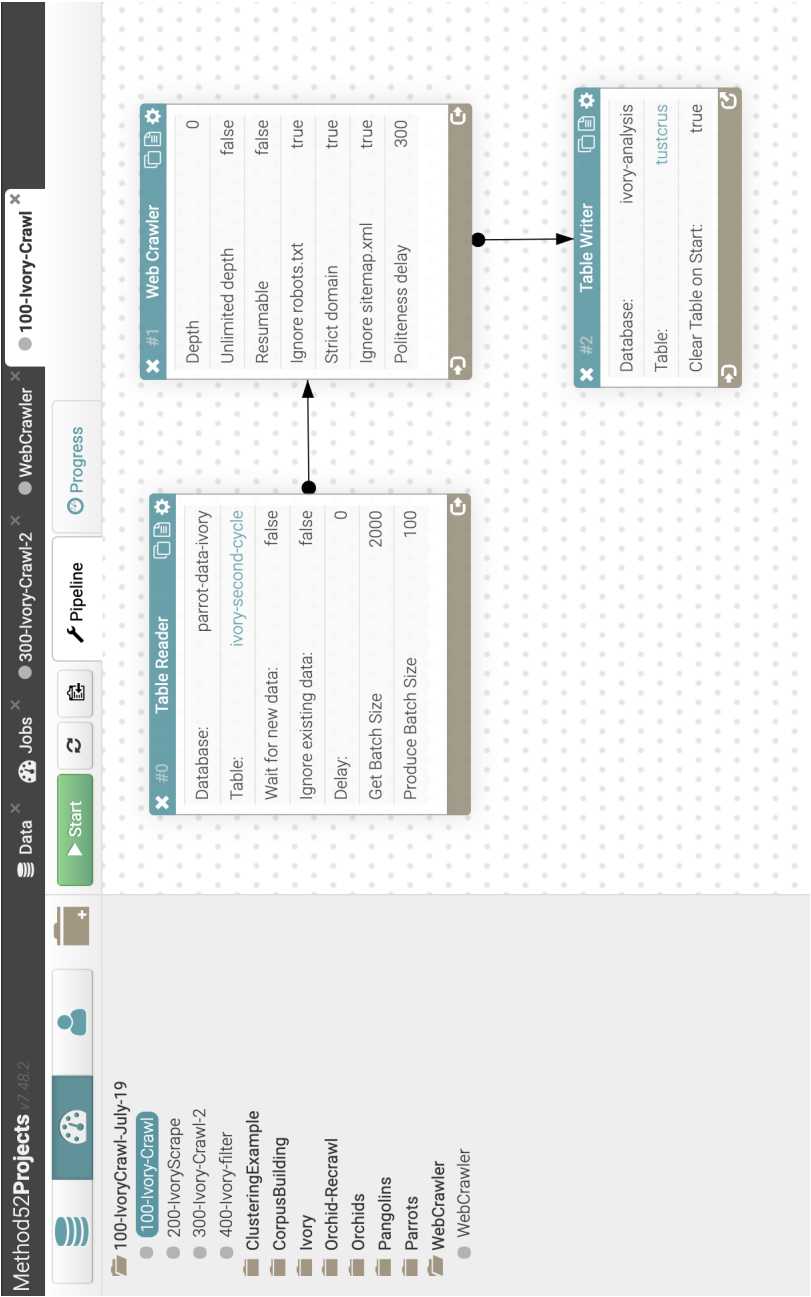


Figure 3.2: The job view within Method52

3.6 COMPONENTS

A component within Method52 is an instantiation of a single data processing software tool. To provide an illustration of a component, the example of a web crawler is used and can be seen in Figure 3.3.

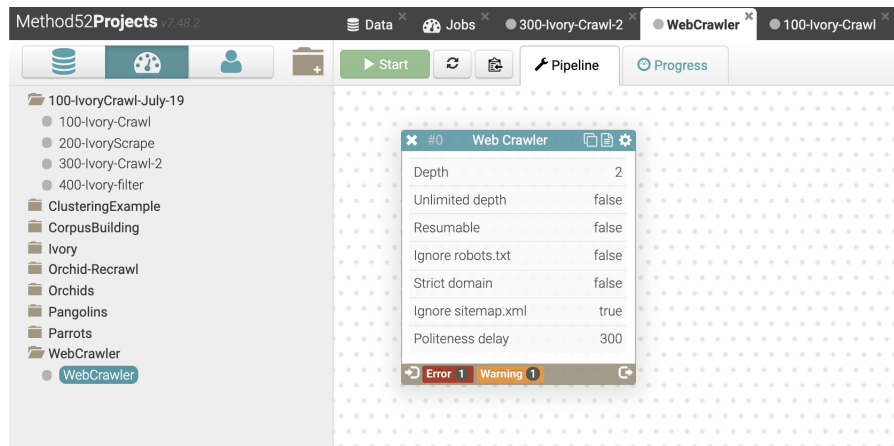


Figure 3.3: A single component instantiated within Method52

Figure 3.3 shows an example of just an instantiated Web Crawler component. Within the view of the component are a number of the configurable parameters. For example, the depth of the crawl, whether it can crawl indefinitely and whether it can be resumed once stopped. For a web crawler to begin crawling it must be provided some seeds. Figure 3.4 shows the Web Crawler component connected to a Table Reader component.

The Table Reader component provides access to a table within a database, in this case one containing URLs that can be used as seeds to a Web Crawler component. The arrow and its direction implies that once the job is run, information will flow from the Table Reader to the Web Crawler. However, this pipeline has not been configured to send its output anywhere. Figure 3.5 shows the Web Crawler outputting to a Table Writer, a component that allows other components to write output to a database table. Figure 3.5 represents a simple example of a complete pipeline, with information flowing from a table, a web crawler crawling the web using the seeds present in the table and all discovered content written to a new table. As previously mentioned, a component typically has a number of parameters that can be configured. To change this information

a components configuration view can be accessed and parameter options made available.

Figure 3.6 shows the configuration screen for the Web Crawler. The configuration screen is unique for each component, depending on its function. Each instantiation of a component can be individually configured via its parameter configuration view. In the Web Crawler example it can be seen that the crawl depth, a number of regex URL filter patterns and the politeness delay are amongst a number of parameters that can be configured by the user to suit the task. Once configured, the output of a component can be fed to another when the Job is run.

Each component of the CET is realised within Method52 as one or a collection of components that can be connected and configured into a Multigent System (MAS) for collecting and analysing online corpora. Each component of the CET is discussed over the course of the case studies in this work and the final form of the CET discussed in Chapter 6. The next chapter presents the first case study that precipitated the development of the CET, which discusses the development of the SAWSAT, a web based application for discovering the external voting rights of diaspora.

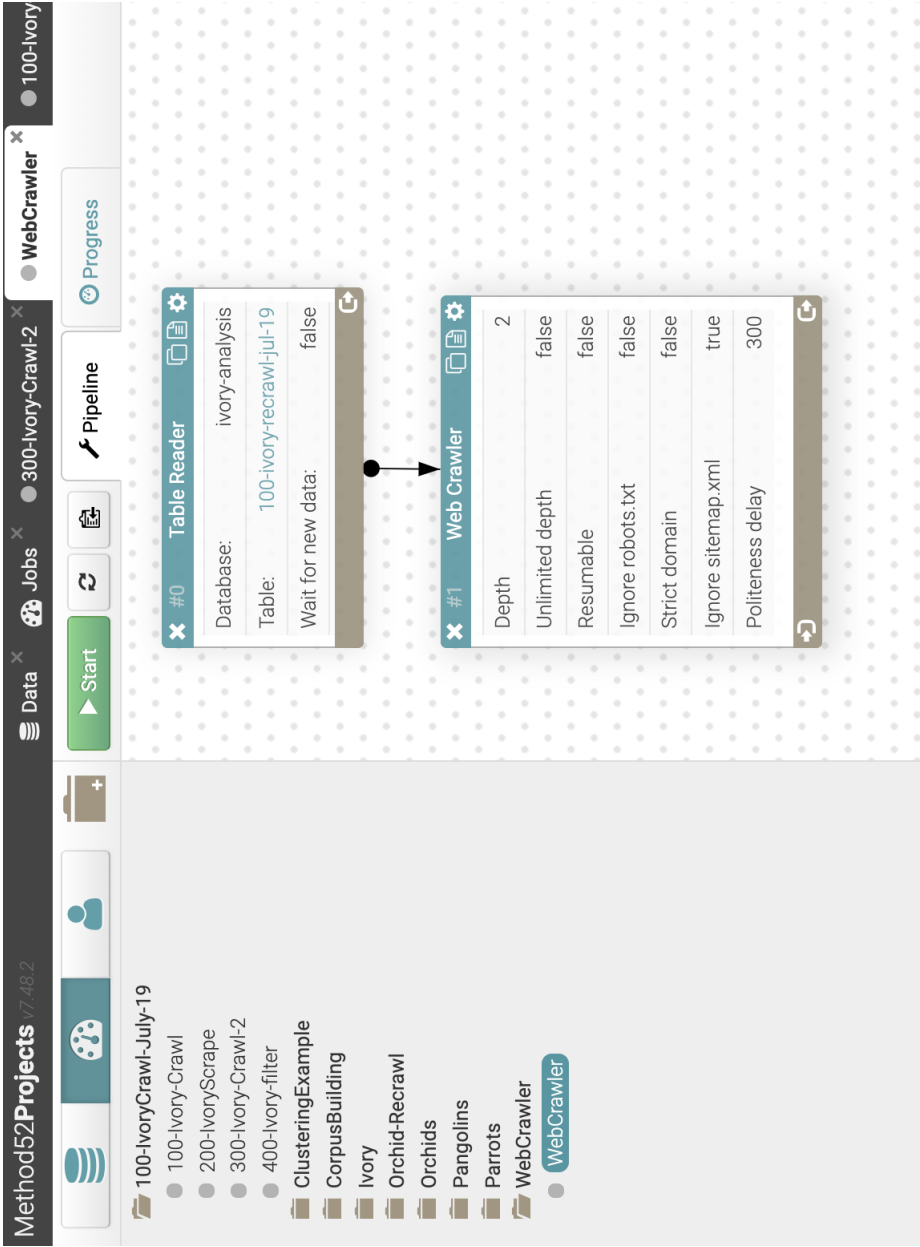


Figure 3-4: Two connected components within Method52

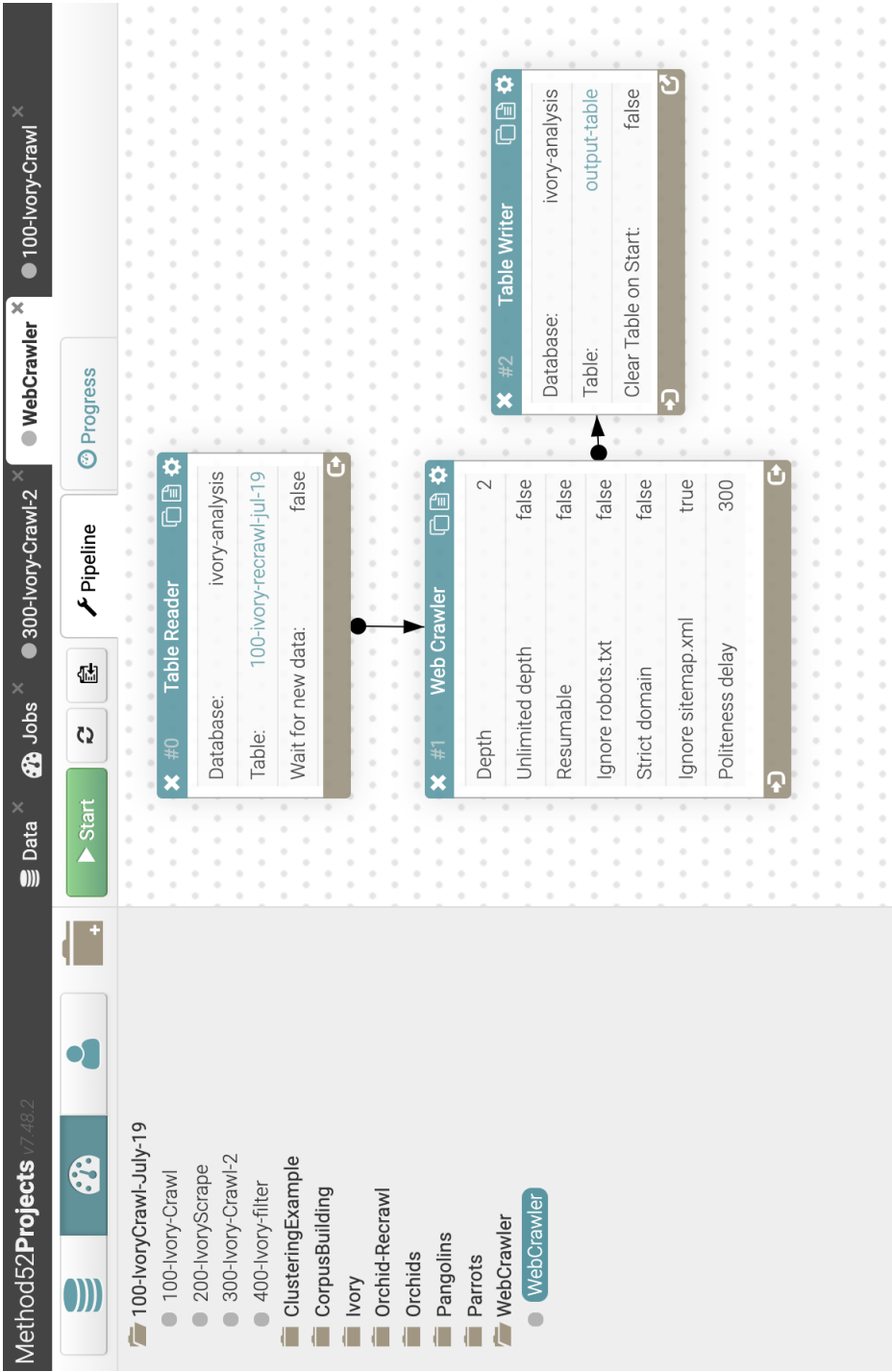


Figure 3.5: A complete pipeline within Method52

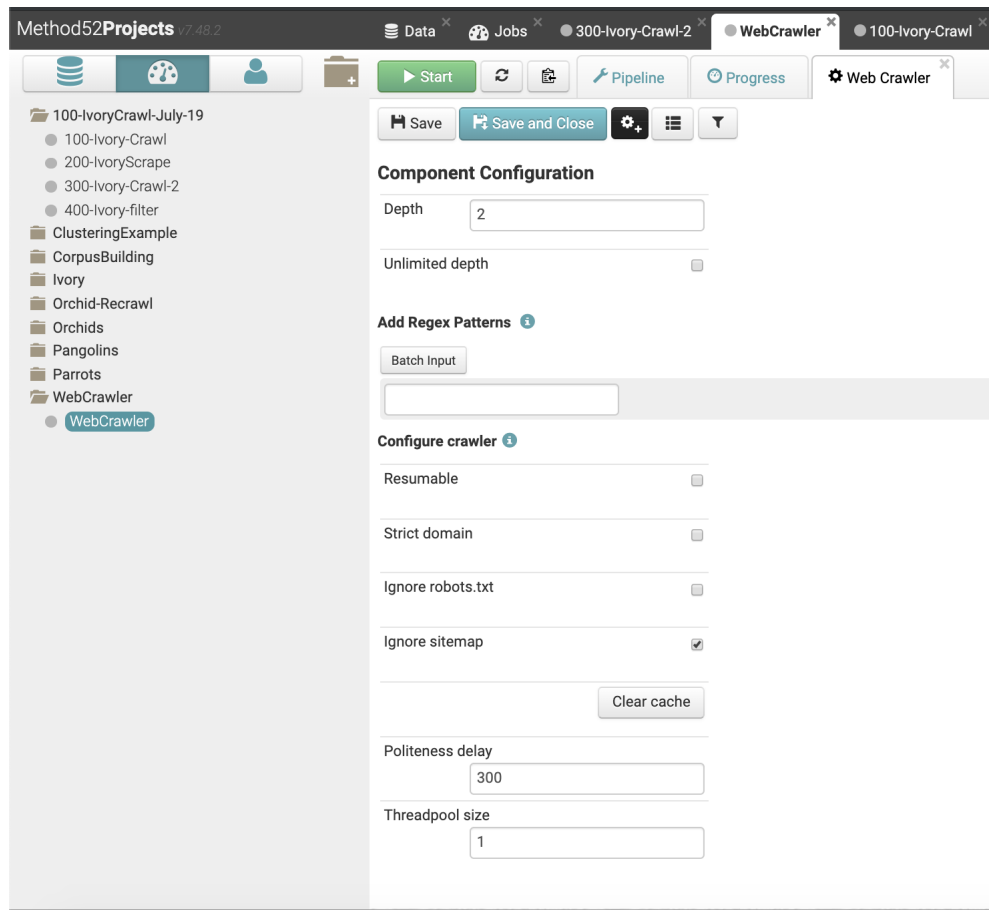


Figure 3.6: The Web Crawler component configuration view within Method52

FILTERING CRAWLED CONTENT

This chapter discusses the first case study and prototype system, the Semi-automated Web-Search and Analytics Tool (SAWSAT), that was the original iteration of the CET. This project represented a collaboration between people from the disciplines of Computer Science and Global Studies and sought to produce a semi-automated system for discovering, recording, annotating and presenting the external voting rights of diaspora in their respective home countries.

This chapter comprises of four sections. Section 4.1 outlines the problems faced by diaspora when attempting to vote externally from their country of origin. Section 4.2 outlines the problems this project sought to address. Section 4.3 presents the SAWSAT, the final method devised to aid analysts in their tasks. This chapter concludes with a discussion of what was learned and the various problems and subsequent solutions that were developed during the project.

4.1 THE EXTERNAL VOTING RIGHTS OF DIASPORA

This case study presents work conducted on creating the SAWSAT, a toolkit developed in collaboration with domain experts in the political and cultural issues surrounding diaspora, to help discover information on the external voting rights of people who have emigrated from their country of origin. In 2007 the International Institute for Democracy and Electoral Assistance (IDEA) released a handbook analysing the issues that arise for diaspora when attempting to vote from outside of their home countries. This work was partly motivated by the heavy seasonal migration that occurs between neighbouring nations who's civilians are looking for work (Reynolds et al., 2005). This mass migration represents a large number of diaspora moving countries both temporarily and permanently. This means that a large number of a countries electorate leave, sometimes only temporarily, and often want to continue participating in their home countries elections (Reynolds et al., 2005). The IDEA handbook discusses the practical and

theoretical issues that surround external voting of diaspora. One of the key issues that are raised is the difficulty diaspora have finding information about how or if they are able to vote in a given election (Reynolds et al., 2005). The reasons most relevant to this work are summarised below, with the rest of this chapter primary focused on the methodology used to develop SAWSAT, the final method that was developed and a discussion on what was learnt as a result of this project.

The IDEA handbook found that 115 states had legal provisions to allow people to vote from abroad, the most common of which were situated in Europe. There were also a total of 5 countries that had legal provisions for external voting but had not implemented any practical means of doing so. These different forms of voting and election type were often found to be subject to change and with some countries, such as Armenia, deciding to abolish external voting due to an ideological belief that people who do not reside in the country should not have a say in the political leadership of the country (Reynolds et al., 2005). The different forms of election type identified are as follows.

1. Legislative elections
2. Presidential elections
3. Referendums
4. Sub-national elections

The first two refer to elections relating to the national leadership and representation of the country. The third refers only to nationally relevant referendums and the fourth refers to anything that is sub-national, such as regional representatives. Listed below are the four possible methods that diaspora could use to vote that were identified by the IDEA handbook.

1. Personal - The voter must vote in person at a designated station.
2. Postal - The voter can register for a postal vote.
3. Proxy - Someone else can cast a vote on behalf of a person who has emigrated.
4. Electronic - There exists a means to vote over the web.

This presented a problem when attempting to centralise information regarding external voting for even those 115 countries identified in the handbook, as these provisions can be complex based on the various types of election and voting type, and as previously mentioned are often subject to change at any time. For example, Afghanistan allowed external voting for its presidential elections in 2004 but not for its legislative elections in 2005. The study found that as many as 45 countries only allowed external voting for one of its election types, and only approximately 20 countries had provision for 3 or more election types (Reynolds et al., 2005). The solution developed in this work was a software tool for researchers to discover, manage, annotate and communicate this information to diaspora on a single bespoke platform designed for the task.

This section presented the scope and background that precipitated the work conducted in this case study. The general theme of this work is that diaspora have a potentially difficult time voting in their home countries, predominantly because it is hard for them to ascertain how, and in what elections, they can participate. The next section presents a brief description of the methodology, and is followed by a section that presents the SAWSAT tool for researchers. This chapter concludes with a discussion on what was learnt over the course of the work and the identification of future work.

4.2 TRACKING DIASPORA VOTING RIGHTS ONLINE

On emigrating from their home country the external voting rights of diaspora can often take many different forms depending on the type of election and their new country of residence. Differences to voting rights can include, but are not limited to: not being able to vote once emigrating, being able to submit a postal or proxy vote or being required to temporarily return to their home country in order to vote. These differences to voting rights also vary depending on an emigrants country of origin and presents a significant problem for those tracking these voting rights as there previously existed no central database containing this information and often no official government documents detailing these rights. The existing solution consisted of maintaining a single, publicly available spreadsheet detailing the external voting rights of diaspora, for each of the 214 countries. The method for aggregating this information involved

analysts manually scouring the internet using keyword search via search engines such as Google and scrutinising sites commonly known to contain such information. For instance, forums with topics relating to diaspora and emigration where members share their knowledge and experience on the topic. This method of storage and discovery presented three problems, listed below.

- Information discovery
- Information reliability and source tracking
- Multiple source tracking

The key issue with discovering relevant information was that it is time consuming and unreliable when performed by human analysts, who do not know precisely where to look or precisely how the information would be presented. Using information processing techniques such as web crawlers and classifiers enables the processing of large quantities of web-based content, which can then be used as a tool by analysts to ensure that only potentially relevant documents are focused upon. It was also inefficient as simply using keyword-based web-searches did not ensure that all content pertaining to diaspora voting rights was captured, as all information was contingent on the keywords being an accurate representation of the language used to discuss such information. Using web-search alone is also problematic as results are biased and based upon criteria not explicitly stated by the service. As discussed in Chapter 1, results are often biased by a number of factors based on the user performing the search, including, but not limited to their location, purchasing preferences and gender, which can ultimately change the search engine's interpretation of search terms.

Information reliability and source tracking were a problem as using a spreadsheet made it hard to keep track of sources that are subject to change, such as the originating URL(s) or the actual content of a web-page which corroborates the information. Storing information in the form of a spreadsheet document meant tracking database changes over time and the reliability of information was scarce. The use of a spreadsheet also made it hard to share to a wider community, such as the general web. These inconsistencies in storage and tracking also meant that analysts were unable to say how certain they were that the information was accurate. For example, a

source found on a forum claiming what they believe is true cannot be considered as reliable as that of an official government web page. This meant that any solution would require the ability to corroborate a source, give an indication of its reliability, add metadata and be able to access the original content and its origin. These problems were addressed in the design of the UI discussed in section 4.4, that presents the SAWSAT system. The next section presents the methodology used in this project.

4.3 METHODOLOGY

The work conducted in this case study was breaking new ground in terms of technology designed specifically for the purpose of collating information about the external voting rights of diaspora. This meant there was no methodology or single piece of technology to draw inspiration and so the design process for the SAWSAT was conducted in incremental, agile stages of development. The agile methodology undertaken involved a collaboration between computer scientists and domain experts in diaspora and their voting rights. The process began with the initial hypothesis of using a combination of web crawling and machine learning based classifiers to collect and filter online content in search of relevant information. What was required was a simple and intuitive platform to curate an annotated collection on the topic.

4.3.1 *Agile method*

The agile methodology in this project involved frequent meetings between the software developer (myself) and a domain expert. Each stage of development involved a discussion of previous work conducted, additional development required and changes to the currently existing technology. Frequent working prototypes were presented to the domain expert for experimentation, in order to formulate a list of any issues needing to be addressed or desired features. Over the course of many iterations the SAWSAT evolved into a method conducive to the tracking and discovery of diaspora voting rights, during which a number of previously unforeseen features were developed as a result of this methodology.

The general method of the SAWSAT employed web crawlers to discover content and trained classifiers to filter the output similar to

work such as [Remus and Biemann \(2016\)](#), [Qiu et al. \(2015\)](#) and [Ester et al. \(2004\)](#).

The next section presents the various elements of the final SAWSAT system.

4.4 THE SEMI-AUTOMATED WEB SEARCH AND ANALYTICS TOOL

The general method employed by the SAWSAT follows a three stage process. First, to crawl the internet, starting from a set of seed URLs known to potentially contain relevant information. Second, classify all web content it found and returning only potentially relevant content to analysts. Third, for analysts to search the database, confirm relevance and annotate the content in a web-based UI. The SAWSAT is a MAS comprised of three agents: the crawler, classifier and UI. A diagram of the SAWSAT can be seen in [Figure 4.1](#).

The overall design of the system was to collect the content from all web pages discovered by a crawler agent, by taking a list of starting *seed* URLs and performing a web crawl outward from this point until some specified depth was reached. In this project a depth of 5 was used as beyond that point there was significant diminishing of return on relevant information. The web crawler implementation used in this project was the fully featured, Java based, Nutch crawler as it provided programming tools to relay captured content to the also Java based search engine Solr; which was used to structure and store the collected content. Both Nutch and Solr are open source APIs developed by the Apache Foundation. All web pages discovered by the crawler are then classified and non-relevant web pages filtered out using an active learning-based classifier, presented below. The classifier in this case was trained by an expert in the domain of diaspora and external voting rights, to identify documents that are relevant to that topic. The key information Semi-automated Web-Search and Analytics Tool (SAWSAT) attempts to collect is any evidence pertaining to the eligibility of an emigrant's right to vote externally from their country of origin, the type of voting available to them and in what forms of election. The following subsections discuss how the collected web pages were analysed by both the SAWSAT and human analysts in more detail.

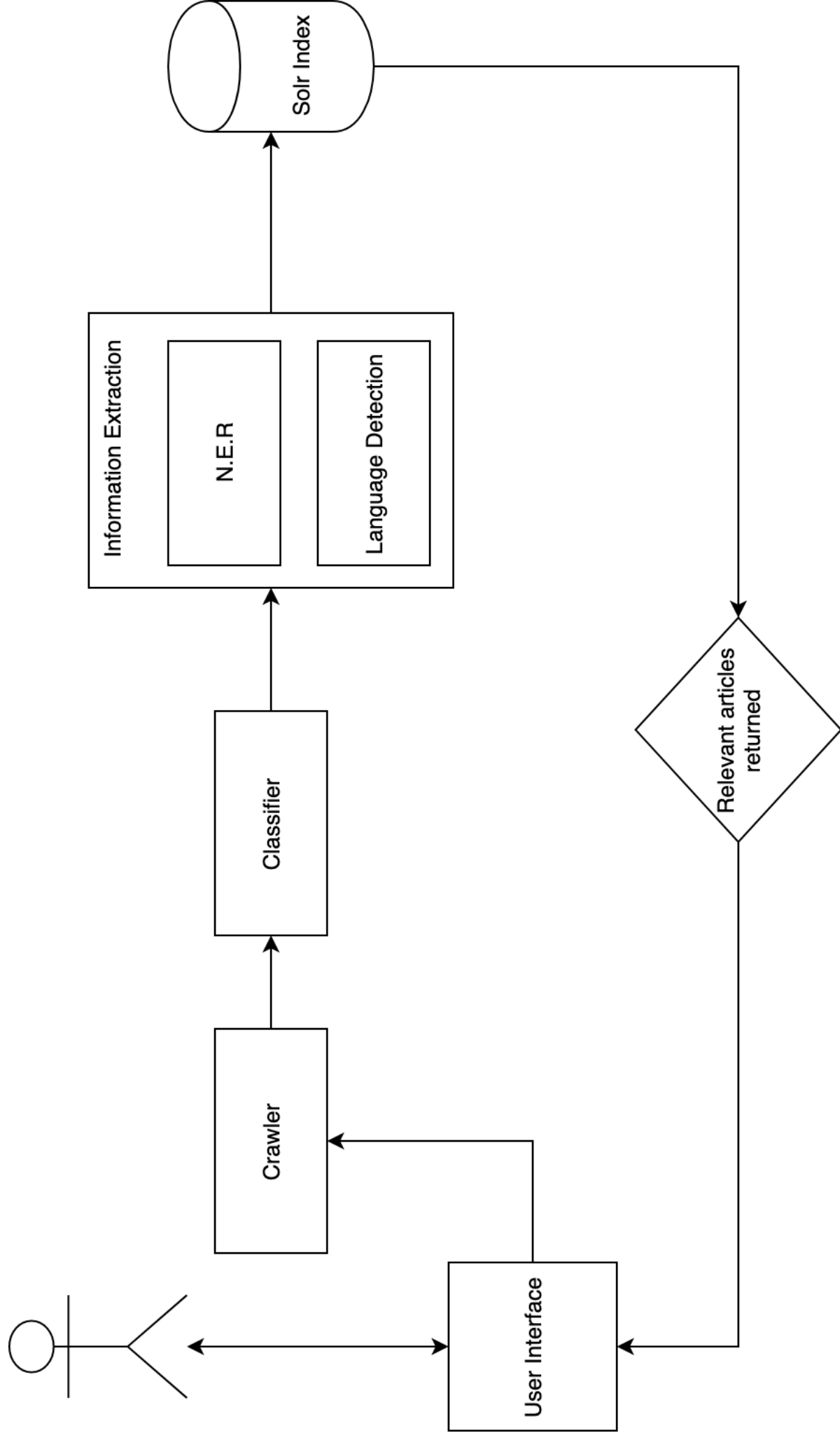


Figure 4.1: The Semi-automated Web-search and Analytics Tool

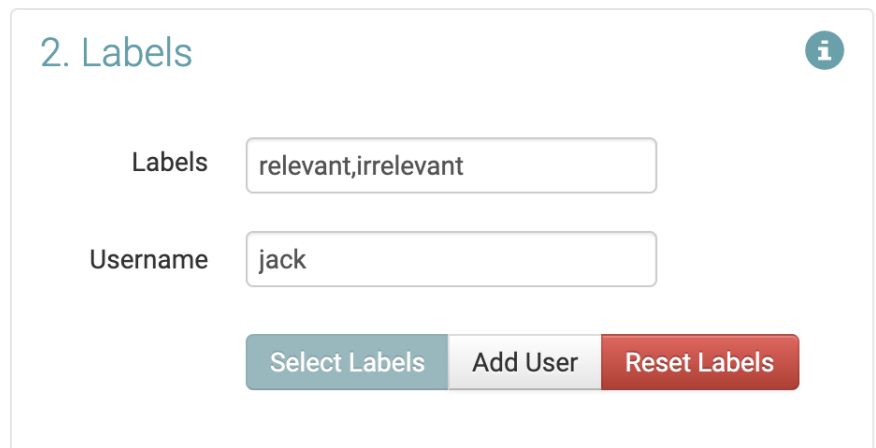
4.4.1 *Active learning based classification*

The method of classification used in this case study and the majority of those presented in this thesis was developed within Method52 by members of the TAG laboratory at The University of Sussex. This method draws heavily from the semi-supervised classifier training method originally developed by [Settles \(2011\)](#) and implemented in the DUALIST framework. This method sees users interactively training Naive Bayes based classifiers on a corpus comprised of documents from the target domain, which in this case is web pages primarily focused on issues surrounding diaspora. The key advantage to using this method of classification is that classifiers can be quickly and easily trained by domain experts with good knowledge of the target domain, but perhaps little or no knowledge of Naive Bayes or machine learning-based methods. There are four key steps to building a classifier in this manner, these are.

- Pick appropriate category labels.
- Create an evaluation gold-standard dataset.
- Train the classifier through manual labelling.
- Pick or add tokens to boost value.

Picking a number of representative category labels is crucial to the final model as these are the categories a classifier will use to make discriminatory decisions about each document it is presented. Choosing these labels is dependant on a number of factors, such as the intended purpose of the model, or what is observed in the data. For the purposes of this work a simple two class relevancy classifier was built, using the labels 'relevant' and 'irrelevant'. [Figure 4.2](#) shows the label selection screen, where users can enter any number of labels they wish.

Second, an evaluation set is randomly sampled from the dataset. The size of this dataset is chosen by the user and removed from the corpus. The evaluation set is used to evaluate the performance of the classifier in real-time during training. Each document is then manually labelled as being from one of the previously chosen category labels. In doing this, a gold-standard evaluation set is created, using documents taken from the target domain. To label



2. Labels

Labels

Username

Select Labels Add User Reset Labels

Figure 4.2: The label selection screen

documents, the text is presented to the user and the option to select one of the preselected classes is available.

Once all documents in the evaluation set are labelled, the classifier can then be trained. Training the classifier follows the same core process as building the evaluation set. The user is presented with a sample of documents taken from the corpus, which are then labelled with one of the category labels. To evaluate the performance of the classifier as it is trained, the classifier is continually classifying the evaluation set and the performance metrics calculated and presented to the user. An example of these metrics can be seen on Figure 4.3.

Label	Precision	Recall	F-Score	Accuracy	Coded	Prior Multiplier
relevant <small>Sample</small>	0.891	0.961	0.925		17	<input type="text" value="1"/>
irrelevant <small>Sample</small>	0.846	0.647	0.733		6	<input type="text" value="1"/>
Unlabelled	0.000	Features	0	0.882		sent out:0

Figure 4.3: The classifier metrics view

Here it can be seen that the labels 'relevant' and 'irrelevant' are represented in the classifier, with the individual precision, recall and f-score for each label presented with the overall accuracy. This allows domain experts to get immediate feedback on the performance of the classifier and to stop training if performance plateaus or reaches and desired level. This method also allows one to fail fast if performance is poor because the metrics are presented in real-time. This allows the user to see instantly if the classifier is working as expected, and if not they are able to start again.

In addition to training and labelling at the document level the user is able to identify or add words that are considered highly indicative of a class to boost performance. These highlighted words are given pseudo-counts within the Naive Bayes distribution over words to increase the likelihood of a class when a chosen word or phrase is present. Using this active learning-based classifier method has a number of advantages in this use case, which are discussed in Section 4.5. Once trained, the voting rights relevancy classifier was integrated into the SAWSAT. This classifier was then used to filter out documents unrelated to diaspora voting rights.

Filtered web pages classified as relevant and potentially containing information on external voting rights were presented to a domain expert via the UI discussed below. The trained classifier judged relevance based on the probability that it discussed the voting rights of diaspora, but did not necessarily need to contain all information required to make an actual assessment about a specific voting type or country. This general definition of relevance was used as the classifier was only trained to make a discriminatory decision on a two class condition, based on the general vocabulary used when discussing this topic. This general definition cast a wide net over a number of different sources found on the web, with any fine-grained analysis being left to the domain expert. The intention was to play to the strengths of the quantitative analysis of web content made by the classifier and the qualitative analysis capability of the analyst. It was also important that the system had a high degree of recall, with as little information lost as possible, which in this case was maximised by having only a general, two class definition of relevance. The corpora used to train the classifier were taken from a shallow crawl of the sites in the original seed list, which can be found in Appendix 7.1.6. These were known to have the highest likelihood of containing relevant information and covered a broad vocabulary over the English language. The SAWSAT used the trained classifier to analyse and classify all subsequent web pages discovered by the crawler and filter out irrelevant pages. These crawl and filter stages were designed to be run at the request of the analyst, via the UI, and only those pages deemed as relevant made accessible. The four components of the UI allowed the analyst to search and discover relevant information from the crawled corpora on a single platform.

ACTIVE LEARNING AND MULTINOMIAL BAYES

The classifier within Method52 is an adaptation of the active learning model DUALIST. At its core, DUALIST trains a model using standard Multinomial Bayes. The formula for classifying documents using Multinomial Bayes is given in Equation 4.1. The condition probability $P(c|d)$ is calculated as the prior probability of class c multiplied by the product of the conditional probability $P(f|c)$ for all features f within d . A normalisation constant $Z(d)$ is used, which is the summation of all classifications given to the features of d . Under Multinomial Bayes, features are represented as a multinomial distribution and so all occurrences of f within a document are considered, even if they occur multiple times. This is represented as $f(d)$, the frequency of feature f in d .

$$P(c|d) = \frac{P(c) \cdot \prod_{f \in d} P(f|c)^{f(d)}}{Z(d)} \quad (4.1)$$

As Multinomial Bayes is a supervised training method the conditional probability of a feature being given a classification $P(f|c)$ is derived from a set of pre-classified documents, in some training data D . The conditional probability $P(f|c)$ is simply the fraction of times feature f is found in documents labelled as being from class c in our training data¹.

The active learning approach adopted within Method52 draws directly from the DUALIST framework which queries the user with documents, and subsequently the features of that document, needing to be labelled. Documents are ranked and presented to users based on their posterior class entropy. In practical terms, this means the model presents documents to the user which it is most unsure of its classification. The intention is to maximise the user's ability to disambiguate the classes during training. Equation 4.2 shows how the entropy of a document is calculated, which models the likelihood of each class c in C , given document d . Entropy is maximised when the probability across all classes is closer to uniform. The documents to be labelled are presented to the user via the UI, described below.

$$H(C|d) = - \sum_{c \in C} P(c|d) \cdot \log P(c|d) \quad (4.2)$$

¹ In practice, Method52 and DUALIST additionally use a Dirichlet prior for smoothing.

SEMI-SUPERVISED LEARNING AND DUALIST

One of the most novel elements within the DUALIST framework is the ability for analysts to not only label documents with a classification, but also individual features. Features are first ranked by highest information gain, based upon how much their presence or absence reduces the classification uncertainty. All features are considered, even those which have no probabilistic label within the current model. Information again, or reduction of entropy, is computed using the formula found in 4.3.

$$IG(f) = \sum_{I_f \in \{f, \neg f\}} \sum_{c \in C} P(I_f, c) \cdot \log \frac{P(I_f, c)}{P(I_f) \cdot P(c)} \quad (4.3)$$

I_f represents the presence or absence of feature f and c is a class within the set of all classifications C . Ranked features are then presented to analysts as a set of lists, with one list of features per class. As the training of the model by the analyst progresses, each list evolves to form a set of features that best correlate with and characterise that list's associated class. The supervised learning approach of DUALIST allows analysts to select features in these lists or add their own. The DUALIST framework introduces a new configurable parameter α which assumes that selecting a feature f_i with class c_j increases the probability of $P(f_i|c_j)$ of f appearing in documents of class c . In practice, α is a pseudo-count added to the multinomial prior M_{ji} , which results in documents containing f_i having an increased likelihood of being classified as c_j .

4.4.2 User interface

The crawler and classifier within SAWSAT were designed to automate the process of discovery and address the problem that a significant amount of human resource is required to personally scour the web for relevant information. The UI of the SAWSAT provides the ability to curate a sub-collection of web pages that provide evidence of diaspora voting rights. To achieve this SAWSAT provides functionality to search and view the collected content and analyse it for pertinent information, such as the country, election type, voting type and the eligibility of specific groups of diaspora. The elements of the UI described below allow the analyst to focus on a corpus that has had a significant amount of erroneous crawled content removed by the classifier. The analyst is then able to organise and annotate a

sub-collection with specific information and present any inferences in a clear and publicly available database that displays the information to interested parties. Using a bespoke UI is common when analysing domain-specific corpora as it presents an efficient means of analysis that is specifically designed for the task. Similar examples of this can be found in the work of [Abel et al. \(2012\)](#), [Krishnamurthy et al. \(2016\)](#), [Qiu et al. \(2015\)](#) and [Sandhaus \(2008\)](#). The four components presented below are the tools made available to achieve this.

4.4.3 *Crawl*

The Crawl page of the UI can be seen in Figure 4.4 and provides the user control over when to perform a new crawl. This part of the UI also provides the option to present a fresh set of seed URLs to start a crawl. This capability was crucial as what relevant and current information was discovered by the crawler was heavily influenced by where on the web the crawl began. During the course of their work an analyst may discover a new site needing scrutiny and therefore be communicated to the crawler². Once the collected web pages are filtered for relevance by the classifier all eligible candidate pages need to be analysed by a domain expert. This functionality was provided by the search page described below.

² The complete list of sites used for the project can be found in Appendix 7.1.6

Search

Crawl

Database

Crawl

Method-51

Email

Cleanup

Blacklist

Seed list

☒

☒

☒

☒

☐

Choose file

No file chosen

crawl

Figure 4.4: The crawl page of SAWSAT

4.4.4 *Search*

An image of the search page of SAWSAT can be seen in Figure 4.5. As mentioned in subsection 4.4, the Java based search engine Solr was used in this project as it allowed for the structured storage of collected web pages, keyword and faceted search, and the addition of metadata within the fields of its index. Figure 4.5 shows the search methods available for the analyst to navigate the collection and discover relevant information based on one or a collection of keyword-searches, date-ranges, language and country-mentions within the text of a document. To achieve the language and country mentions search functionality, the Java-based Tika³ document parsing toolkit was used to identify the language of text and discover country-mentions using standard NER. This was useful to analysts needing to perform faceted searches based on their native language or country of focus. Search results are presented as a list below the search parameters, with each search result displaying a summary of the text, a link to the original source on the web and a button which navigates to the document view so that the document can be read and annotated, based on the text content of the originating web page.

³ Also developed by the Apache Foundation

Search

Crawl

Database

Search Filters

Relevant Only: ☒

Unread Only: ☒

Show Deleted: ☐

Date From:

Date To:

Country:

Australia

Language:

English-en

Query:

Submit

Search Results

4 results found. Page 1 of 1

Voting from Abroad — Skip to content. | Skip to navigation Search Site Advanced Search... English Español العربية Swahili Home Elections Today Encyclopaedia Comparative Data Electoral Materials Support and Advice Regions & Countries About ACE Personal tools Log in Practitioners' Network You are here: Home → Encyclopaedia → Topic Areas → Voting from Abroad Encyclopaedia Topic Areas Electoral Systems Electoral Management Boundary Delimitation Voting Operations Voting from Abroad Legal Framework Media and Elections Parties and Candidates Civic and Voter Education Voter Registration Electoral Integrity Results Management Systems Elections and Technology ACE On Focus On Series Measuring Electoral Quality International Election Observer

Voting in the Australian Federal Elections | USA

Voting in the Australian Federal Elections | USA USA Home Moving Housing Working Money Family Healthcare Transport Lifestyle Blogs You are here: Home > Moving to the USA > Voting > Vote: Australia Moving Country File Residency Moving Pets Relocation Checklist Leaving the USA Voting Vote: Australia Vote: Canada Vote: Ireland Vote: New Zealand Vote: South Africa Vote: United Kingdom Useful Contacts Show Worldwide AngloINFO Home Argentina Australia Azerbaijan Belgium Brazil Bulgaria Canada Colombia Costa Rica Cyprus Denmark Dubai France Germany Greece Hungary India Indonesia Ireland Italy Japan Luxembourg Malaysia Malta Mexico Monaco Netherlands New Zealand Panama Portugal Qatar Russia Saudi Arabia Singapore South Africa South Korea Spain

Ways to vote - Australian Electoral Commission

Ways to vote - Australian Electoral Commission Skip to content Contact us | Media centre | About us | Employment | Australian Electoral Commission Search the AEC website Search AEC website Search the AEC website Search Enrol to vote Enrol to vote Change address Change name Check enrolment Enrolment FAQs Special category enrolment I'm going overseas Enrolment statistics About the electoral roll Voting Voting overview Make sure you count Ways to vote How to vote At a polling place Informal voting Counting the votes Practice voting Indemnities Australian Electoral Commission Find my electoral division Current

Detail

Figure 4-5: The search page of the SAWSAT

4.4.5 Document Analysis

When navigating from the search screen to a particular document of interest in the search results the analyst is taken to the document view screen. An example image of a single document view can be seen in Figure 4.6 and Figure 4.7. This part of the UI presents the text content of a single page, so that an analyst is able to discover and keep a record of any relevant information found. In the document view, an analyst can mark the document as 'read-relevant' or 'read-irrelevant', which signals to SAWSAT whether to retain the document within the database and that it has been seen by the analyst. As previously mentioned, one issue that arose from the analysis of these documents was author confidence. Author confidence refers to how confident a reader is that the original author was accurate in their statement of facts about voter rights. Some aspects that can affect the judgment of this reliability can be the specificity of the statement, evidence given and age of the claim. To address this a *traffic light* system was implemented to allow analysts the ability encode and convey author confidence to potential consumers of the information. If an analyst found what they believed to be information about the voting rights of diaspora for a specific country, election and voting type they could annotate a document as such; along with one of three colours to indicate their confidence in the information. Red, if the information seemed unreliable, vague or potentially out-of-date, amber, if the information seemed reliable and specific but did not necessarily provide any evidence or was potentially out-of-date, and green if there was clear evidence that the facts given were true and current. This information across all countries and voting types was then summarised in the database page described below.

Search

Crawl

Database

Delete Document

Delete

<http://usa.angloinfo.com/moving/voting/vote-australia/>

Country: -- Select --

Relevant ☒

Election-Class: -- Select --

Read ☐

Voting-Type: -- Select --

Comments

Coding: High

Update Database

Submit

Country	Election-Class	Voting-Type	Coding	Delete
Australia	Presidential	Post	High	<div>Delete</div>

Figure 4.6: The document analysis page of the SAWSAT

Voting in the Australian Federal Elections | USA USA Home Moving Housing Working Money Family Healthcare Transport Lifestyle Blogs You are here: Home > Moving to the USA > Voting > Vote: Australia Moving Country File Residency Moving Pets Relocation Checklist Leaving the USA Voting Vote: Australia Vote: Canada Vote: Ireland Vote: New Zealand Vote: South Africa Vote: United Kingdom Useful Contacts Show Worldwide AngloINFO Home Argentina Australia Azerbaijan Belgium Brazil Bulgaria Canada Colombia Costa Rica Cyprus Denmark Dubai France Germany Greece Hungary India Indonesia Ireland Italy Japan Luxembourg Malaysia Malta Mexico Monaco Netherlands New Zealand Panama Portugal Qatar Russia Saudi Arabia Singapore South Africa South Korea Spain Switzerland Taiwan Thailand Turkey United Kingdom United States of America Vietnam Inside Emergency Numbers Foreign Exchange Public Holidays Translate Online USA Maps USA News Voting in the Australian Federal Elections from overseas The absentee ballot: how to register and vote from overseas in the Australian Federal Election...

An Australian citizen living abroad may register as an overseas elector and vote in Australian national elections provided that: He or she intends to return to Australia within the next six years AND Has not been out of the country for more than three years British subjects who were enrolled to vote in Australia on or before 25 January 1984 are also eligible, provided they also meet the conditions above The three year limit does not apply to children living overseas with their parents.

Expat children who have recently turned 18 and wish to enrol to vote must only declare that they intend to return to Australia within the next six years.

A voter's enrolment address can be confirmed (and other election-related information is available) on the Australian Electoral Commission website (AEC) The relevant registration form can be downloaded.

This page also explains what to do if unable to vote while overseas General enquiries on voting and enrolling to vote can be emailed to the AEC Note: It is not possible to enrol at an overseas address.

Enrolment must be within the last Australian electorate in which the person was eligible to vote.

Voting overseas Votes may be cast in person at an Australian Embassy, Consulate, or High Commission.

Figure 4.7: The document view of the SAWSAT

4.4.6 *Database*

The purpose of the database was to provide a summary of all collected information for each country and voting type discovered using SAWSAT. An image of this view can be seen in Figure 4.8. In addition to recording if diaspora could vote and how, a link back the originating document is given along with the reliability coding for each piece of information. The intention was to provide a publicly available, read-only view of the database for diaspora and a second, live version, editable by the system users/analysts.

This collection of pages provided a single platform for the use of domain experts and diaspora to track the changing landscape of external voting rights. This is similar in concept to Krishnamurthy et al. (2016) who provided a means to explore, annotate and store document metadata via its bespoke UI.

The final section of this chapter discusses both the findings of this pilot research, how this work contributed to the CET and identifies future work that led to the inception of the CET.

Search

Crawl

Database

Country	Election-Class	Voting-Type	Date (dd/mm/yyyy)	Coding	Link
Afghanistan					
Angola					
Albania					
United Arab Emirates					
Argentina					
Armenia					
Antarctica					
French Southern and Antarctic Lands					
Australia	Presidential	Post	27/07/2019	High	<div>Detail</div>
Austria					
Azerbaijan					
Burundi					
Belgium					
Benin					
Burkina Faso					
Bangladesh					
Bulgaria					
Burkina Faso					

Figure 4.8: The database view of the SAWSAT

4.5 DISCUSSION

To conclude this case study, four issues that arose as a result of this work are discussed. These helped inspire the further development of the CET discussed in later chapters. First, comes as result of the cross-disciplinary nature of the work and the tendency for domain experts to be highly specialised in their respective field, and less so in that of their collaborators. For instance, my own work consisted of developing the software system that created the SAWSAT, but with little knowledge of what was required to best implement the work. This also required the expert knowledge of those in the area of migrant issues to guide that development. It was decided that the domain experts should be the individuals who train the models used to classify documents for relevancy. It also became apparent that in spite of the UI provided by Method52 the domain experts had little or no knowledge on how best to train a classifier. This resulted in a number of failed attempts at coding documents used to train the models. It is also worth noting that for the experts to remain abreast of the work there needs to be sufficient training from those who understand the technology to help train those that do not. To address this issue, two potential solutions present themselves via the CET. It is still believed that a domain expert is the best person to use the system and drive the evolution of the corpus but they should be trained and involved in the entire process. Over the course of this study the experts were only involved in the training of models, with little or no context given for their purpose. With a single expert in control of the entire pipeline, the CET, they are much better positioned to understand the purpose of each agent in the system and better equipped to drive the creation of the final corpus.

The second issue was one of classifier performance, as it became apparent that a more sophisticated method of content extraction was required to better capture the main article(s) or content of a page, whilst removing boilerplate and any other superfluous content web pages contain. This was an issue in this project as any additional content, such as boilerplate, added significant noise to a classifier model. A weakness of training classifiers on noisy web content is that it is hard to train them to a sufficient level of precision. This results in a situation where documents can only be classified at a coarse-grained, high level in order to maximise recall. This resulted

in this project only using a simple two class relevancy classifier. The CET and Method52 are a platform that allow users to define layers of classification steps that increase in granularity until they reach a final set of document sub-collections. However, in the case of this project it was found that once a model had been successfully trained there was a significant quantity of potentially relevant information being returned by the model via the UI. The large quantity of potentially relevant information in combination with the search functionality provided by the UI made the process of searching, and coding information far easier, because all potentially relevant candidates were presented on a single platform, thus removing the need to spend vast amounts of time searching the web.

Third, the use of a web crawler allowed analysts to widen the scope of search, which was originally a small and fairly static list of URLs. Using these seeds to begin a crawl allowed the scope of search to include the wider web, whilst reducing the time spent on search by analysts. One caveat to this method was that the seeds used to perform the crawl remained fairly static over the course of the project, as the means of discovering these seeds rested solely on the domain experts knowledge. Although the system provided a means to add new seed URLs the overall quantity of seeds changed little from the original. This presented a problem as where a crawler visits on the web is heavily dependant on what seeds it is given and providing too few or not changing that list will confine the crawlers search space and potentially miss other relevant areas of the web. From the perspective of this project it meant that although there was a large amount of relevant information being returned, it was possible there was more being missed as a result of not providing a seed list that allowed those sources to be discovered. This illustrated the need for a more sophisticated method of seed generation which the CET provides as a composite part of its functionality.

Finally, the bespoke UI provided an excellent means for analysts to explore, analyse, store and annotate the information that was discovered as a result of the SAWSAT, but it was highly specialised. This extremely specialised solution made the SAWSAT hard to translate to other problems that do not specifically match the same criteria. To design a system that could generalise across many cases required a reduction of focus on a specialised UI and a means of

being able to discover relevant online content at scale with a primary focus on discovery, collection and classification.

It is worth discussing here that the agile methodology employed to develop SAWSAT resulted in a piece of software that was deemed best suited to the task by the domain expert, Michael Collyer. His original expectation was to have a simple system with which analysts to could more easily obtain information regarding the topic and a means to record that information. Instead, through an iterative process of prototyping and collaboration a complete ecosystem from which to search, annotate, record and share information was designed. Something of particular interest to Michael was that ability to not just find information but link back to the source content, annotate the documents and provide a confidence rating. Providing this ecosystem reduced the workload of respective analysts, in addition to discovering less obvious instances of people reporting information, such as mentions in forums posts. These advantages were due to three reasons. First, the crawler recovered data at a much higher rate than a human analyst. Second, the classifier picked up on these less obvious instances of information reporting. Third, the search functionality of the UI allowed analysts to analyse content from one environment, whilst providing tools to facet the search. For instance, the use of NER to discover country names allowed for a focused search for information on a specific country.

As a result of the findings and issues discussed above, the CET presents a generalised solution to discovering relevant online content at scale. The next chapter presents the results of project conducted in collaboration with the Wellcome trust, a case study intended to discover and analyse forum posts related to mental health issues, where the information extraction method of the CET was refined.

Chapter 4 presented a solution for discovering the external voting rights of diaspora online, that used a combination of web crawlers to collect data and classifiers to filter the output. This solution addressed the use case of domain experts knowing what they are looking for within a collected dataset but not necessarily where to look. The initial seed list of URLs represented a starting point to begin that search. One key problem that was identified during the project was that web pages often have a large quantity of superfluous information surrounding relevant content, such as other articles, adverts and other boilerplate. This can be seen as a significant issue because it obfuscates relevant content and introduces a considerable amount of noise when attempting to train or utilise relevancy classifiers. In this chapter a number of case studies are presented that benefit from two contributions to the CET, that allow for scraping and appropriate storage of content from complexly structured sites, such as forums. These two solutions come in the form of automated and manually configurable web scraping technologies that allow users to obtain only the information they need from a web page. Web scraping technologies compliment web crawlers by parsing web pages collected by a crawler and extracting articles or other relevant meta data for structured storage and analysis, whilst removing unwanted boilerplate.

In addition, two of these case studies present a requirement for specific sites to be continually crawled for new content, resulting in contribution that sees the implementation of an incremental web crawling strategy within the CET. One key difference between the case studies presented in this chapter and the previous is that we know where to look for the content and what we are looking for, but need the ability to collect and parse it regularly. To summarise, this chapter presents several contributions to the web crawling and scraping strategies of the CET, which are illustrated using three case studies in which they were applied.

The work presented in this chapter is organised into four sections. First, each of the three case studies are summarised to contextualise the work and identified issues to be addressed. Second, the technologies and their reasons for development are presented. Third, the results of one case study are presented to illustrate these contributions. Finally, a discussion of the findings, technologies and future work are presented to introduce the final chapter of this thesis, which sees the introduction of feature extraction and web search technologies to discover new domain specific content online.

5.0.1 *Case study: Mental health discussions in public forums*

The purpose of this research was to provide insight into the ways in which people use online medical health forums. The focus of this work was centred on the analysis of forums with topics specifically surrounding mental health, precipitated by an identified need for official health bodies, such as the NHS, to have an increased awareness of the public use of online communities and social media. More specifically this work sought to address a growing fear that inaccurate information is being shared and disseminated among members of these forums and presents the potential for harm if incorrect advice is followed. This project was originally reported in the pilot study [Smith et al. \(2017\)](#) and is summarised here.

In recent years there has been a dramatic rise in the use of online forums to discuss and share information and advice on many medical issues, including mental health. This rise can be attributed to a number of factors, such as the speed at which information can be gained across a wide range of opinions and experience garnered from online communities. Through the use of online forums, people are also able to get this information with relative anonymity by using pseudonyms and providing no personal information. These forums also allow people to seek advice without having to consult a GP, which some people may choose to avoid for a variety of reasons, such as embarrassment or lack of trust in health professionals. It has been observed that this anonymity presents a disinhibition, causing people to be more honest and forthright with their questions and requests. This project sought to provide a valuable insight into a number of these increasingly used, publicly accessible forums to better inform health professionals and members of the health industry on

how people are using these forums. This project was a pilot study into the field as previously very little work had been performed on this widely untapped, unregulated landscape of online health forums. The technology developed for this project that contributed to the existing CET framework included web scraping methods for extracting structured forum threads, posts and associated meta data. This addition to the framework provided a better means to use the pre-existing active learning based classifiers and provided more easily accessible information to be explored and analysed in the custom UI developed by members of DEMOS. The culmination of this work was a focus group at the King's Fund with a group of health professionals from the Department of Health who used the results of this work to perform an investigation on the collected data.

One key reason for generating corpora on mental health topics from forums is that the rise in their use has generated a large quantity of unregulated, publicly accessible content and communities discussing these health topics. These include the discussion and sharing of advice on topics such as drugs, treatments, experiences and opinions. This content is spread across a number of officially accredited online forums and small community run projects that are largely untapped and their exact contents unknown. The discussions found on these forums have a number of qualities that could help health professionals better understand the landscape of experience and practice in the area of mental health. The four key qualities identified by this research can be described as follows. First, the large quantity of information contained in these forums holds the potential to gain a large scale understanding of how people from a variety of backgrounds are dealing with a range of mental health issues. Second, the sense of anonymity provided by these forums means that people are potentially more likely to be open and honest in their discussions than perhaps they would be in traditional interview environments, such as GP consultations. Third, people post within these sites regularly, giving access to a large quantity of current information as people share and discuss things of concern on weekly, daily or even minute-by-minute basis. Finally, the organisation of these forums into topics and threads of conversation provide a simple means to collect document collections centred around the same theme, which are publicly available and provide a communal mentality. The advantages provided by these online communities

have the potential to learn more from the experienced world of mental health than could have been garnered from traditional methods of collecting patient information. However, the means of collecting and analysing such large and diverse datasets present their own sets of challenges that are outlined below.

First, collecting information of this type in a structured manner is difficult as most forums structure their messages using esoteric methods of configuring forum management technologies. In this project a method to perform this task was developed, that addressed the issue of esoteric web content that was originally raised in Chapter 4. In addition, the quantity of information collected and its eclectic and noisy nature, consisted of often short posts written in colloquial language. To address this problem a number of active learning-based classifiers were trained to filter the content and a custom UI¹, was used to analyse and explore the collected information during the focus group.

RESEARCH AIMS AND OBJECTIVES

In response to the advantages of collecting and analysing online forums this project consisted of two main objectives. First, to provide an insight into the themes present in online mental health forums, what information was being sought and how information was shared and discussed. Specifically, to use machine learning-based techniques, such as classifiers, to identify 'cries for help', where individuals are seeking immediate help due to a mental health related issue. Second, to develop methods to explore the possibilities of learning and computationally identifying conversations and themes surrounding Cognitive Behavioural Therapy (CBT), as this was identified as a subject of great importance for the Department of Health. To what degree these research objectives were achieved is discussed in section 5.3. The next subsection presents the second case study of this chapter, which has similar research aims, but progresses the work to one of continual monitoring of online forums.

¹ developed by Josh Smith of DEMOS

5.0.2 *Case study: Regularly collecting structured content from the Childline message boards*

The contribution to the CET presented in this chapter describes a new data collection element that is intended to collect all content from the Childline messaging forum, found at <https://www.childline.org.uk/>, and to collect all new posts made to the forum as they are posted². The reason for conducting this work was to aid the NSPCC in gathering current information on what children are discussing, and gain insight on what threats currently pervade society that could be seen as a threat to the safety of children. Similar to the mental health forum project described above, the intention of crawling the Childline forum and collecting recent posts was to analyse them for what topics affect children, by discovering what they discuss and seek advice or help for online. The main parallel with the mental health forum project is that it is believed that the relative anonymity and access to peers in possibly similar circumstances and experience, provide a means for children to seek advice and discuss issues in safety. These qualities provide the potential means to access content that is extremely current and may expose issues that require immediate reaction, such as harm to a child. The threats to children and topics of discussion or interests to children are often considered to be continually changing and evolving. For instance, the increasing acceptance and acknowledgement of the transgender community can be seen as something of extreme difficulty for a child coming to terms with their gender and work of this kind could help the NSPCC navigate and understand how younger people are dealing with these difficult subjects; which may be seen as too hard to share with parents, family or friends in the immediate vicinity.

RESEARCH AIMS AND OBJECTIVES

To conduct this work, collect forum posts and associated meta data this project had three main objectives. First, to be able to capture forums posts as individual documents, whilst retaining meta data regarding the post, for example anonymised member username and information that retains its position within the thread. Second, to continually crawl the site for new posts and only returning newly created posts for analysis by the NSPCC. How this was implemented is described in the methodology section following the final case study

² This work was conducted with the express permission of the Childline organisation

summary. Third, to perform feature and phrase extraction on the extracted content and present posts containing phrases of interest on a weekly basis to the NSPCC via a bespoke UI. The feature and phrase extraction method is discussed in more detail in Chapter 6.

5.0.3 ACLED: *Crawling and scraping news websites for reports of conflict*

The third project discussed in this chapter concerns aiding the organisation ACLED track, collect and share of information relating to instances of conflict across the globe. The purpose of this work was to provide a significant level of automatisisation to the data collection, filtering and coding processes of the organisation. In this instance, the sources of collection are known, as are the desired topical themes needing to be discovered. Similar to the NSPCC project described above, this work requires the continual, incremental collection of data from all identified sources. The main challenge of this work is the scale at which the system must perform the task; which must be able to index and continually crawl approximately 2000 sites, simultaneously.

For ACLED to maintain consistency in their output, researchers and coders are provided clear definitions of the categories and subcategory of conflict types used to annotate documents. The definitions of conflict events ACLED fall in to one of 3 major categories: violent events, demonstrations and non-violent actions.

The general method currently employed by ACLED is to task a large number of researchers assigned to observe specific regions of the world via a number of news reporting sources, including but not limited to online news websites. Typically, a large proportion of these relevant online news articles are obtained by daily scrutiny of each site in question or performing a keyword search³ on news aggregation sites, such as <https://www.lexisnexis.co.uk>. Each researcher then reads and assesses all matching articles for information relevant to one of the categories. The exact coding review process can be summarised into three processes listed below, which is taken from the ACLED coding guidebook.

³ The exact keyword query can be found in Appendix 7.1.6

1. Sourcing and review source materials
2. Collecting and inputting data
3. Cleaning and reviewing those data and resources

The method of sourcing and collecting information has been summarised above. On discovering a relevant news article, that article is coded with meta data detailing the type of conflict it describes. All discovered sources are then curated and shared online via a number of datasets and visual mapping tools. The cleaning and review process on coded data occurs on a daily and weekly basis and includes coders liaising with research managers to correct information and clarify coding decisions. This work is conducted regularly to ensure consistent definitions and understanding of all conflict types is maintained.

RESEARCH AIMS AND OBJECTIVES

This work has typically been extremely time consuming and complex work for researchers, and as ACLED's coverage both locally and globally has increased, the task has become harder to manage with their current methodology. In addition, ACLED were concerned that this method of curation missed many crucial instances of conflict due to both the lack of time researchers had available to search these sources and the collection methods used by news aggregation sites. As a result, the work conducted in this project sought to automate the task by crawling ACLEDs online news source list and collecting all current news articles which match ACLEDs keyword query. To achieve this an incremental crawler has been implemented within the CET to increase its capability to include this use case and collect all new articles relevant to this query. Within the scope of this project, an incremental crawler is restricted to crawling only a single domain, but crawls continuously, indexing all newly published content on the site. The exact implementation details of the incremental crawler can be found in section [5.2.1](#). This work is similar to both the previously discussed case studies because in order to capture all desired content from a news article a number of scraping technologies were required that are discussed in the coming sections.

5.1 METHODOLOGY

The three case studies in this chapter all share the characteristics that the location of the data to be collected is known and there is also a clear definition of what is being looked for within the collected data. The challenges shared by each of these projects is the scale of collection, the removal of boilerplate or the division of each page into sub-documents and metadata to prepare them for filtering and analysis. In the case of the NSPCC and ACLED projects an additional requirement is that these sites be continually crawled for newly occurring relevant content indefinitely. Work of this type is in keeping with others that have a particular need to extract very specific pieces of relevant content and metadata from web pages. For example, [Qiu et al. \(2015\)](#) produced a system for discovering product websites and in particular particular products types. One key contribution of this work was the implemented capability to automatically extract product specifications based on the common HTML markup features that encapsulate them ([Qiu et al., 2015](#)). These challenges have precipitated the development of the CET to include methods for achieving this, which are summarised in Figure 5.1.

Figure 5.1 shows the user interacting with Method52 to initiate one of two forms of crawl over a pre-selected set of sites. Crawled pages are then post-processed by either typical information extraction methods, or one of two web scraping components to extract the desired content from the crawled html. Once stored, documents can be pre-filtered using a number of methods⁴ such as keyword matching or date filters and the results returned via a custom UI or the data view of Method52. Figure 5.1 shows a distinct evolution of the domain specific corpus building software demonstrated by the SAWSAT in the previous Chapter 4. These additions originally occurred for two reasons. First, Chapter 4 identified that common boilerplate found in the html of web pages poses a significant problem when attempting to perform analysis, both manually by analysts or attempting to train classifiers due to a considerable amount of noise introduced by the confounding, erroneous content. Second, the case studies presented here precipitated the further development of the CET due to their own requirements and issues

⁴ Discussion of the keyword matcher and feature/phrase extractor can be found in the next chapter

to overcome. The next subsection provides a detailed discussion of the implementation of the methods needed to achieve this work and improve the capability of the CET.

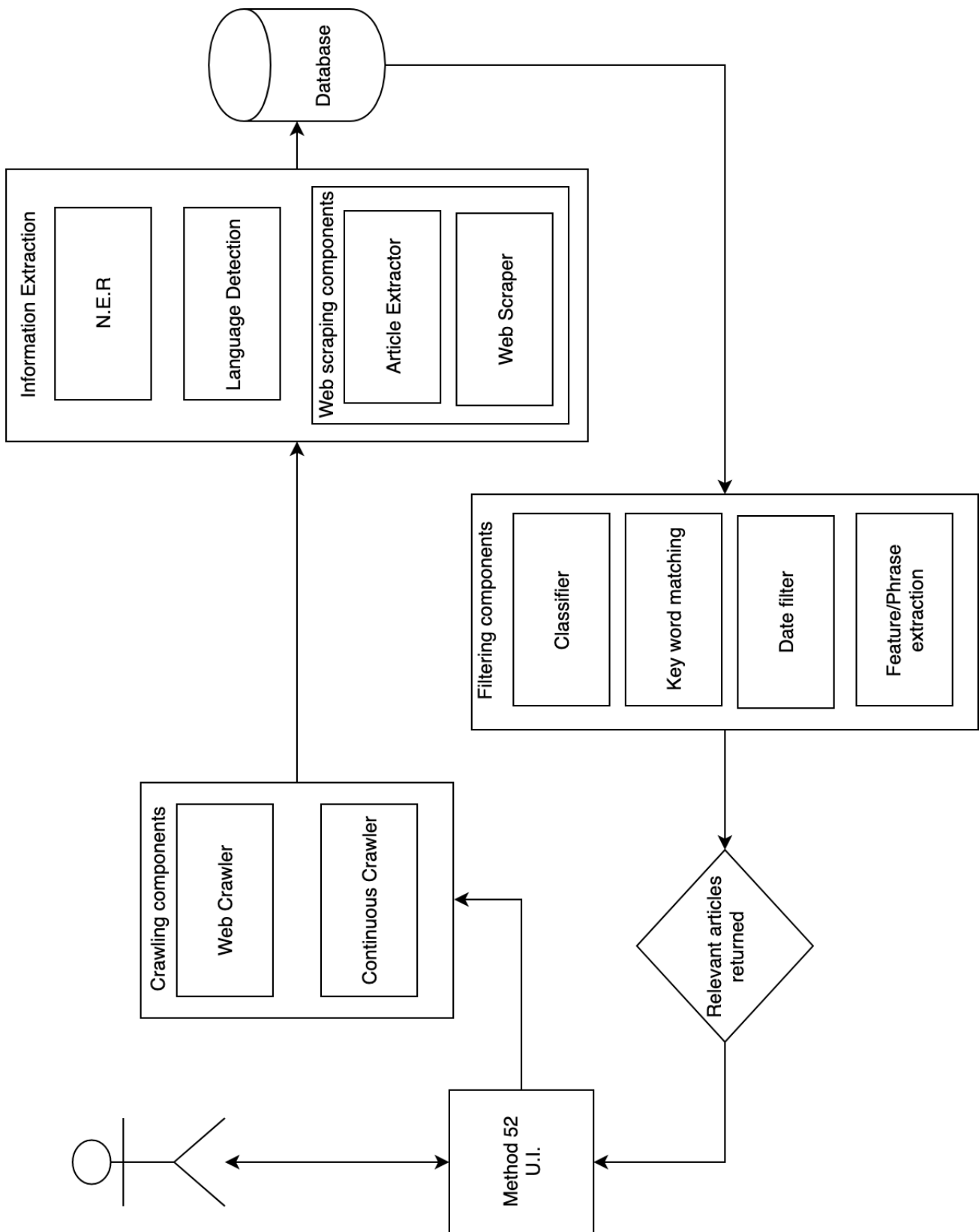


Figure 5.1: The flow diagram of the utilised sub-system of the Corpus Expansion Toolkit

5.2 WEB SCRAPING: AUTOMATIC AND CONFIGURABLE

The HTML markup of web pages is often organised heterogeneously from one site to another which makes it extremely hard to extract only the desired content to construct domain-specific corpora free from erroneous content or content that confounds the analysis. For instance, the main article found on a page taken from a news site may be relevant to the domain of our target corpus being generated but that content is often obfuscated by other site information, such as links, adverts and summaries of other articles found on the site. With HTML markup organised heterogeneously from one site to another this boilerplate cannot often be easily differentiated from the main content. This presents an issue when building corpora from the web as the user does not necessarily have any prior knowledge of what sites the crawler will visit, or what markup those sites will use to organise the content. To address this issue a number of methods for automated article extraction have been devised (Medrouk et al., 2016; Kohlschütter et al., 2010; Bakari et al., 2017). This problem is particularly prominent when building domain-specific corpora intended to be analysed or filtered using computational and machine learning methods as they are sensitive to noise and confounding content. To aid in this form of work, one main contribution to the CET was the development of an automatic content extraction software component that any and all crawled web pages could be passed through.

The underlying library used to implement this was the Java toolkit Boilerpipe that provides four robust means of extracting the main article from HTML markup (Kohlschütter, 2019). The reasons for choosing this method was that it is a fast, computationally efficient and commonly used means of extracting articles from web pages. The four extraction types provided by Boilerpipe are listed below.

- All text content
- Ruled based
- Shallow text features
- Canola full-text extractor

The first two simply extract all text content from a web page or from those HTML commonly used to demarcate text, such as the (P) tag. The most powerful and commonly used method is *Shallow text features*. This method is an implementation of boilerplate detection using shallow text features originally developed by (Kohlschütter et al., 2010). In their work, Kohlschütter et al. (2010) identified that web pages often contain adverts, out-links and other boilerplate content, which can negatively affect search performance. In their work they also observed that at the time some of the most powerful boilerplate detection methods were computationally expensive, as they evaluated content based on the topic or semantic relevance of the text, meaning that the solution was also domain dependant. In their work Kohlschütter et al. (2010) developed a domain independent means to identify the main content of a page on shallow text cues. What this means is to simplify the process of html markup analysis and avoid over-fitting to a specific site or domain by only assessing certain html tags and text features. To differentiate text blocks Kohlschütter et al. (2010) only used headline tags (H1, H2, H3, H4, H5, H6), paragraph tag (P), division tag (DIV) and anchor text tag (A) when analysing text blocks on a page. As this method of detection is domain independent and only assesses for boilerplate at a functional level Kohlschütter et al. (2010), evaluations are made at a higher level than the language or token level. Using Quantitative Linguistics a text block is only quantified by its average word length, average sentence length and absolute number of words. Using these simple features Kohlschütter et al. (2010) established that text blocks on a web page can be organised into one of two classes of long and short text with the former indicating the main article of a page. As previously mentioned, this presents a simple, topic independent and efficient method of extracting the main textual content of a page. The speed and simplicity of this method made it practical for its implementation within the CET, which potentially needs to extract content from millions of web pages from anywhere on the web. Finally, the Canola option is an implementation of a full-text extractor trained on the KrdWrd dataset, a gold standard dataset consisting of web pages that have their main content identified by hand.

The simplicity of this library and range of options made it a practical choice to implement within the CET, as people could easily

instantiate the software and experiment with the four options. An image of the configuration options for the final software component can be seen in Figure 5.2. Adding this software component to the CET allows users to place an automated Article Extractor at some stage between a Web Crawler and Table Writer, and extract the main article content from a page. This significantly reduces the noise added by boilerplate when training classifiers and reduces false positives when performing other forms of filtering, such as keyword search.

One significant drawback of this method is that it only identifies a single instance of a *large* body of text, such as an article, but does not have the ability to discover more nuanced information or metadata which may be relevant such as article titles, published dates and authors. To present a computationally efficient and practical solution to this problem, the Web Scraper was developed.

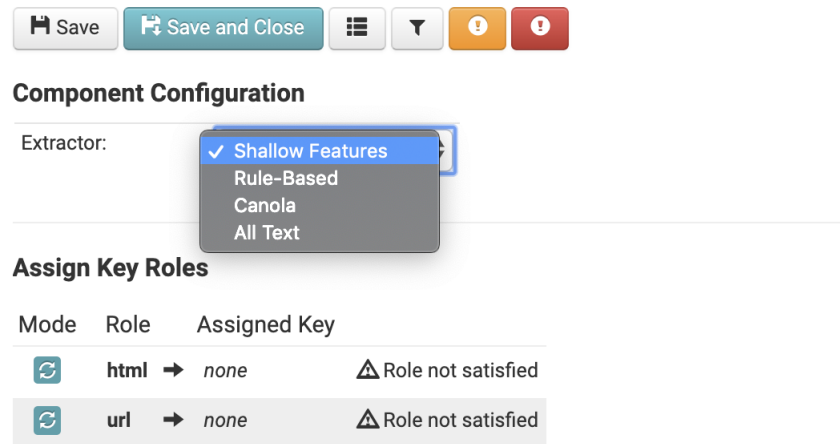


Figure 5.2: The configurable options for the Automatic Extractor component

MANUAL SCRAPING TOOL

The Article Extractor described above provides a general, domain independent solution to extracting the main article content from a web page, that typically excludes most, if not all confounding boilerplate. This was originally implemented with the intention of extracting the main content from web pages such as those found in the online news sources provided by ACLED. However, this method was not sufficient to extract complex structured information from sites, such as forums. In the case of the mental health and

Childline forums, a solution was implemented to allow bespoke web scrapers to be configured for any desired forum or structured website. This new contribution to the CET came in the form of a Web Scraper software component, an example of which is given below. The underlying programming API to implement this solution was the Java-based markup parsing toolkit JSoup (Hedley, 2019). In Figure 5.3 an example of a news article taken from the BBC news website can be seen with its html markup directly adjacent to it (taken from Bbc (2019)). The use case for this component is when the site of interest is known and one wishes to crawl and extract a large quantity its content, but also retain a large quantity of metadata in a structured manner. For instance, the mental health forum and NSPCC projects required all forum posts and associated metadata to be stored as separate sub-documents.

Figure 5.3 shows three coloured circles on each image, where the colour of each circle represents a single part of the page and its respective markup in the html. Within these two images it can be seen that within the blue circle the article title is captured and encapsulated within the headline HTML tag (H1). The main body of text, surrounded by a green circle, is show to be encapsulated within the paragraph (P) tags and the published date is contained within the html (DIV) tag in the red circle. From this example it can be seen that capturing the article text, title and published date as separate pieces of metadata surrounding the article may be useful when it comes to analysis. The page also contains a large amount of unwanted 'boilerplate', such as links to other areas of the site, adverts and summaries of other articles.

The Web Scraper software component was implemented to provide an efficient and simple solution that would allow fine-grained parsing of web pages. The Web Scraper component allows a user to instantiate a new instance and configure a bespoke web scraper for a specific site in a short space of time. Figure 5.4 shows the configuration screen, which shows four configured metadata instances: root, article, title and date. The 'root' tag definition is a special case that defines the root HTML tag-set for each sub-document within a web page, assuming that all sub-documents have the same repeated pattern of HTML markup. For instance, the root tag-set would differentiate the surrounding

html markup for a single forum post within a page containing multiple posts. The root tag-set is where all proceeding tag-sets are searched and must always be configured by the user, even if there is only one sub-document within a page. All proceeding tag sets within a page can be configured and named arbitrarily by the user, where the naming convention used is translated to annotations within the documents of the underlying database table. Figure 5.4 shows a Web Scraper configured to extract the title, article and published date from the news article displayed in Figure 5.3. Comparing the two images it can be seen that the main content of the entire article is defined in the root tag-set, which is contained in a dividing (DIV) HTML tag, with the HTML (class) 'story-body'.

The general configuration of a Web Scraper is achieved by a user scrutinising the HTML markup of a sample of the required pages on a site, finding the root and subsequent metadata tag-sets and communicating them to the Web Scraper software tool within Method52. Any number of tags can be configured for a given piece of metadata, resulting in a tag-set which is recursively searched within the root tag-set of a page. To ensure the Web Scraper is correctly configured a test suite is available to the user in the same configuration screen which can be seen in Figure 5.5. To test the configured scraper, a user enters the URL of a page that matches the configured tag-sets and selects the 'Scrape Page' button. The Web Scraper component then retrieves the page and attempts to scrape the content using the configured tag-sets, displaying the output in the 'Output' window below. In Figure 5.5 it can be seen that date, title and article of the page were successfully scraped. If this were not the case a message explaining the failure for a particular piece of metadata is displayed, indicating that the user must attempt to correct the configuration. The Web Scraper represents a simple and practical solution to full site scraping, with the ability to retain almost any amount of information from a page. To simplify the configuration process, the main options available to define a single tag definition are the HTML tag, which is required, a class or attribute name. If a more complex query is required, the user also has the option to construct a custom query which conforms to the JSoup *select* syntax.

The above automatic and configurable web scraping solutions represent a significant contribution to the CET. These components provide a considerable reduction in noise to the collected documents, whilst also providing the ability to collect from sources such as forums, which contain large quantities of structured sub-documents. The final contribution to the CET is the implementation of an continuous crawling strategy.



Figure 5.3: An example web page and corresponding html markup

Save

Save and Close

Y

!

!

Add HTML tag sets

Field Name:

field.name

name

+

Add Field Definitions:

root

➤

root

TAG = div -> CLASS = story-body

x

Enter a new tag set:

div

+

story-body

attribute name

custom jsoup query

field.name

TAG =

html ta

class n

attribu

custom

Testing Area

URL:

https://www.....

Scrape Page

Figure 5.4: Web Scraper component configuration within Method52

Testing Area ⓘ

URL:

https://www.bbc.co.uk/news/bus

Scrape Page

Output:

field.name/date
26 July 2019

field.name/title
'Shambles' as Sports Direct's results delayed

field.name/article
Media playback is unsupported on your device
Media captionMike Ashley's High Street empire
There is confusion surrounding the release of results from Mike Ashley's Sports Direct, after the firm failed to publish them throughout Friday. In a statement, Sports Direct said it was "still finalising" the results, and would give a further update at 16:00. However, that deadline has now passed and the firm's share price has closed down 3.9%. It is extremely unusual for results to be delayed i

Figure 5.5: Web Scraper component test area within Method52

5.2.1 Continuous web crawling

In both the ACLED and NSPCC projects there is a requirement to have each identified site be continuously crawled for new content. What constitutes new content for the NSPCC is any page that contains a new forum post and in the case of ACLED it will be any newly published news article in the provided online sources. To create an initial epoch from which to begin continuously crawling a specific site, a crawl of the entire site is conducted and an index of all visited pages created. This initial epoch provides the crawler with a starting index from which to discover new content. Any subsequent crawls that discover pages not in the index are typically considered to be new to the site and therefore of potential interest to the analyst. A second alternative that can identify new content is to use the Web Scraper component to extract the publication date and to filter pages by their age. The data flow diagram in Figure 5.6 shows the process which is followed for all subsequent epochs of the crawler. The exact definition of each step is discussed below.

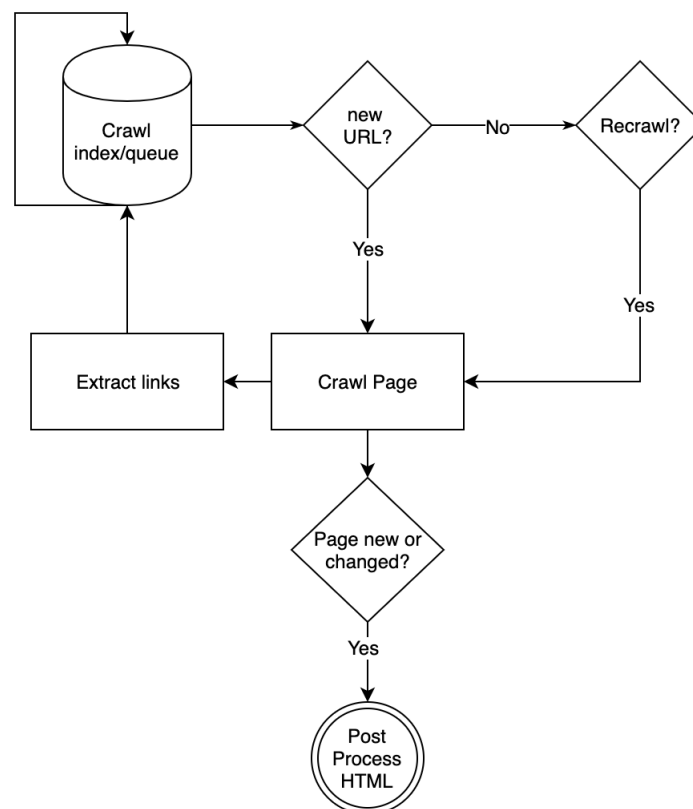


Figure 5.6: Flow diagram of the continuous crawler

In Figure 5.6 the process illustrated is continuous. Once all URLs in the queue have been assessed for crawling, the process begins again from the top of the index and the process repeated. If a URL is discovered that was not previously present in the index it is automatically crawled, extracted links added to the queue and the HTML markup at that address downloaded and sent for post-processing. Post-processing in this case refers to the html document being sent to a scraper and stored in a database table in Method52, to then be analysed within the CET. For instance, in the case of the ACLED project the article, title and published date are extracted and if the published date implies a new article and a word matching ACLED's keyword query appears in the extracted content, then the document is retained and sent for processing by an analyst.

In the case where a URL is already present within the crawl index it is only crawled if it passes one of two criteria. First, if sufficient time has passed between its last estimated scheduled recrawl date. The recrawl date is calculated using the recrawl strategy which is discussed below. Second, if the page is to be recrawled the HTML is downloaded and analysed for changes and only passed for post-processing if the main content of page has changed. How this assessment of change is estimated will now be described.

ASSESSMENT OF CHANGE IN A PAGE

Within a typical crawler implementation the content of a web page at time t is passed through a hash function to generate a numeric value representing the content found on that page. On subsequent crawls of the page a new hash value is generated and compared with that of the value generated at time $t - 1$. If the values do not exactly match it is assumed that the page has changed since the last time it was crawled and therefore is considered to contain new content and sent for post-processing. If the hash values at t and $t - 1$ match then the page is considered to not have changed and is therefore discarded and not sent for post-processing. One issue that arises is that due to dynamic content, such as adverts, and fluctuating boilerplate being inserted, page hash values will in some cases always be different at all values for t . To solve this problem a solution was implemented to generate hash values only on the main content of a page, so that the crawl by crawl hash value comparisons are only made on the main text content of a page. To achieve this,

the shallow-text-features article extractor was implemented within the crawler as a contribution to the CET. This means that when a page is downloaded its main content is extracted using Boilerpipe's shallow text features implementation, and a hash value generated only on that content, whilst ignoring the removed boilerplate. This significantly reduces the possibility of duplicates appearing in the post-processing output because assessments of change are now only based upon the main text content of an article, which typically only changes when the author updates that content. For instance, if a change occurs in a news article and its corresponding web page is updated, we wish to capture that change but ignore subsequent crawls where the article remains the same but the boilerplate perhaps changes. Whether a page is recrawled is based upon an estimated recrawl date generated using the implemented recrawl strategy described below.

RECRAWL STRATEGY

Recrawl strategies are methods of estimating when a page should be recrawled. More specifically, a typical recrawl strategy creates an estimate on the likelihood of a page changing from one crawl to the next or as in the case of this implementation the rate of change. The rate of change λ in this instance is an absolute value representing how many complete iterations over the crawl index will pass before a specific web page is likely to have changed. To calculate and use this rate of change the recrawl strategy originally developed by [Cho and Garcia-Molina \(2003a\)](#) was implemented as a contribution to the continuous crawler of the CET. In their work [Cho and Garcia-Molina \(2003a\)](#) calculate a practical, domain and problem independent rate of change that could be easily implemented and is both computationally efficient and requires little storage overhead. In their work [Cho and Garcia-Molina \(2003a\)](#) present an equation for calculating the rate of change based on the ongoing statistical analysis of individual pages. The method for calculating the absolute rate of change according to [Cho and Garcia-Molina \(2003a\)](#) can be seen in equation 5.1 and is described below. This particular implementation was chosen due to these computational and storage efficiencies as many other common solutions require the training of an estimator or classifier.

$$\lambda = -\log\left(\frac{z - y + 0.5}{y + 0.5}\right) \quad (5.1)$$

The above equation 5.1 shows the absolute rate of change λ as the log ratio of the number of times a page does not change and the total number of times a page is checked for change. Here, z is the count of times a page has been checked and y is the number of times a page has been checked and found to have changed. This rate of change is recalculated each time a page is downloaded and its hash value at $t - 1$ and t compared. The estimated rate of change for a specific page provides an absolute value representing the number of complete iterations over the entire crawl index needing to occur before that page should be re-downloaded. The constant value 0.5 is added as Cho and Garcia-Molina (2003a) found that this constant hyper-parameter value yielded estimates closest to the actual rate of change for the average page found on the web. To implement and use this metric within the CET, four implementation decisions were made.

To implement a practical use of this rate of change estimation within the CET a number of statistics need to be added to the original crawl index. First, the number of times a page is downloaded is stored for each page, this is the value for z in equation 5.1. Second, the number of times a page is checked and found to have changed is stored, which represents y in equation 5.1. Third, the average amount of time a single crawl c of the index is maintained. The average crawl time is used to calculate an exact date when the next crawl of a page should occur. The estimated amount of time that should pass between crawls of a specific page is calculated as λc , which is added to the current date and time the page is downloaded to project a future date at which the page should be checked.

Considering the size of the crawls that are anticipated within the projects of this chapter and the typical use of the CET that would see possibly a few thousand to a few hundred thousand web pages per crawl index, a burn-in period of 20 crawls was built into the system. This means that a page within a crawl index is required to have been downloaded and checked for change 20 times and associated statistics stored, prior to any change rate and projected re-crawl date being estimated.

This concludes the implementation details of the Continuous Crawler within the CET and below a description and illustration of how a user is able to make use of it is discussed.

CONTINUOUS CRAWLER IN METHOD52

The Continuous Crawler component was implemented as a software tool for users to quickly instantiate within Method52 and the CET. Below in Figure 5.7 the configuration screen for the Continuous Crawler component can be seen.

Figure 5.7: Configuration screen of the continuous crawler

Figure 5.7 shows that a list of URLs can be entered that the crawler will continuously crawl. This is limited to a small number of URLs as it is a software component intended to comprehensively crawl one or very small number of specific sites, as opposed to a general crawl of the web from some potentially large seed list. For example, Figure 5.7 shows a starting URL <https://www.childline.org.uk/get-support/message-boards/>, showing that the crawler will never crawl outside the [childline.org.uk](https://www.childline.org.uk/) domain. The second parameter allows users to enter a number of URL filters, specified in regex. This parameter constrains the crawler to only collect pages from specific areas of the site that match these filters. For instance, Figure 5.7 shows a filter that constrains the crawler to just the message boards of [childline](https://www.childline.org.uk/).

[org.uk](#) and not the entire site. The depth, robots.txt and sitemap parameters work similar to the Web Crawler, allowing the user to specify a depth to crawl or an unlimited depth, in addition to being able to ignore or adhere to a site's robots.txt and sitemap.

5.2.2 Named Entity Recognition and OpenNLP

In addition to crawling and scraping forums, Named Entity Recognition (NER) was also used to quantitatively assess the main entities discussed within the various collected threads. The NER was performed using the open source java API OpenNLP, developed by the Apache Foundation. OpenNLP uses a standard Maximum Entropy (MaxEnt) approach to train models for NER in a specific language. The principal of MaxEnt states a model, or distribution $p(a, b)$, is one that maximises the "uncertainty" or entropy based on the constraints. Within this distribution a is one of all known possible classes in A and b is one of all possible contexts in B . The constraints in the case of MaxEnt are born from the features in the known evidence. The evidence, or training data, is known as "partial information", because they are a finite set of examples used to estimate $p(a, b)$. Equation 5.2 provides the formula so that p maximises the entropy. The form of p is dependent on the feature representation of the task.

$$H(p) = - \sum_{x \in \epsilon}^{max} p(x) \log p(x) \quad (5.2)$$

The feature representation of evidence in OpenNLP takes the form of a set of gold-standard, tokenised training sentences which have all named entities of the target type labelled. The OpenNLP library uses a MaxEnt based classifier to discover named entities in previously unseen target sentences. The NameFinder package typically uses a model trained to find one of seven specific entity types within a sentence; People, Locations, Organisations, Dates, Times, Percentages and Currency amounts. However, the library provides developers the ability to train custom, multi-class models. Each sentence presented to the NameFinder is tokenised and iterated over, and the classifier makes a discriminatory decision about each token as being either the start, end or part of an entity within the sentence. Tokens that have been identified as part of an entity in previous sentences of the same document are given an increased weight, as this is considered to be

additional information⁵. It should be noted that these discriminatory decisions are heavily dependant on the domain and grammatical makeup of the training data, meaning performance will be greatly diminished if the training and target corpus are significantly different in structure and features.

5.3 WELLCOME TRUST: RESULTS

This section reports the results of the mental health forum project, which helps to illustrate the potential of the use cases and additional capabilities of the CET. In addition to what was learned in the below results a considerable amount of knowledge was gained over the course of both the ACLED and NSPCC projects with regards to web scraping and crawling strategies.

For instance, methods for automatically scraping web pages, and devising a practical and efficient means for users to configure bespoke web scrapers has provided a powerful means for analysts to access structured content from across the web. In addition, the implementation of a continuous recrawl strategy that can be easily used by non-experts has proved a valuable asset to the overall system. The combination of these two contributions has also led to a significant amount being learnt about the difficulties and solutions for dealing with large heterogeneously structured document collections.

The results presented in the remainder of this section are divided into four subsections each covering a different learning objective achieved as a result of the work on mental health forums.

5.3.1 *Health forum case study*

In this project, 1,070,469 forum posts were collected across 6 popular health forums, which are summarised in table 5.2. The exact names of the forums used have been anonymised for ethical consideration, so the names displayed describe the type of forum each one represents. These forums were chosen as they represented a wide and varied array of related topics, forum sizes, patterns of use and user-base. For example, at the time of the project the patient advice forum contained

⁵ This information can be cleared after each sentence and effectively ignored, if desired

360,000 posts, from 25,000 unique users, whereas the medication focused forum contained only 13,000 posts. All posts were submitted to one of the forums between June 2004 and May 2016.

5.3.2 *Ethical review*

Before embarking on this project an ethical review was submitted and given approval to collect forum posts of this sensitive nature. To ensure the ethical considerations of this work were adhered to, all members aliases were anonymised by creating a unique string as a replacement for their actual username. In addition, no individual user was profiled for the work and all posts were only kept for the duration of the project. At the outset it was established that there is an ethical consideration about whether it was permissible to crawl these forums without notification to the site owners. If any forum forbid the crawling of its domain via its robots.txt or NOINDEX and NOFOLLOW instructions, then the data was not collected and that forum not used.

5.3.3 *Data collection*

A bespoke web scraper was configured for each site and the posts and their metadata collected. A short description of each forum is given and the relevant figures presented in table 5.1, which shows the calculate average number of posts each user makes, and table 5.2, which shows the raw count of posts made during the collect period. These tables show statistics for the six forums chosen for this case study and begin to shed some light into the different ways people use these them. For instance, it can be seen in Figure 5.1 that both Mental health forums and the Carer's forum have significantly higher average posts per user, potentially implying an increased likelihood of users engaging in conversations.

All of these forums focused on topics surrounding mental health, but each was chosen due to its particular focus on a specific area within the context of mental health. Mental Health Forum 1 and 2 were both two medium sized forums, with several hundred thousand unique members posting on each site. The topics discussed on this forum were any and all subjects surrounding mental health experiences, problems and requests for advice or help. They were

Forum	Average posts per user
Mental health forum 2	45.3
Carer's forum	42.66
Mental health forum 1	35.87
Mental health medication forum	20.45
Patient advice forum	14.65
Depression forum	3.51

Table 5.1: Average number of users per forum

both commonly known and used forums, with a large number of active topics at all times. The Patient Advice Forum had the largest user base as can be seen in table 5.2, with topics focused predominantly on users asking for advice regarding their disease, condition, medication or experience. From this forum only users and forum topics regarding mental health were collected. The Carer's forum was predominantly occupied by users sharing advice and experiences regarding those they care for and again only those threads regarding mental health were collected. The Mental Health Medication Forum focused on topics for those seeking advice and experience on their medication. These threads often discussed topics such as, non-adherence to a drugs, side effects, dosage and reassurance from experiences. Finally, the Depression Forum contained topics focused solely on mental health issues that cause or come as a result from depression. Depression in this case referring to acute depression caused by external forces or events, or chronic, medically diagnosed depression.

Site	Date of collection	Unique posts	Unique users
Patient advice forum	05/2015 - 06/2016	362,623	24,746
Mental health forum 1	09/2005 - 06/2016	339,752	9475
Mental health forum 2	10/2007 - 06/2016	192,101	4239
Carer's forum	02/2007 - 06/2016	106,656	2495
Mental health medication forum	12/2010 - 06/2016	56,040	2744
Depression forum	06/2004 - 06/2016	13,297	3781

Table 5.2: Table of forum collection statistics

5.3.4 *Analysis*

Three methods of analysis were used in this project that utilised the features provided by the CET. All analysis was performed on each forum separately to differentiate discussions and highlight the ways in which each forum is used. Second, a bespoke UI was developed using the analytics platform Qlik⁶. This second method of analytics was used to provide a visual analytical tool for interested health professionals and representatives of the NHS in a half-day workshop intended to present the project findings and allow these parties to explore the data. Third, a focus group was conducted at the King's fund to explore the data with a number of health professionals from the NHS and a number of researchers in computational linguistics. This focus group was conducted to discuss the efficacy of the data collected, explore the collected content using the UI. Fourth, the same named entity recognition implemented in Chapter 4 was used to discover entities of interest within the dataset. For instance, mentions of organisations, such as the NHS.

5.3.5 *Cries for help*

To help discover cries for help a dataset was constructed using just the first posts from all threads in the entire collection as it was considered more likely that these posts contain questions and statements intended to begin a discussion, or ask directly ask for advice or help. This resulted in a dataset consisting of 158,548 unique forum posts from across all forums crawled. Members of the research team at DEMOS then used Method52 to build active learning based-classifiers to identify these cries for help. The labels used to classify each post were 'needs help' or 'does not need help', creating a two class classifier trained to only identify posts relevant to cries for help. A cry for help within the scope of this project would often manifest itself as one of two main types. First, people directly asking for help or advice about a particular mental health issue they or someone the care for is suffering. These were considered non-urgent cries for help and often involved statements of interest and advice about certain treatments or to share experiences with others. Second, a more urgent cry for help was identified, where people were not necessarily asking for advice voice exasperation

⁶ Developed by Josh Smith, another member of the research team.

and statements detailing a failing ability to cope. This results in users turning to forums to help them cope, as they may not feel they have anyone to turn to, can seek support and potentially receive immediate responses from others who sympathise or share similar feelings. The latter was of particular interest and importance to discover in this project as it represents a cry for help that could result in immediate harm to an individual.

The relevancy classifier identified 95,544 as 'cries for help' which comprised of 60% of the entire collection. However, the classifier only reached 65% accuracy. Although the classifier displays relatively accuracy, it still provided significant help in the filtering of posts and subsequent analysis discussed below.

To filter these first posts identified as cries for help further a number of other classifiers were built. The posts of particular interest were any deemed to be discussing treatment or contact with the help services. The intuition behind identifying these types of posts were cries for help as a result of a failing in the health service. The three classes trained in this second classifier were as follows: 'sought help' implies a forum user who has sought medical help already, 'not sought help' indicated someone had not sought medical help but should have, 'other' was a general classification used to classify those that did not necessarily need for medical help with their problem. The trained classifier identified 50,333 of the original 95,544 as posts from users who had sought medical care and 24,132 as members who had not. The overall accuracy of the classifier was again quite low at just 56% but the 'sought help' classifier came in at a significantly higher 69%. In spite of the poor performance of the classifiers, there is strong evidence to show that there are a significant number of people going on to online forums for help and advice about a medical condition that is relevant medical care. These results, although exaggerated by the low classifier performance indicate a likelihood of members being dissatisfied with the medical care they have been provided, or more worryingly using these forums without consulting a health professional at all. It is worth noting that again in spite of the low performance of the classifiers the task of an analyst was significantly reduced using these classifiers as filters and allowed qualitative analysis of a sample of posts that contain a high number

of relevant posts that could be used as a means for discovering how people are using these forums.

5.3.6 *Cognitive behavioural therapy*

During the initial consultation and focus group at The King's Fund, Cognitive Behavioural Therapy (CBT) was suggested as a topic of focus due to its increasing prevalence in mental health treatments, especially using talking therapy for those suffering from anxiety and depression. The aim of the work was to see if it was possible to identify forum posts discussing CBT and gain a measure for the demand and perceived desire for the treatment in online communities. This resulted in an experimental corpus of 8374 posts which mention CBT.

The CET was utilised to build the corpus by finding all posts in the entire dataset that contained of the three key phrase 'CBT', 'Cognitive Behavioural Therapy' or 'Exposure Therapy' a variant of the method. A classifier was then built to identify posts as being in one of two classes, 'have-had' or 'other'. Posts classified as have-had were any found to be authored by anyone sharing current or past experience of having CBT or sharing information regarding a planned future treatment. Posts classified in the 'other' category were intended to be those that discuss CBT but did not contain any information regarding personal experience. Of the 8374 posts, 5582 posts (66%) were classified as 'have-had', with the classifier evaluated as having an 77% accuracy in the 'have-had' category and 53% accuracy. Some of the key findings from the 'have-had' dataset analysed by members of was that people who share experiences of the CBT treatment tend to focus discussion on very personal feelings regarding the affects of the treatment, associated medications and other means of finding information or methods to augment the treatment in the form of apps or other websites that contain information regarding treatment.

5.3.7 NER

In addition to using classifiers and keyword search, the forum posts were also analysed for named entities using the NER technologies originally presented in Chapter 4. This provided a simple means to quantify how much certain organisations are discussed within mental health forums. The most prominent organisations and the count of times mentioned is summarised in Table 5.3.

Organisation	#mentions
NHS	5332
Social Services	1626
Department of Work and Pensions	1164
BBC	1018
Food and Drug Administration	951

Table 5.3: Count of organisation mentions across all forums

As shown in Table 5.3 above, the main focus of discussion was the NHS, which had a significantly larger number of mentions suggesting that forums could potentially prove a useful resource for gaining in sight into the views, opinions and experiences of getting treatment from the NHS and other institutions. One significant observation that was made is the prevalence of the Food and Drug Administration which is a US based institution. This is of particular interest as all forums chosen for this work were based in the UK, as it was the experience of UK patients that was the primary focus of study. This high number of mentions of the Food and Drug Administration indicate the global audience and participants that use these forums, which was a not originally highlighted at the outset of the work, but only serves to widen the scope of the population sample created by scrutinising these forums. This 'global' quality also highlights another potential advantage to analysing online forums, due to the wide variety of people sharing information.

Through the methods of crawling, scraping, keyword filtering and classification, the features provided by the CET allowed a domain-specific corpus relating to CBT was built and more easily analysed, by finding potentially relevant documents and filtering them to a mere few thousand. This allowed a new insight into the

experiences and behaviours of patients in a specific area of mental health to be analysed.

This concludes the results section of this chapter, which ends with a discussion on the capabilities provided by the CET and how they have been utilised to collect domain-specific corpora across a number of uses and topics.

5.4 DISCUSSION

Over the course of this chapter three use cases for the CET have been presented in the form of three case studies that make use of its capabilities or precipitated its further development. These three case studies can be summarised as follows. One of the outcomes of Chapter 4 was the discovery that the boilerplate that surrounds the desired text content of a page can add significant noise when attempting to train classifiers. As a result, it was identified that a solution was needed in order to mitigate this issue. In addition, both the mental health forum and NSPCC project required that individual forum posts and associated metadata be crawled from specific sites as annotated sub-documents. For example, during the ACLED project it was necessary that the main news article of a page, its title and published date be taken from every online news source. The three case studies presented in this chapter each present a similar use case that the sources and domain to construct a corpus are known but very granular recovery of page content is required. The solution to this use case was the implementation of the Manual Scraper software component, that allows users to scrutinise a sample of pages taken from a target site and configure a custom scraper to retain and structure any number of text elements from that site's pages. For example, the individual forum posts, author username, and post date from a forum or message board website, such as those found at www.childline.org.uk.

The above use case sits in contrast to a second use case illustrated by the diaspora voting case study presented in Chapter 4, which requires a general crawl of the web, with no knowledge of what sites will be visited or what the HTML markup of any discovered site may be. To address the boilerplate issue for this case study the Article Scraper component was added to the CET, which provides a domain

independent solution to boilerplate removal. Although individual metadata cannot be recovered in this instance, the main text content for a page can be more reliably retrieved, whilst removing most other irrelevant content found on a page.

The ACLED and NSPCC case studies presented in this chapter both share the need to continuously crawl the same list of sites, whilst returning only new content. This requirement precipitated the addition of the Continuous Crawler to the CET in addition to the standard Web Crawler software tool. The combination of the Web Scraper and Continuous Crawler gives users a simple but powerful tool to continuously search a domain for pages of interest and store a significant amount of structured metadata.

All three of the case studies that have been presented in this chapter share one key similarity, that the sites of collection are known and fixed, meaning that there is no capability or need for the CET to be able to discover or include newly discovered information found during analysis. This makes the data collection process much simpler as the location of the information is known and simply needs to be collected, stored and analysed. Chapter 6 presents project Parrot, a case study intended to discover instances of people advertising, discussing and directly selling illicit wildlife across the internet. In this instance, very little information is known about the vernacular used by those selling items and the location of sale is also an unknown. To address this issue the next chapter details the use of a phrase and feature extraction toolkit originally developed by [Robertson \(2019\)](#) and the implementation of web search technologies to expand or construct a corpus from the general web.

To summarise, the subsystems of the CET introduced up to this point have been developed to handle two forms of use case. First, where the source of information remains unknown but the required domain of the collected corpus is known. Second, where the location of the information and the domain of the collected corpus is also known. In addition, the CET is now capable of a number of methods for scraping to remove boilerplate, and crawling to collect a domain-specific corpus continually from known or discovered sources of interest.

The next chapter presents the development of the CET to manage use case where the true domain of the required corpus is unclear and the location of where to find that information on the web is also unknown.

DISCOVERING DOMAIN CONTENT FROM UNKNOWN SOURCES

In the previous chapters we have seen 2 possible use cases where the CET can be applied, and in this chapter we introduce a third. Each case study presented across all chapters fall in to one of the three cases listed below.

1. Domain known but source unknown.
2. Domain and source known.
3. Domain unknown and source unknown.

In this chapter, we present the third use case, one in which the domain is poorly defined or unknown and the location of where to collect documents online is also unknown. To address this situation two new software components to the CET are introduced in the form of phrase/feature extraction and web search which builds on the original work presented by [Baroni and Ueyama \(2006\)](#) for bootstrapping corpora using the web. This chapter is intended to present the complete flow of information envisioned through the CET, which advocates the use of a human-in-the-loop strategy. The purpose of this chapter is to illustrate how this philosophy allows full control over the evolution of a domain-specific system, generated over multiple iterations. To demonstrate the efficacy of the complete CET a case study is presented that uses the method to discover instances of people selling illicit wildlife products online. The main challenge with this particular project was that those selling illegal wildlife products wish to avoid detection, whilst continuing to advertise their products. The results will show that to find instances of people selling these illicit items, a vocabulary of esoteric domain vernacular must be built and learnt. In this example, it is demonstrated through a combination of automated computational search methods and qualitative analysis, which evolves a model of language that fits the domain for a particular animal product being sold. This chapter illustrates the complete power of the CET and the virtues of human-in-the-loop methodologies in instances where

little is known of the target corpus needing to be built. This chapter is organised in to five sections. First, the details of the case study, referred from here onwards as Project Parrot (PP) are given. Second, a review of the literature that conducts similar work in the field of illicit wildlife sale is discussed. Third, the complete methodology of the CET used in PP is described and the methods used in its implementation. Fourth, the results of this work are given. This chapter concludes with a discussion of the findings and efficacy of the CET.

6.1 CASE STUDY: PROJECT PARROT

This pilot project was a piece of research done at the request of the Global Initiative Against Transnational Organized Crime (GIATOC). The purpose of this work was to explore the possibility of identifying instances of people selling illegal wildlife products online, using computational collection and machine learning methods. For this project, the CET was chosen as its unique feature set and capabilities made it suitable for the task of building domain knowledge and collecting a corpus from the web; whilst providing a means to organise, filter and analyse the results for positive examples of illicit trade.

The illicit sale of online wildlife is common place but sometimes hard to discover due to sellers obfuscating their advertisements. Three animal products, listed below, were chosen as case studies to test the efficacy of the method.

- Pangolin scales
- Ivory
- Orchids

These cases were chosen as they represent three of the most widely trafficked animal products in the world. Rare and illicit orchid species are often the target of collectors wishing to obtain examples of wild-grown species of endangered orchid breeds. In recent years it has been shown that these orchid breeds can be grown in captivity, but their difficulty of cultivation and demand for wild-grown specimens often finds illicit wild orchids being added to legally grown collections for sale, as demonstrated by [Hinsley and](#)

Roberts (2018). In a similar case, ivory is one of largest markets of internationally traded illicit wildlife items and can be found on sites as commonly used as eBay. Pangolin scales are a common ingredient in Traditional Chinese Medicine (TCM), which is practised by many people across the world. This market for pangolin scales has made the Pangolin one of the most endangered species on the planet. As will be shown in the results of this study, the problem is so significant at this point that products containing pangolin scales can be commonly found on some e-commerce sites that focus on TCM medicines and treatments. To illustrate the findings achieved by a non-domain expert using the CET, this case study focuses on the example of finding pangolin scales being traded online. The next section presents the literature most closely related to this work.

6.2 BACKGROUND

Since the rise in globalisation and growing power of the internet there has been an ever increasing online market for illegal wildlife and wildlife products. Recently reported figures state that this market is estimated to be worth an estimated 91 to 250 billion USD, a market which rose in value by 26% between 2014 and 2016 (Yeo et al., 2017). In recognition of these growing markets, the Trade Records Analysis of Flora and Fauna in Commerce (TRAFFIC) in 2012 began the surveillance of 15 Chinese language websites across China, Vietnam and Hong Kong. By 2014 the list of sites had increased to 25. In their work, TRAFFIC continually crawled these sites, searching for a known set of code words used by the sellers of illicit items and reporting pages containing potential hits to the site owners for moderation (Yu and Jia, 2015). This work illustrated the size of these online markets and has since motivated organisations such as TRAFFIC, CITES and ICPO-Interpol to increase their surveillance and action in identifying the extent and locations of these markets, in order to begin disrupting them. This need for greater investigation has precipitated a number of research methods to combat these issues.

This work can be encapsulated by three scopes of research. First, the need to identify the main exporters and importers of these products and any intermediary members of these networks (Patel et al., 2015). Second, establishing the legality of a controlled

wildlife product. For instance, a thriving market for artificially bred endangered species of certain plants species has arisen to produce a legal means to alleviate the illegal trade of wild species, but have been know to be used to launder wild specimens (Hinsley and Roberts, 2018). Third, identifying incidences of trading on online, notably those selling, where and how.

The most recent work in this latter area has attempted to perform this work at scale by combining qualitative analytical approaches with computer science and machine learning. A large proportion of this work has focused on establishing the extent and workings of these markets on popular sites such as eBay and social media sites, such as Facebook and Instagram. Due to the relatively recent inception of such efforts to investigate and stem the online market of illicit wildlife, people within these markets have enjoyed relatively uninterrupted business on such openly available platforms. These approaches often prove successful due to their ability to potentially identify the sellers in these markets, the type of language they use and the relative ease these sites can be crawled and analysed using machine learning techniques (Yeo et al., 2017; Hernandez-Castro and Roberts, 2015). For instance, Yeo et al. (2017) performed an 8 week study intended to establish the extent of sellers of illegal ivory on the popular auction site eBay UK. Starting from 28th march 2014, a daily search for the keyword 'ivory' was conducted on the eBay website. All potential hits were assessed by two former law enforcement officers and any that did not comply with the terms of the AWPP, were recorded. Their results showed that there were between 528 and 633 postings of illicit items per week over the course of the 8 week period; thus exposing the extent of the problem, and by recording the sellers details, a potential solution. Similarly, Hernandez-Castro and Roberts (2015) curated a similar collection over an 8 week period. Hernandez-Castro and Roberts (2015) used the data mining tool Orange to discover features of the text to allow a two stage process of analysis and capture (of Ljubljana, 2019). The first stage simply discerned if the document discussed ivory and the second stage discerned if the item in question was illegal.

Social media sites, such as Facebook and Instagram, have also been the object of much study due to their wide-spread use in these markets. In a five month study during 2016, TRAFFIC monitored

14 Facebook groups in Peninsular Malaysia and found that on average 46 posts per month concerned the sale of a number of endangered species. Similarly, Di Minin et al. (2018) found that Instagram was often used to post images of items for sale as a means for establishing interest in potential buyers, and proposed that the recent developments in deep-learning could be used to identify these images at scale across the site (Di Minin et al., 2018). It should also be noted that poachers and traffickers have been known to trawl Instagram accounts and holiday makers on safari looking for images that could provide clues on potential hunting spots (Rosen and Smith, 2010). Social media has the added advantage of having closed groups and the ability to move the actual discussion of item sales to more private messaging services, such as WhatsApp or WeChat, that use encrypted peer-to-peer channels. This is leading the owners of these companies to act in collaboration with investigative organisations, such as TRAFFIC (Traffic, 2015).

Although this work is crucial to the development of these methods and technologies, they do however exclude much of the wider internet and those sites which are currently undiscovered. What the above work does show is that just on simple searches for the most basic of terms yields a huge number of results, but does not necessarily capture other aspects, such as seller code words that could obfuscate further instances of illicit items for sale. To address these two problems of scope and scale, work has been conducted to develop automated approaches that cast a wider net.

6.2.1 *Multiagent systems for identifying illegal wildlife trade*

Oliveira (2014) proposed WATES, a MAS for identifying posts related to the trafficking of wild animals. WATES is a system comprised of many agents which communicate between each other and ultimately relay the results through a number of analysis experts. The example they use for an expert is The Fetching Expert, a sub-system designed to retrieve more content from the web for analysis. The main agents within this system comprise of a spider — or web-crawler — page-filter — for filtering useful information — and a page-exporter for conveying the information in a structured manner. This would encapsulate not just web-content, but also structured metadata and fields that may exist in the source that could be of use.

The work of [Vicente et al. \(2018\)](#) can be considered as the MAS most relating to the CET. In their work, [Vicente et al. \(2018\)](#) developed a system for the real-time analysis of social media consisting of 6 discrete experts: crawler, keyword discovery, topic extraction, sentiment analysis, IXA-pipes and a UI. In this system information flowed from the crawler, which collects new information, to keyword discovery, topic extraction and sentiment analysis, depending on the task. The intermediary steps between the crawler and UI act as means to identify key information within the texts, and to discern relevant documents to be presented to user.

The CET proposed in this thesis, attempts to develop beyond the above work as it allows a more dynamic relationship between the analyst and system by allowing the analyst to influence and change the definition of relevance by each agent over multiple iterations. In doing so, this provides a means to access the wider internet and dynamically define domains and discover instances of wildlife trafficking, related content and language.

6.3 METHODOLOGY

As previously mentioned, the main methodology used in building illicit wildlife corpora followed a human-in-the-loop strategy. This sees an analyst having an integral role during each stage of the process, to best utilise the methods and tools made available in the CET. The methodology in this case study followed a two stage process. First, a small, high quality corpus was built by iteratively cycling over phrase extraction, search and analyst filtering. Second, this corpus was then expanded further through the use of crawlers, classifiers and additional filtering phases. The following subsection introduces the flow of the complete Corpus Expansion Toolkit (CET), with a brief introduction to its general purpose and method. The two remaining subsections present each of the two stages used in the methodology and includes the methods and contributions to the CET that made them possible.

As mentioned in the introduction to this chapter, the methodology and components of the CET presented here provide a solution to the final use case addressed in this thesis that. This use case sees corpora generated from online sources when the exact domain is poorly

defined or unknown and the location of relevant documents is also unknown. The choice of a human-in-the-loop strategy is integral to this method as the involvement of an analyst at all stages allows for the incremental definition of the domain, and the continued education of the analyst. This symbiotic relationship allows the analyst to become a quasi-expert in the target domain language as the system evolves. One of the most significant contributions of the CET presented in this form is the ability to generate not just a corpus, but a system for continually generating more content. By retaining extracted phrases and extracting more from newly performed search results a vocabulary characteristic of the domain can be built, maintained and adapted over time. Using the combination of crawlers and classifiers to filter out content, allows the analyst to continually search for more content with little overhead beyond updating the vocabulary once the initial work has been conducted. A second contribution presented in this chapter is the capability to discover potential seeds to conduct a crawl and find more relevant documents, utilising classifiers once again to filter out erroneous content. As previously mentioned the next subsection introduces the subsystem of the CET used in this case study.

6.3.1 *The Corpus Expansion Toolkit*

Over the course of three chapters a number of the CET's features have been introduced. In this final chapter the complete toolkit is presented, the flow diagram of which is illustrated below in Figure 6.1

There are two key contributions introduced in this chapter and illustrated in Figure 6.1. First, the analyst sits at the centre of all phases of the pipeline, providing support in defining the domain as the system evolves. Second, the introduction of cyclical iterations over the documents and extracted information provides the ability to revisit previous stages in the pipeline to improve yield and conceptual specificity as the target domain becomes better defined. In Figure 6.1 these feedback systems are represented by the dotted arrows. Prior to discussing this in more detail, it is worth mentioning the bootstrapping methodology introduced as a contribution in this chapter. The inclusion of the Surprising Phrase Detector and implementation of the Web Searcher, discussed below, allow a

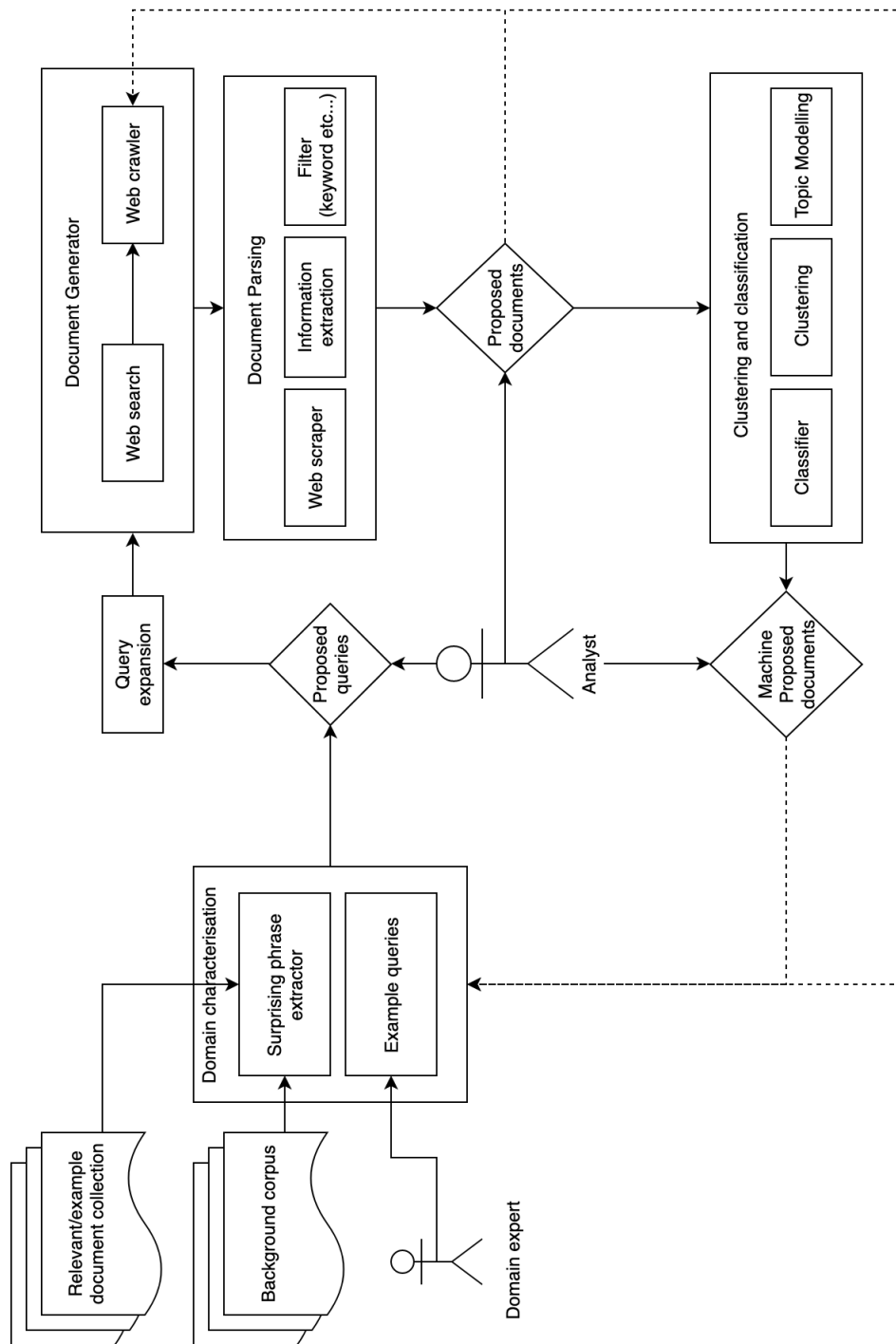


Figure 6.1: The Corpus Expansion Toolkit

method similar to that of BootCaT to be included in defining a domain and its vocabulary.

The intention of a system devised in this way is to address use cases where the domain and source locations to find documents is unknown or poorly defined. Having the potential to revisit previous stages presents a solution to this problem, by allowing the analyst to revisit previous steps from a more informed position and to filter out unwanted information generated by the automated computational components. To give an example, the analyst may decide to look at a sample of system proposed documents that may have been returned from a search or crawl and manually choose those which best fit the domain of interest. These chosen documents can then be used to bootstrap the original corpus by having further information extraction performed and thus creating a set of more informed search queries. To give another example, a user may choose to train one or a chain of multiple classifiers to find multiple facets of a fine-grained domain of interest within a collection. Choosing to then expand a sub-collection by extracting further phrases relevant to that class, or crawl the web under the assumption that more relevant content may be linked to pages found in that class. In addition to further informing the system of the target domain, an analyst is able to become personally informed of the domain. The example shown in this chapter is an analyst with no prior knowledge of a domain, in this case the use and illicit sale of pangolin scales, progressively becoming a quasi-domain expert who is informed by the results of many iterations over the CET. The coming subsections present the various implementation details of previously undiscussed components of the CET.

6.3.2 *Human-in-the-loop strategy*

As previously stated the CET implements a human-in-the-loop strategy that allows analysts to work with the various agents in the system to guide information discovery and extraction. This strategy manifests itself in a number of ways within the main agents of the CET, but also between agents via a number of software components present within Method52. These additional components are listed as follows:

- Keyword annotator
- Category annotator
- Frequency annotator
- Custom logic annotator
- Finder Replacer

6.3.3 *Frequency Annotator*

The Frequency Annotator allows users to make discriminative decisions based on the frequency of a chosen document annotation. For instance, ignoring documents with an annotation type that falls below some threshold to mitigate for data sparsity.

6.3.4 *Custom Logic Annotator*

The Custom Logic Annotator allows users to define rules and criteria that document annotations must pass in order to be considered a positive match. For instance, filtering out documents that do not have a particular annotation present in their record.

6.3.5 *Category Annotator*

The category annotator is a software tool that allows users to define a number of categories and manually annotate documents with those categories in a manner synonymous with creating gold-standard classification datasets. One common example within the work presented in this project is to define two simple categories 'relevant' and 'irrelevant', manually reviewing all documents and using these category annotations to filter out all documents irrelevant to the domain. Although this example presents a two class categorisation, the Category Annotator can be used to organise a document collection into any number of arbitrary sub-collections.

6.3.6 *Keyword Annotator*

The Keyword Annotator is a software component that allows users to build lists of keywords to search for in target documents. These

keyword lists of can be ordered into categories that subsequently allow document collections to be organised into categories based on the keywords they contain.

6.3.7 *Duplicate Annotator*

The Duplicate Annotator uses exact matching or shingling to filter out near or exact duplicates from a single dataset (Broder et al., 1997). This component is often useful when parsing large corpora built from a web crawl or combining corpora that are known to contain the same documents.

The next subsection describes the first stage of the corpus expansion methodology used during this project and the methods used to perform the task.

6.3.8 *Phase One: Bootstrapping*

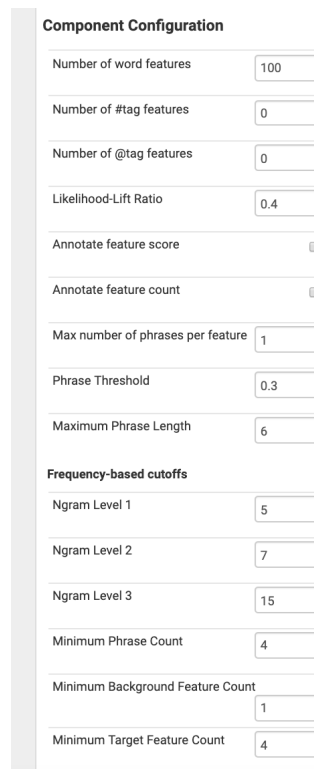
Phase one of this two stage process involves the generation of a small but ‘high’ quality corpus, expanded from an initial set of seed documents or a set of characteristic phrases provided by a domain expert. For example, in the case of Pangolins, a small set of seed words and phrases were originally provided by Dan Challender of Oxford University, who is a researcher on the Oxford Martin Programme on Illegal wildlife trade and specialises in the field of the illegal Pangolin trade. This first stage uses a method very similar to that of Baroni and Bernardini (2004) and Baroni and Ueyama (2006), which bootstraps a corpus through a process of feature and phrase extraction, starting from a small *seed* corpus taken from the target domain, and web search to expand the collection. This process is repeated by performing phrase extraction on the search results to generate a new set of phrases to use as queries to the search engine. Using this method generates both new phrases and new documents that have a high likelihood of originating from the target domain. The contributions to this methodology presented by the CET are the inclusion of an improved form of the surprising phrase detector, developed by Robertson (2019), the inclusion of manual phrase and search result filtering, plus a new method of combining phrases¹.

¹ This was originally achieved by randomly combining extracted phrases to form queries.

These contributions are presented in more detail below, by describing each of the software components developed to achieve this task.

SURPRISING PHRASE DETECTOR

At its core the Surprising Phrase Detector (SPD) is an implementation of the characteristic phrase extraction method originally proposed by Baroni and Bernardini (2004) and originally discussed in 2.1. To find characteristic phrases within a target corpus the Surprising Phrase Detector takes as inputs, a target and background corpus for comparison and builds a model of likelihood for the occurrence of each word in the background corpus and compares that with the likelihood of each word in the target corpus. Words appearing with significantly higher likelihood in the target corpus are proposed as characteristic features. Multi-word phrases are then recursively expanded from the discovered features based on the likelihood of co-occurrence within the target corpus. The parameters to the method can be seen in Figure 6.2, and are discussed below.



Component Configuration	
Number of word features	100
Number of #tag features	0
Number of @tag features	0
Likelihood-Lift Ratio	0.4
Annotate feature score	<input type="checkbox"/>
Annotate feature count	<input type="checkbox"/>
Max number of phrases per feature	1
Phrase Threshold	0.3
Maximum Phrase Length	6
Frequency-based cutoffs	
Ngram Level 1	5
Ngram Level 2	7
Ngram Level 3	15
Minimum Phrase Count	4
Minimum Background Feature Count	1
Minimum Target Feature Count	4

Figure 6.2: The configuration parameters for the Surprising Phrase Detector component

The first key parameter is the number of highest likelihood word features to expand into phrases from the target corpus. Second, the likelihood lift-ratio is a hyper-parameter originally proposed

by Sievert and Shirley (2014) that provides the analyst a means to re-rank the features list. The Likelihood-lift ratio ranking of a word is the ratio between its *lift* and likelihood within the target corpus. A features *lift* is how frequently it occurs in the target corpus versus its occurrence in the background corpus. The hyper-parameter λ provides a means to rank features by re-weighting the contribution of *lift* and likelihood to a words score. A λ value of 1 weights a feature's score solely on its lift and 0 purely on its likelihood within the target corpus (ignoring the background corpus completely). Values between 0 and 1 allow for features score to be a weighted ratio of the two. The affect λ has over the original method is it provides the means to affect a feature's ranking by increasing or decreasing the effect a word's likelihood within the target corpus has on its *lift*, which could be heavily skewed when compared with the background corpus. In addition to the Likelihood-lift ratio a number of threshold parameters are provided to affect what features are extracted based on raw feature counts within the target and background corpus.

As previously mentioned, the core statistical framework for the SPD is based on the work originally proposed by Baroni and Bernardini (2004), which attempts to discover phrases by recursively expanding surprisingly occurring unigrams, to $N + 1$ grams that fit certain criteria. These criteria include words which occur with the unigram above some threshold, that are not stop-words and occur as a longer phrase, rather than a shorter one, above some threshold. These thresholds are presented as parameters to the user as part of SPD's UI. A unigram feature f in a target category t^2 is considered surprising based on a measure of their likelihood in the target corpus weighted against the background corpus, using the *lift* hyper-parameter. The formula used to calculate relevance is presented below in Equation 6.1.

$$Relevance(f) = \lambda \cdot \log(P(f|t)) + (1 - \lambda) \cdot \log\left(\frac{P(f|t)}{P(f)}\right) \quad (6.1)$$

When establishing a features relevance to a target corpus, the marginal probability $P(f)$ is estimated using the background corpus, which acts as a comparator. This results in features that have a higher marginal probability being considered less relevant to the

² In this instance the domain of the target corpus

target corpus, as they are assumed to be examples of more general vocabulary than being characteristic of the target domain.

The parameters to control the expansion of features to a phrase, based on the proportion of times n -grams appear together compared with the original feature alone, culminate into a set of features that allow for the fine-grained construction of corpus-specific phrases tailored on the size and variance of vocabulary within the target corpus. The choice of background corpus is important as providing a corpus originating from the same domain as the target corpus will typically result in features being returned that are very specific to the target corpus. This is because similar corpora share a similar vocabulary, and words or phrases that could be considered surprising would have to be very specific to the target corpus and not common within the background corpus. A background corpus from a very different domain will often result in features more generally common to the domain of the target corpus as there is much wider scope for words in the target vocabulary to be considered surprising.

WEB SEARCH

The Web Searcher is implemented to communicate with the Microsoft Azure Bing API (Swhite-Msft, 2019). Queries and results are communicated via http requests using Bing's RESTful API, that allows developers to send search queries and receive the ranked search results via http requests. The implemented solution used the Java interface provided by Microsoft to communicate with the Bing API. The Web Searcher software component is implemented in such a way as to allow other search engines to be implemented and potentially provide users a number of options.

Figure 6.3 shows the Web Searcher configuration screen. Here the user can set the 'Query results', the number of results Y to return per query and 'Safe Search', which specifies whether to filter results and prevent undesirable, such as pornography, appearing in the results. There is also the 'Top results' parameter, which refers to the total number of results returned across all searches. The Web Searcher takes a list of N search queries as its input that are then presented to Bing and the Y results for each query collected, resulting in NY results. The output is the set of all unique results across all searches, ranked in order of their frequency of occurrence across all N searches.

The screenshot shows a configuration window for a 'Web Search' component. At the top, there is a toolbar with five buttons: 'Save', 'Save and Close', a settings gear icon, a list icon, and a filter icon. Below the toolbar, the title 'Component Configuration' is displayed. The configuration area contains four settings, each with a label and a control element:

- Query results:** A text input field containing the number '10'.
- Search engine:** A dropdown menu with 'Bing' selected.
- Top Results:** A text input field containing the number '0'.
- Safe Search:** A checkbox that is checked, indicated by a blue checkmark icon.

Figure 6.3: The configuration screen for the Web Search component

This ranking is performed as it is assumed that the search queries are often topically related and that multiple occurrences across multiple queries implies a high degree of relevancy to the topic in question. This provides the user with a compressed view of all searches, that are ordered according to their perceived relevancy to the overall search queries/ If the queries are generated from the Surprising Phrase Detector they also have the potential to be sites and pages relevant to the originating corpus.

FEATURE COMBINER

Once the extracted phrases are filtered by the analyst and prior to search the extracted phrases and features from an example corpus are often expanded using the Feature Combiner. The Feature Combiner was implemented specifically for the CET as phrases and features extracted from a corpus are often too short or seemingly no longer related to the target domain outside of the context in which they appeared in the original text. To address this issue the Feature Combiner performs a pairwise combination of all phrases and features that are passed to it. This creates an expanded list of queries which contain a greater number of words from the same or topically related source. These phrase combinations can either be pairwise or triple-wise, which is specified by the first parameter shown in Figure 6.5. The second 'Bag-of-words' parameter is a boolean option which specifies whether to keep each of the combined phrases quoted as separate phrases, or as an independently occurring bag-of-words. The example shown in Figure 6.5 shows that the user has specified to set the phrases be a bag-of-words.

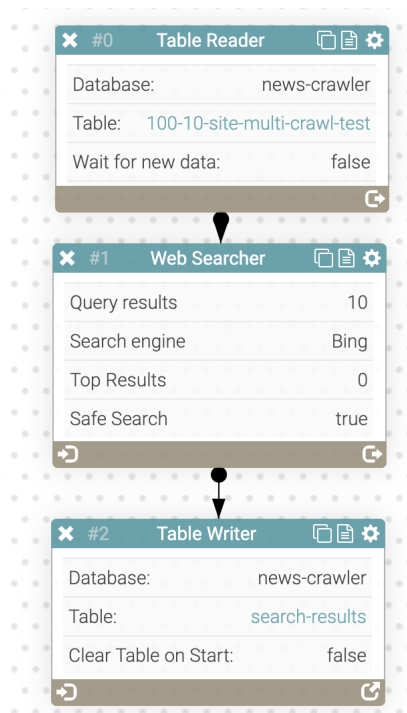


Figure 6.4: A complete pipeline that will perform a web search using the input queries.

Component Configuration

Feature combinations

1

Bag-of-words ☒

Figure 6.5: An image of the configuration parameters for the Feature Combiner component

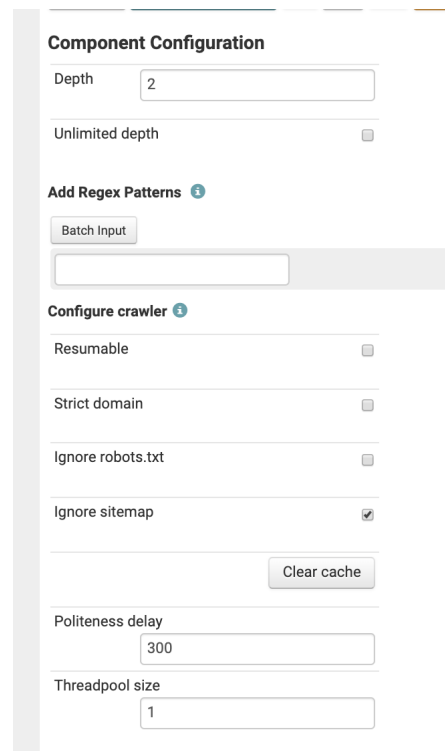
6.3.9 Phase Two: Web Crawling, scraping and machine learning

In phase one, an analyst iterates the process until a corpus of sufficient size and domain specificity is built. This forms a starting corpus from which to expand further in phase two. Referring to Figure 6.1 the document generator includes a web crawler and during phase two the source URLs in the corpus generated in phase one are used as seeds to a crawl. The assumption is that the relative high relevance of this corpus is likely to link to other relevant pages on the web. One consequence of performing a general crawl of the web is that there is also likely to be a large amount of irrelevant

content collected. The solution originally posed in Chapter 4, uses active learning- based classifiers trained by the analyst to filter the output. To generate a training and evaluation dataset for the classifier a single crawl of limited depth is performed, using the phase one corpus URLs as seeds. The precise implementation details of the newly introduced crawling components are presented below.

WEB CRAWLER

The Web Crawler component is based on a web crawler implementation using the Norconex programming API, that allows the programmatic implementation of custom web crawling applications (Essiembre, 2019). The Web Crawler takes as inputs a list of 'seed' URLs that act as a starting point for the crawler to begin scraping content and following links. The parameters to this component are summarised below.



Component Configuration

Depth

Unlimited depth ☐

Add Regex Patterns ⓘ

Configure crawler ⓘ

Resumable ☐

Strict domain ☐

Ignore robots.txt ☐

Ignore sitemap ☒

Politeness delay

Threadpool size

Figure 6.6: The Web Crawler configuration options

The depth of a crawl specifies how many times the steps of following the links found on a page, and extracting and following links discovered on subsequent pages. The parameter 'Unlimited depth', allows the analyst to specify whether to crawl indefinitely until it is manually stopped. The 'Resumable' parameter keeps a

record of the web-crawlers state so that if stopped it can be resumed from the last set of links in its crawl queue. The 'Strict domain' parameter restricts the crawler to only being able to crawl within the sites present in the original list of seed URLs. This is useful when attempting to crawl a specific website as opposed to performing a general crawl outwards across the web. The 'Ignore robots.txt' and 'Ignore sitemap' parameters specify whether to honour the crawler instructions often provided by sites via these files.

In addition to those components discussed and used in this chapter there is a number of other software components within the CET that are briefly presented below.

6.3.10 *Clustering and topic modelling*

One contribution made to the CET to add more analysis and machine learning based document proposal systems come in the form of k-means clustering and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). These software components provide unsupervised methods from semantic analysis, document clustering and classification at both the word and document level. These components use the Java based Spark machine learning suite at their core and provide a number of parameters to control the output, such as the number of topics or clusters and the hyper-parameters for LDA. Once a model is trained using either method, previously unseen documents can be streamed (such as a web crawl), decoded by the model and separated into sub-corpora depending on their topic composition or classification. An analyst is able to view a representation of the clusters or topics discovered and therefore choose specific sub-corpora to analyse further.

Although not utilised in the work presented in this thesis, these components are included within the CET as they provide a potential means to perform course-grained thematic analysis of corpora through their latent topical structure. This can be useful when exploring corpora in a new domain when an analyst may not know anything about its characteristics or topics of discussion.

6.3.11 *Twitter and Reddit API*

In addition to a general web crawler the CET has software components that provide access to the Twitter and Reddit APIs provided by each platform as a means to collect tweets from particular hash-tags or chosen Reddit threads.

The next section presents the results of project Parrot, with specific attention made to the discovery of Pangolin scales for TCM.

6.4 RESULTS: PANGOLINS

In each case study the CET was utilised in a different manner to experiment with the potential for different potential methodologies. In this section the methodology and results for the Pangolin case study will be focused upon as they provide the best demonstration of the CET. The process of data collection in phase one began by performing feature combination and web search using an initial seed list of phrases provided by the pangolin domain expert, Dan Challendar³. Over multiple iterations of search, phrase extraction and analyst filtering an initial 'high quality' corpus was build, consisting of 11,972 documents. The documents of this corpus were considered to have a high likelihood of pertaining to Pangolins and/or TCM. In addition, the use of the Category Annotator within the CET allowed for the creation of five characteristic sets of keywords and phrases representing distinct sub-topics in the domain of Pangolins.

1. **Pangolins** - Names for Pangolins (e.g. code names or Latin binomials).
2. **Brands** - Brands known to potentially sell Pangolin products.
3. **Comms** - Communication types (social media,bidding).
4. **Purchase** - Words associated with sale ('buy', 'purchase', 'add to basket', etc...).
5. **Sites** - Site names/domains known to potentially sell Pangolin products.

³ The exacts list can be found in Appendix [7.1.6](#)

KEYWORD ANALYSIS

Creating these five distinct categories allowed for a more focused investigation, generation and expansion of individual corpora relevant to each sub-topic, or to generate combination corpora using keyword searches. For example, in all instances where this was attempted each set of keywords contained at least one name for a Pangolin (category 1) and at least one phrase taken from categories 2-5. These investigations and size of generated corpora are summarised below in Table 6.1, which displays the keyword category combinations that yielded a corpus significant enough in size for analysis.

Keyword combination	Corpus size	% of final corpus
'Pangolins' and 'Brands'	342	8.03%
'Pangolins', 'Brands' and 'Purchase'	320	7.77%
'Pangolins', 'Brands' and 'Sites'	3	0.07%
'Pangolins', 'Sites' and 'Comms'	74	1.79%
'Pangolins' and 'Comms'	122	2.96%
'Pangolins' and 'Purchase'	503	12.21%
'Pangolins' and 'Sites'	3	0.07%

Table 6.1: Table of keyword categories and % of corpus recovered

This method of creating distinct categories and manageable sub-corpora, using keyword search, allowed for the qualitative analysis of the sites identified. A number examples of what was found using this method are discussed below.

EXAMPLES OF SALE

Figure 6.7 shows two images taken from a site selling a product described as 'Armadillo Counter Poison a.k.a Chuan Shan Jia Qu Shi Qing Du Whan'. The discovered phrases, names for pangolins and those provided by the domain expert enabled the CET to discover a site using the code name, 'Armadillo Counter Poison' and one given name for pangolins, Chuan Shan Jia. The ingredients for the drug also list the product as containing 21% 'Chuan Shan Jia Anteater Scales', where Anteater is another code name used for pangolins. The image does not however show where or how one can purchase the item, only that the site advertises it. Directly adjacent to the name, the number '15100' is visible and it can be seen that it is a link. Selecting this link

navigates to another part of the site advertising the actual product, how it can be purchased, the price and that it is in stock. It makes no mention of the ingredients here, meaning that the site organiser has been able to separate the potentially incriminating evidence of the ingredients from the actual product. This example shows that the CET was able to find potential examples of sale but it took the scrutiny of an analyst to make a positive assessment, demonstrating the purpose of the CET's emphasis of a human-in-the-loop strategy.

Figure 6.8 shows another example of a site selling the same product as that shown in Figure 6.7 and using the same codewords. Once again there is no mention of any ingredients the product contains at the point of sale. This page, however, provides information that further informs the analyst. For instance, the types of ailments it is meant to treat, such as improving circulation. This sort of information can be easily added to the evolving system to better inform the CET of the target domain, which in this case is the area of TCM that Pangolins are often used.

Figure 6.9 shows an example of a seller that has removed pangolin scales from their product and represents an interesting development that was discovered over the course of the project and that alternatives are available.

Figure 6.10 shows more information regarding the area of TCM pangolins are often used, in this case common ailments of pregnant women. It is also interesting that Pangolin is repeatedly misspelled as 'pongolin', but confirmed as Chuan Shan Jia and Squama Mantis. The means of sale is also available but is again presented as a separate link. Although it cannot be seen in this image, scrutiny of the site shows strange activity when navigating to the site, including browser checking and a circuitous means of ordering the product, which may be a variety of means to protect the seller and buyer.

The final image of positive examples shown in Figure 6.11 was collected as the domain became more clearly defined within the system. On analysing this item it was not immediately clear to the analyst what had caused the CET to collect this document. When the ingredients of the product are searched it can be seen that one of them is dried Squama Mantis Scales (Pangolin). This shows that in

EXCELLENCE

TRUSTWORTHY
Trusted Commerce
G-Site is Verified

Catalogs.com

Secure shopping
made faster.
Pay with your credit card,
bank account, or
payment details.

UnionPay

Armadillo Counter Poison(15100)
aka: Chuan Shan Jia Qu Shi Qing Du Wan

Ingredients

Pin Yin/ English/ Percentage
Chuan Shan Jia Anteater Scales 21.1%
Bie Jia Turtle Shell 10.1%
Dang Shen Codonopsis Root 6.5%
Niu Huang Cow Gallstone 6.1%
Huang Qi Astragalus Root 5.1%
Sheng Di Huang Rehmannia Root - raw 5.1%
Ze Chi Euphorbia 5.1%
Chuan Bei Mu Fritillaria Cirrhosae 4.1%
Ju Hua (Hua Ju) Chrysanthemum Flower 4.1%
Bai Xian Pi Dictamn Cortex 4.1%
Tu Fu Ling Smilacis Glabra 4.1%
Niu Bang Zi Arctium Fruit 3.5%
She Chuang Zi Cnidium Seed 3.5%
Lian Qiao Forsythia Fruit 3.5%
Jin Yin Hua Honeysuckle Flower 3.1%
Bai Shao Peony (White) Root 3.1%
Cang Er Zi Xanthium Seed 3.1%
Gao Ben Ligusticum Rhizome 3.1%
Huang Qin Scutellaria Root 3.1%

Indications: Dry Skin, Itchiness, Red Skin Eruptions, Eczema, Acne, Hives
Chinese Symptoms: Wind Damp Heat affecting the Skin

Home > Formulas from China > Poria & Gardenia Combo Extract (Amadio Skin Formula)

Poria & Gardenia Combo Extract (Amadio Skin Formula)

<< Previous in Formulas from China

List Price: \$42.55
Your Price: \$3.95
You Save: \$3.00 (29 %)

category

2017

Essential

with Salts

Medicine

to Tea

China

orel

g

alls

tsang

rmulas

Item Number: 15100

brand: Chinese Patent Medicines

Availability: In Stock

Email this page to a friend

Go

anufacturer

Category

2017

Essential

with Salts

Medicine

to Tea

China

orel

g

alls

tsang

rmulas

Home > Formulas from China > Poria & Gardenia Combo Extract (Amadio Skin Formula)

Poria & Gardenia Combo Extract (Amadio Skin Formula)

<< Previous in Formulas from China

List Price: \$42.55
Your Price: \$3.95
You Save: \$3.00 (29 %)

category

2017

Essential

with Salts

Medicine

to Tea

China

orel

g

alls

tsang

rmulas

Item Number: 15100

brand: Chinese Patent Medicines

Availability: In Stock

Email this page to a friend

Figure 6.7: An instance of a seller providing a link to products containing pangolin scales

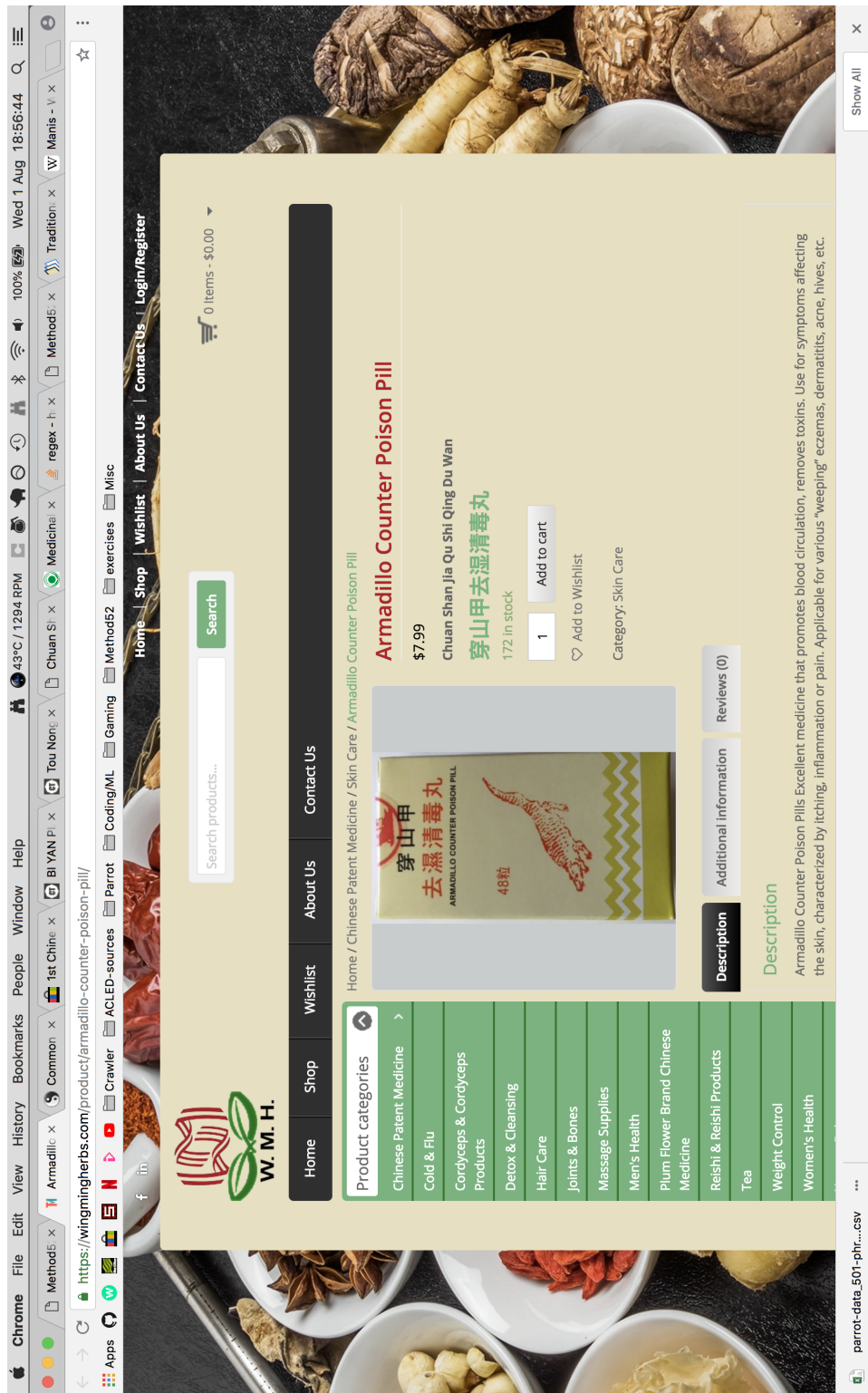


Figure 6.8: An example web page selling Armadillo Counter Poison

Zhen Ren Huo Ming Yin (Qu Chuan Shan Jia)



Code : 1024

Pin-Yin Name : Zhen Ren Huo Ming Yin (Qu Chuan Shan Jia)

Chinese Name : 真人活命飲 (去穿山甲)

English Name : Angelica & Boswellia Combination (Minus Pangolin)

Capacity :200g

Tags : **Pimples** / **acne** / **rosacea** / **dermatitis** / **carbuncles** / **swelling** / **ulcers** / **boils** / **rash** / **allergic reactions**

Share:



Figure 6.9: An example of a page selling products with pangolin scales removed.



Figure 6.10: An instance of a site selling ‘pangolin’ scales

some cases the system is able to discover more obscure information that a human analyst might miss, and requires the analyst to investigate the potential sale of items discovered by the CET.

When performing this research a number of different types of site, besides e-commerce, were commonly found in the generated corpora. These were typically 'Wiki' sites that provide information about the Pangolin, social media and conservation groups. A summary of these sites are presented below.

AMERICAN DRAGON

Home | Shop | About Us | Contact Us | Privacy Policy | Terms of Service | Affiliate Program | Newsletter | Treatment Protocols | Training | Book | Contact

Search Products

Back to herb formula index

HERB FORMULAS

TOU NONG SAN - 通經 - DISCHARGE PUS POWDER

English: Discharge Pus Powder
Also Known As: The Anti-Pyretic Powder

HERBS AND ACTIONS


Pharmaceutical Latin	Pinyin	Dosage	Actions
Rx. Atractylodes	Huang Qi	9-15g	Tonifies Qi and Blood, tonifies Wei Qi, disperses cold, promotes the discharge of pus, generates flesh and resolves abscesses. For Qi and Blood deficiency with chronic non-healing sores, itchy sores, and sores that do not perforate or only produce a thin clear discharge.
Rx. Asparagus Root	Long Ji	6-9g	Drains dampness and promotes the discharge of pus, generates flesh and alleviates pain. For chronic non-healing sores that do not perforate or only produce a thin clear discharge.
Rx. Chuanxiong	Chuan Xiong	6-9g	Invigorates the Blood, promotes the discharge of pus, generates flesh and resolves abscesses. For chronic non-healing sores with or without pus in the sores.
Dried Red Squama Herile	Chao Chuan Shu	3-9g	Invigorates the Blood, disperses Blood Stasis, reduces swelling, promotes the discharge of pus, generates flesh and resolves abscesses. For chronic non-healing sores with or without pus in the sores.

Figure 6.11: An instance of a seller obfuscating the ingredients of the product.

SHEN CLINIC

Dr. Shen's Chinese Herbal Products | Shen Clinic Herbs | Health Concerns | Special Formulas | Guan Ci Tong / Actonel | Spring Wind Herbs | Pain Medicines | Whole Herbs | Formulas | Yin-Care | Yao | TCM Guide | Contact Us | Dr. Shen's Blog | International Sales | Practitioner Discounts

Home > Tou Nong San - Discharge Pus Formula



Tou Nong San - Discharge Pus Formula

Formulas

\$ 49.00

Quantity: - 1 +

CHOOSE: TOU NONG SAN - DISCHARGE PUS FORMULA GRANULES or WHOLE HERBS:

Granules - 100 grams, Tou Nong San - Discharge Pus Formula CUSTOM MADE - NOT RETURNABLE

ADD TO CART

WIKI-FILTERING

In the entire collected corpus a large proportion of sites were dedicated to information sites, such as Wikipedia and organisations involved in the conservation of Pangolins. As these sites contain a lot of language relevant to their trade they were often found to produce false positives in the sub-corpora and produce results that obscured actual instances of illicit sale. To mitigate for this, a black-list of sites and keywords relating to these forms of sites were compiled and used to filter them out of the main corpus.

SOCIAL MEDIA

During a second phase of keyword analysis it was identified that a number of positive hits for TCM and pangolin products were found on social media sites. This observation precipitated the individual analysis of these pages. This was achieved by filtering out all pages originating from social media sites, such as Facebook, and scrutinising them. Although there were no instances of people selling pangolin-based products there was a lot public groups referring to pangolin products in TCM or conservation groups providing information on their endangered predicament. Examples of these are given below.

Figure 6.12 shows images of people having a heated discussion on drugs, treatments, risks and effects of TCM. Although not directly related to the p Pangolin these images show people sharing information and pictures of drugs used, indicating that social media sites and forums may be a rich source of information regarding potentially illicit items.

The most predominant sites found on social media were activists and Non-Governmental Organisation (NGO)s concerned with the conservation of an endangered species. These groups and sites often give information about instances of poaching, descriptions of pangolins and generally attempt to inform and advertise the plight of the pangolin. Two examples of these groups taken from Facebook and Twitter can be seen in Figure 6.13. The next section presents the results of the second phase of the methodology using the CET, which sees the expansion of the final corpus using the high quality seed corpus and a more clearly defined domain within the system.

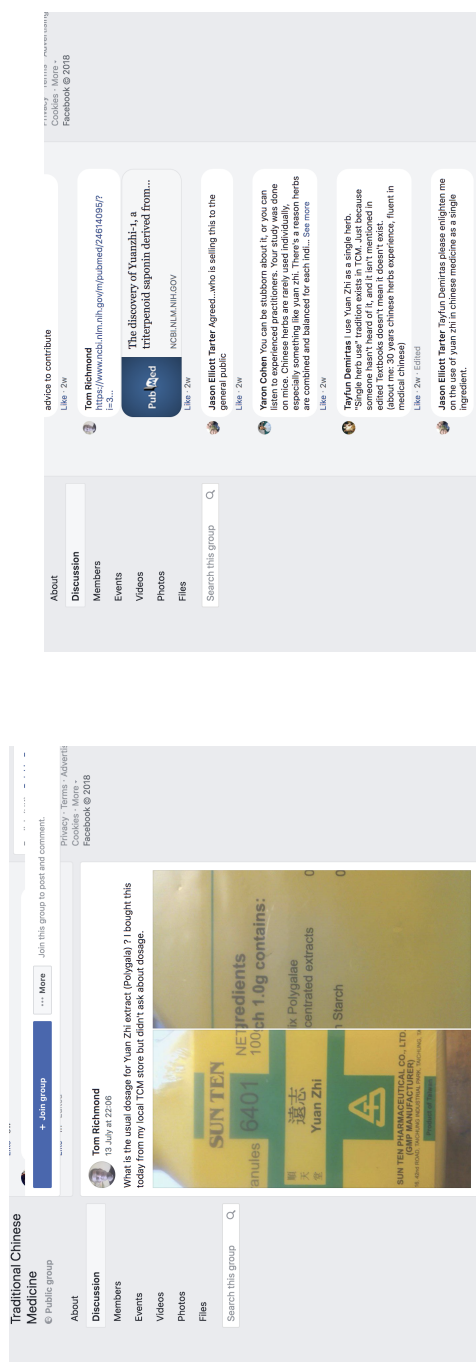
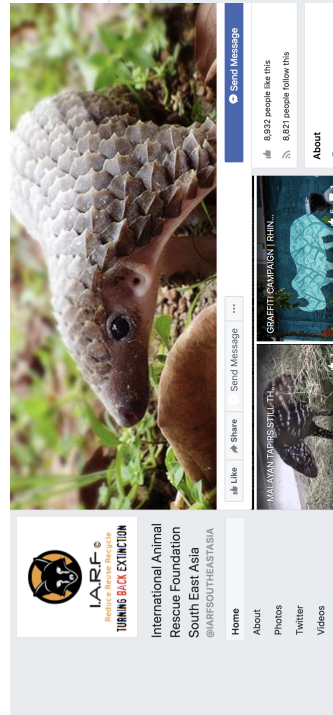


Figure 6.12: An instance of groups discussing TCM treatments

#back2thewild



Follow

Tweets 407
Following 225
Likes 666
420

People for Pangolins

Our mission is to support, encourage and raise awareness of projects and people across the globe who are committed to conserving pangolins.

London, UK
peopleforpangolins.org
Joined February 2015

Tweets

People for Pangolins @peopleforpangolins · 4 Mar 2017
We NEED YOUR SUPPORT to help rescued pangolins get back2thewild. Please DONATE generously: jungle.org/crowdfunding/p... and be a Pangolin Hero!



Tweets & replies

Media

Who to follow

Scotscape @ScotscapeW...
IUCN Pangolins @IUCN_Pang...
United for Wildlife @UfW...

Figure 6.13: Examples of NGOs and activists interested in saving the pangolin

PHASE TWO: CLASSIFICATION AND EXPANSION Due to the relatively low yield and obscure means by which pangolin based products are sold there was a significant emphasis on Phase One and the qualitative analysis of small sub-corpora to find examples of discussion and illicit sale. The subsequent web crawl and classification stages of phase two consisted of limited depth crawls from each sub-corpora, which were combined to create a final corpus in this case study of 39,823 documents. The trained classifier was used to identify potentially missed instances of sites selling illicit pangolin products, both within the original corpus and in subsequent web crawls. Below some examples of those found are presented and discussed.

EXAMPLES OF SALE IDENTIFIED BY THE CASCADE

Qualitative analysis of those documents classified as relating to the sale of pangolin products yielded new results previously not found using keyword analysis. In the majority of cases these were found to be much more obscure examples of sale and are discussed below. The evolved system created in phase one allowed for the use of the CET as a cascade. This cascade consists of crawling and classifying documents, which continually expands the corpus by providing computationally proposed documents for analysis. These proposed documents generally contained large quantities of 'Wiki' information sites, NGO sites concerned with pangolin conservation and news sites reporting instances of those illegally eating, trading or poaching pangolins. The CET also identified a number of new sources of sites selling illicit items on increasingly obscure, small sites found through a process of crawling and classifying that may not have been found using merely search. Figure 6.14 shows one such example of a site willing to sell a large number of CET items wholesale, including pangolin scales.

Figure 6.15 presents another wholesale site intended for TCM practitioners to order quantities at scale and in this instance requires a login to purchase, which implies a further level of obstruction or protection to access the illicit product. What is key in this analysis is that the CET is able to automatically identify previously undiscovered incidences of people selling illegal wildlife products.

Wholesale Herb Order Form One

Your email:

(needed*)

Your company (or personnel) name, address, zip code, Tel and fax so that we could export the herbs to you. Also let us know the port you prefer.

Please tick off every item you need, and fill out the weight of each item in kilogram. The price is in US\$ per kilogram.

☐ kg A1, a jiao^a ass hide glue^a Asini Corii Gelatinum

☐ kg A2, a jiao zhu^a ass hide glue pellets^a Asini Corii Gelatini Pilula

☐ kg A3, a wei^a asafetida^a Asafetida

☐ kg A4, ai di chu^a Japanese artisia stem and leaf^a Ardisiae Japonicae Caulis et Folium

☐ kg A5, ai ye^a mugwort leaf^a Artemisiae Apyi Folium

☐ kg A6, chuan mu xiang^a sichuan saussurea root, Vladiniriae Souliei Radix

☐ kg A7, chuan mu xi^a cyathula root, Cyathulae Radix

☐ kg A8, chuan wu^a prepared aconite main tuber, Aconiti Tuber

☐ kg A9, chuan xiong^a ligusticum root, Ligustici Rhizoma

☐ kg A10, chuan shu jia^a pangolin scales, Manitis Squama

☐ kg A11, chuan xin lian^a andrographis, Andrographidis Herba

☐ kg A12, chui pen cao^a hanging stonecrop, Sedi Sarmentosi Herba

☐ kg A13, chun gen pi^a toona root Bark, Toonae Cortex Ailanthi

☐ kg A14, ci shi^a loadstone, Magnetum

☐ kg A15, ci shi^a loadstone, Magnetum

☐ kg A16, ci li li^a tribulus fruit, Tribuli Fructus

Figure 6.14: An instance of seller providing a sale catalogue that includes Pangolin scales

Figure 6.15: A second instance of a seller providing a catalogue that includes pangolin scales

Figure 6.15: A second instance of a seller providing a catalogue that includes pangolin scales

The classifier also identified instances of people discussing the legality and use of pangolin scales in TCM, as shown in Figure 6.16.

The most noticeable effect of the system that had evolved from a small number of seeds and a classifier trained on the generated corpus, is the focus on TCM. With a large quantity of proposed documents being on the subject, but not always relating to treatments specifically involving pangolins. This was not something inherent to the original seed phrases or documents at the outset and became integral to the discovered domain. This came as a result of a large quantity of documents discussing treatments involving pangolins being in the results. For example, 6.17 presents one instance of a treatment which mentions their use. Information such as this is important as it helps to further define the domain and reasons why people may feel motivated to use a product taken from an endangered species.

This concludes the results section of this chapter which provides an example use case that benefits from the methodology of the CET. The final section of this chapter presents a discussion and conclusions of the project.

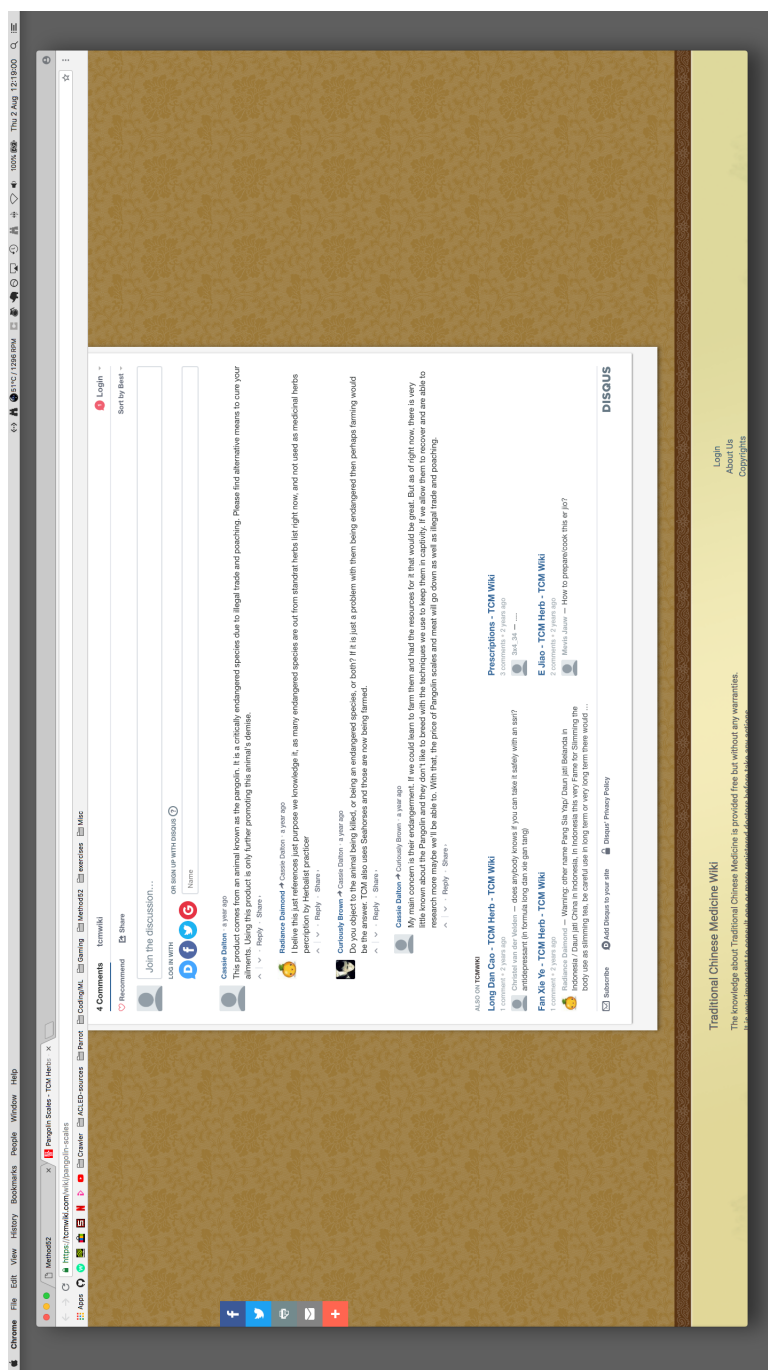


Figure 6.16: An example of those discussing pangolins



Figure 6.17: An example of sites advocating TCM treatments involving pangolin scales

Mu Xiang (Auckland Root) — 12 g
Qing Huo (Notopterygium) — 12 g
Fang Feng (Silur Root) — 12 g
Chi Shao (Red Peony) — 12 g
Kuan Jin Teng (Tinospora Stem) — 12 g
Zi Lan (Bagile Weed) — 12 g
Tao Ren (Peach Kernel) — 12 g
Mu Tong (Akebia) — 9 g
Rou Gui (Cinnamon Bark) — 9 g
Gua Lou Ren (Tricocarpites Seed) — 9 g

NOTES: Multiple variations of this popular formula exist. Herbs such as Lu Jin, Hu Gu, Bai Hua She, Qian Nian Jian, **Chuan Shou**, Hai Long, and San Leng appear in various versions while measurements of herbs such as San Qi, Bai Hua She, Tao Ren, Zi Lan Tong, and Gu Zhi are adjusted. Wu Ling Zhi, Long Xue, and Gua Lou Ren may also be substituted for other herbs. Chuan Shou, Chuan Nian Jian, and all these herbs, however, are not available in the United States. Chuan Shou is a species of the genus *Chuan Shou* and herb pack includes Bai Hua She and Hai Long in the above formula unless a specific variation is requested.

Shaolin External Iron Palm Wine
Qing Huo (Notopterygium) — 4 oz
Du Huo (Pubescent Angelica) — 4 oz
Mu Gua (Chinese Quince) — 4 oz
Da Huang (Rhubarb) — 3 oz
Chi Shao (Red Peony) — 3 oz
Wu Jia Pi (Carthagenae) — 2 oz

6.5 DISCUSSION

In this chapter the last components of the CET have been introduced. In addition, this chapter has illustrated its potential to iteratively build a system for discovering and characterising a domain, in order to generate continually expanding corpora with new content from the web. This was illustrated using the pilot Project Parrot (PP), which successfully discovered instances of people selling illicit wildlife online. One key advantage of the CET shown here is the emphasis on a human-in-the-loop strategy and the considerable number of software tools available to the analyst to more easily influence the expansion of the original corpus. This strategy and toolkit provide a means to expand or build a corpus even in cases where there is little or no knowledge of the target domain, or location on the web. The main disadvantage to this approach was the considerable amount of time required to manually analyse the search and crawl results for indicators of sale. However, this came not as a result of the methods inability to find content but the low frequency of items for sale that presented themselves, relative to the size of the collected corpus. This was compounded by the small amount of explicit and relevant information present on these sites that indicate sale, and the high variance in how instances are presented. For example, the use of code words, hidden links to items for sale, or in some cases just contact details. These are potentially very hard for any classification system to find automatically and therefore require the intervention from an analyst. The benefit to the human-in-the-loop strategy in this instance is the informed position of the analyst made possible by their high degree of involvement during the process. To mitigate for this, future work might include more attention being paid to clearly classify and identify all other categories in the domain, such as Wiki pages and conservation sites, so that all remaining documents can be scrutinised for instances of sale.

Where the CET succeeded in this instance, was its ability to provide a means of finding instances of those selling illicit wildlife products online, in spite of a relatively low frequency of their occurrence on the web when compared with more typical, legal content and items for sale. In addition, the CET proved successful in a variety of domains ranging from plant life, to animal parts originating from wildly different species and the ability of the

system to pick up on code words, such as 'mammal teeth' to identify potential sellers of illegal ivory.

One other potential avenue for future work is the inclusion of other means of classification that could be potentially more robust and yield better results. For instance, an SVM trained on a gold standard corpus labelled by a domain expert may have a higher recall and precision. One observation that was made during the course of the project was the existence of forums intended for enthusiasts to discuss and share information regarding treatments, in the case of TCM, or hobbyists discussing methods and experiences in the case of ivory. In cases such as this, it was observed that the site owners often warned people not to discuss the sale or means of purchasing illicit items, which would imply that this does occur. Another avenue of investigation might include scraping these forums for information regarding illicit sale, by utilising the capability of the CET to quickly configure web scrapers to retrieve forum posts and its capability to continually crawl a specific domain. This concludes the last case study chapter in this thesis. The final chapter discusses the conclusions of this work and details a number of future additions to the CET.

CONCLUSIONS

Over the course of the previous three chapters, five case studies have been presented and used to introduce the various components of the Corpus Expansion Toolkit (CET). The case studies each fell into one of three use case types ranging from a clearly defined domain and source of document collection, to needing to discover both the domain and location of desired documents. The main contributions of the CET have been to provide a generally applicable toolkit for researchers to build domain specific corpora from the web, starting from just a small number of example documents or phrases. This is evidenced by the variety of practical applications presented through these case studies. The CET provides a toolkit and iterative process to both discover what we want from online sources and to analyse the discovered content for desired information. This culminates into a system that provides an analyst with the capability to actively define the domain of interest, and make more informed decisions on future iterations. This iterative methods of the CET allow for a continually evolving system with a persistent state that allows for the continued search, discovering, collection and analysis of domain-specific content. The methodology of the CET places the responsibility on the analyst through a human-in-the-loop strategy, that keeps the human at the centre of the entire process. This is key to the evolution of a system that starts with little or no concept of a domain and allows the researcher to eventually become a quasi-expert in the target domain language.

The implementation of the CET within the Method52 environment provides a user friendly UI, making it easy to quickly create workflows and custom methodologies utilising any desired subset of components present in the toolkit. However, the key components of domain discovery, search, crawling and machine learning-based analysis remain at the core of the overall method. The key contributions of this work are the implementation and practical application of the complete system, the introduction of new and pre-existing methods for researchers to utilise, and the ability to keep the human-in-the-loop at all stages.

The two main questions posed of the work in this thesis are presented below.

1. Can the MAS presented in this thesis provide the means to thoroughly explore the web for domain specific content.
2. Considering the CET is a general solution to domain-specific web research, can the CET be used on a wide variety of tasks?

The practical applications of the CET over the five presented case studies evidences the efficacy of the CET as a means to discover domain specific content online and perform web-as-corpus research in digital methods. The inspiration for the CET came from an initial method of crawling and filtering discovered content using classifiers. Over the course of this thesis it has been shown that from this initial inspiration the development of the CET has created a complete and versatile tool for discovering desired content from the web across many topics and possible sources. This was demonstrated in Chapter 6, which saw the evolution of a system to discover instances of those selling illicit wildlife online. This is typically a hard problem to solve due to the sparsity of occurrence of sale across the web and the high degree of variance in their presentation. In addition, we saw that the CET is also capable of performing continual content discovery on known sources and is to extract sub-documents using simple but powerful web scraping tools as shown in Chapter 5.

One key drawback of the method has been the relatively poor accuracy of the active learning based-classifiers. However, this can be potentially explained by the high degree of variance in presentation and heterogeneous nature of online content, making fine-grained analysis hard, at least using Naive Bayes-based methods. Training domain experts to utilise these methods is also problematic, but necessary to the process when little is known of the target domain. The implementation of other classification methods is a topic of discussion in the future work, presented discussed below.

7.1 FUTURE WORK

The development process of the CET and its practical application has often illuminated its limitations and avenues for potential improvements. This final section details a number of additions to the CET that could improve its efficiency and precision. This future work is divided into three subsections. First, additions to the classification framework are discussed. Second, changes and additions to the crawling framework are covered. Third, two methods are presented for augmenting the seed generation and feature extraction methodology.

7.1.1 *Classifiers*

One area requiring significant improvement within the CET is the classification layer. The practical results of each case study that used classifiers has evidenced their effectiveness and use within the system, however it was plagued with consistently low accuracy, precision and recall. The use of Naive Bayes-based methods has also been shown to produce sub-standard results in other methods. For example, Ester et al. (2004) produce a focused crawler methodology that used a mixture of k-means with k-nearest neighbours to classify pages based on their proximity to the centroids of a given class. The choice of k-means over Naive Bayes was due to k-means consistently outperforming Naive Bayes. Although the CET contains a number of unsupervised clustering methods within its framework, such as topic models and as k-means, there remains a requirement for a more robust supervised method for defining domain relevance in the computational filtering of crawled content.

To improve performance some simple solutions would be the use of Support Vector Machines (SVM)s, such as those implemented by Barbosa and Freire (2007), which have proven to be a robust and powerful means to make discriminatory classifications over documents. Focusing research on a number of potential classification methods could provide a useful insight into optimal methods for classifying web-based content. Work in this area would of course focus primarily on overall performance in classifying web content, but additional points of interest could include the ease of training and implementation, robustness to multiple domains and amount

of training data required to attain acceptable performance. These are empirical questions that would require a significant review over multiple tasks and domains, something that the CET is well suited.

7.1.2 *Focused crawling*

In some ways the CET could be considered a partially-focused crawler as it uses search to discover relevant starting seeds, crawlers to collect content and machine learning methods to filter content at scale. However, a true focused crawler typically follows a fully automated method of choosing sites and pages to prune from the graph of web links to explore, simultaneously making the crawl more efficient and increasing the yield of relevant documents returned. Implementing a focused crawler, such as ACHE developed by Santos (2019), into the CET could potentially increase the efficiency of the system and provide an additional alternative that reduces the onus on a human-in-the-loop strategy.

7.1.3 *Web scraping*

The two means of web-scraping provide a solution to two use cases. Automatically scraping the main text content from a page when the HTML markup of the domain is unknown, or the ability for users to manually configure a scraper when the structural makeup of a site is known. The CET does not currently provide the means to automatically scrutinise pages and discover structured content. A proposed solution would be to use a method similar to that of Qiu et al. (2015), who analysed patterns of commonality in HTML markup to discover product specifications on sites selling products. An example solution synonymous with the work presented in this thesis would be that of scraping forum content. The solution would be to train a model or machine learning-based classifier on html tags and markup common to forum sites, such as forum post delineation, post date or member information. Using the CET to find forums of interest and the proposed solution would provide a fully automated system for discovering and collecting structured content from sites such as forums.

7.1.4 *Feature extraction and seed generation*

The means of extracting features, phrases and the generation of seeds follows the work of [Baroni and Bernardini \(2004\)](#) and others which look to bootstrapping corpora from the web. Although this is a proven method of seed generation it works at a surface level, and often only focuses on building corpora and extracting features from the search results provided by the Bing API. Although the CET is designed to have documents at any stage be used for extraction and search, there is the potential for more elegant methods to be incorporated within the framework. For instance, [Ester et al. \(2004\)](#) implemented a two level means of internal and external crawling to discover domain-specific sites. Adapting this method to perform internal crawls of sites found by a search could be used to quickly generate more relevant documents for feature extraction and corpus bootstrapping. Assuming that the originating site of a page found by a web search may contain more relevant content, the search results could be expanded prior to return through an internal crawl and domain discovery.

7.1.5 *Search*

The reasons for using the Bing API in the CET was practical in nature as very few search services provide a publicly available API to access their search index. However, the implementation of ranking algorithms differ between companies and it is safe to assume that different pages will have been indexed and more importantly, some missed. Another issue with using the Bing API is that there is a small financial cost to using the service. To mitigate for these problems it would be desirable to have a number of search options available, that are preferable free of charge. For instance, the DuckDuck Instant Answers API provides a means to send queries and have a number of topically related search results returned ([DuckDuckGo, 2019](#)). Although these results are often taken from common sources such as Wikipedia (their full index is unavailable for legal reasons), this API could be used to help disambiguate a topic or query. The intention would be to provide as many search options as possible to increase the variety of possible results.

7.1.6 *User interface*

Although Method52 provides a UI to implement a system for discovering domain-specific documents, and a number of tools for their analysis, it lacks the ability to explore datasets in a more organic manner. For instance, it provides the means to perform topic modelling and clustering of a collection, whilst outputting a list of words characteristic to the discovered clusters. It does not however, provide a visual representation of the clusters or the cluster space. An envisaged solution might look similar to that of [Krishnamurthy et al. \(2016\)](#) who present the Domain Discovery Tool, which has a primary focus on the user's ability to visually see and explore a collection at scale. It achieves this by using a method of dimensionality scaling over clusters, that was originally proposed in LDAvis ([Sievert and Shirley, 2014](#)). The proposed solution would be presented as an entirely separate system from Method52, that would act as a means to more efficiently present a visualisation of an entire collection; whilst providing the means to view documents of interest, or summaries of topics and clusters. An interactive visualisation method of this sort has the potential to better inform the analyst at the heart of the system, and places them in a more informed position to develop the domain representation. This concludes the proposed future work and this thesis.

Appendices

DIASPORA VOTING RIGHTS SITE LIST

<http://aceproject.org/ace-en/topics/va/onePage>
http://www.idea.int/elections/vfa_search.cfm
<http://allafrica.com>
<http://thinkafricapress.com>
<http://diasporaalliance.org>
<https://www.overseasvotefoundation.org/vote/home.htm>
<http://www.nigeriansinamerica.com>
<http://www.nigeriavillagesquare.com>
<http://www.canuk.org.uk>
<http://nido.net>
<http://www.afford-uk.org>
<http://sand-uk.org>
<http://www.nigerianwatch.com>
<http://newtelegraphonline.com>
<http://www.elombah.com>
<http://www.thisdaylive.com>
<http://diasporacommittee.com>
<http://diasporavotefornigerians.org>

ACLED QUERY TERMS

conflict
fighting
aggression
attack
attacks
conflicts
kill
kills
killed
massacre
bomb
bombed
bombing
bombs
dead
died
rebel
attacked
riot
battle
protest
clash
violent
demonstration
strike
wound
injure
casualty
casualties
vigilante
war
torture
displaced

PROVIDED PANGOLIN SEARCH PHRASES

Pangolin

Manis (genus)

Squama (this and the two terms immediately below – used to describe pangolin scales in TCM)

Squama manis

Squama manitis

Incense sticks

Nagi (how scales are referred to as an ingredient in incense sticks)

Chuan Shan Jia (pinyin)

Chuanshanjia keli (pinyin)

Xuyen Son Giap (English version of Vietnamese for pangolin scales)

Plastic pieces (some labelled as in illicit trade)

Plastic sheets (some labelled as in illicit trade)

Armadillo counter poison pill

LIST OF FIGURES

Figure 1.1	The Corpus Expansion Toolkit data flow diagram	17
Figure 2.1	The BootCaT flow diagram	24
Figure 2.2	The DISCO algorithm	32
Figure 2.3	The focused crawler expansion model	40
Figure 2.4	The ACHE crawler	42
Figure 2.5	The WATES multiagent system	45
Figure 2.6	The Talaia component view	46
Figure 2.7	The AutoNet process	48
Figure 2.8	The U.I for the DDT	49
Figure 3.1	The data view within Method52	62
Figure 3.2	The job view within Method52	63
Figure 3.3	A single component instantiated within Method52	64
Figure 3.4	Two connected components within Method52	66
Figure 3.5	A complete pipeline within Method52	67
Figure 3.6	The Web Crawler component configuration view within Method52	68
Figure 4.1	The Semi-automated Web-search and Analytics Tool	75
Figure 4.2	The label selection screen	77
Figure 4.3	The classifier metrics view	77
Figure 4.4	The crawl page of SAWSAT	82
Figure 4.5	The search page of the SAWSAT	84
Figure 4.6	The document analysis page of the SAWSAT .	86
Figure 4.7	The document view of the SAWSAT	87
Figure 4.8	The database view of the SAWSAT	89
Figure 5.1	The flow diagram of the utilised sub-system of the Corpus Expansion Toolkit	102
Figure 5.2	The configurable options for the Automatic Extractor component	105
Figure 5.3	An example web page and corresponding html markup	109
Figure 5.4	Web Scraper component configuration within Method52	110
Figure 5.5	Web Scraper component test area within Method52	111
Figure 5.6	Flow diagram of the continuous crawler	112
Figure 5.7	Configuration screen of the continuous crawler	116
Figure 6.1	The Corpus Expansion Toolkit	135
Figure 6.2	The configuration parameters for the Surprising Phrase Detector component	139
Figure 6.3	The configuration screen for the Web Search component	142

Figure 6.4	A complete pipeline that will perform a web search using the input queries.	143
Figure 6.5	An image of the configuration parameters for the Feature Combiner component	143
Figure 6.6	The Web Crawler configuration options	144
Figure 6.7	An instance of a seller providing a link to products containing pangolin scales	149
Figure 6.8	An example web page selling Armadillo Counter Poison	150
Figure 6.9	An example of a page selling products with pangolin scales removed.	151
Figure 6.10	An instance of a site selling 'pongolin' scales .	152
Figure 6.11	An instance of a seller obfuscating the ingredients of the product.	154
Figure 6.12	An instance of groups discussing TCM treatments	156
Figure 6.13	Examples of NGOs and activists interested in saving the pangolin	157
Figure 6.14	An instance of seller providing a sale catalogue that includes Pangolin scales	159
Figure 6.15	A second instance of a seller providing a catalogue that includes pangolin scales	160
Figure 6.16	An example of those discussing pangolins . .	162
Figure 6.17	An example of sites advocating TCM treatments involving pangolin scales	163

LIST OF TABLES

Table 5.1	Average number of users per forum	120
Table 5.2	Table of forum collection statistics	120
Table 5.3	Count of organisation mentions across all forums	124
Table 6.1	Table of keyword categories and % of corpus recovered	147

BIBLIOGRAPHY

- F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Twitcident: Fighting Fire with Information from Social Web Streams. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion*, pages 305–308, New York, NY, USA, 2012. ACM. (Cited on pages [5](#), [44](#), and [81](#).)
- R. Agerri, J. Bermudez, and G. Rigau. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3823–3828, Reykjavik, Iceland, May 2014. European Languages Resources Association (ELRA). (Cited on page [45](#).)
- L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing Trending Topics in Twitter. *Sensing Trending Topics in Twitter*, 15 (6):1268 – 1282, October 2013. (Cited on page [44](#).)
- P. Arora and T. Bhalla. A Synonym Based Approach of Data Mining in Search Engine Optimization. *A Synonym Based Approach of Data Mining in Search Engine Optimization*, 12(4):201–205, June 2014. (Cited on page [28](#).)
- W. Bakari, P. Bellot, and M. Neji. A logical representation of Arabic questions toward automatic passage extraction from the Web. *A Logical Representation of Arabic Questions Toward Automatic Passage Extraction From the Web*, 20(2):339–353, April 2017. (Cited on page [103](#).)
- L. Barbosa and J. Freire. Searching for Hidden-Web Databases. In *WebDB*, pages 1–6, 2005. (Cited on page [41](#).)
- L. Barbosa and J. Freire. An Adaptive Crawler for Locating hidden-Web Entry Points. In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 441–450, New York, NY, USA, 2007. ACM. event-place: Banff, Alberta, Canada. (Cited on pages [41](#), [42](#), [49](#), and [168](#).)
- L. Barbosa, S. Bangalore, and V. K. Rangarajan Sridhar. Crawling Back and Forth: Using Back and Out Links to Locate Bilingual Sites. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 429–437, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing. (Cited on page [41](#).)
- M. Baroni and S. Bernardini. BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC 2004*, volume 1313,

- Lisbon, Portugal, 2004. (Cited on pages 3, 7, 23, 24, 25, 26, 27, 28, 29, 38, 138, 139, 140, and 170.)
- M. Baroni and A. Kilgarriff. Large Linguistically-Processed Web Corpora for Multiple Languages. In *Demonstrations*, 2006. (Cited on page 24.)
- M. Baroni and M. Ueyama. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the NIJL International Workshop on Language Corpora*, pages 31–40, Tokyo, Japan, 2006. (Cited on pages 7, 24, 38, 128, and 138.)
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*, 43(3):209–226, September 2009. (Cited on pages 6, 7, 24, 26, and 39.)
- M. Bauer and G. Gaskell. *Qualitative Researching With Text, Image and Sound*. SAGE Publications Ltd, 6 Bonhill Street, London England EC2A 4PU United Kingdom, 2000. (Cited on page 38.)
- Bbc. ‘Shambles’ as Sports Direct’s results delayed, July 2019. (Cited on page 106.)
- N. Bel, V. Papavassiliou, P. Prokopidis, A. Toral, and V. Arranz. Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform. In *The 5th Workshop on Building and Using Comparable Corpora*, volume abs/1303.1932, 2013. (Cited on page 41.)
- V. Benko. Aranea: Yet Another Family of (Comparable) Web Corpora. In *TSD 2014*, volume 8655, pages 247–254. Springer International Publishing, 2014. (Cited on page 26.)
- C. Biemann, F. Bildhauer, S. Evert, D. Goldhahn, U. Quasthoff, R. Schäfer, J. Simon, L. Swiezinski, and T. Zesch. Scalable Construction of High-Quality Web Corpora. *Scalable Construction of High-Quality Web Corpora*, 28(2):23–59, 2013. (Cited on page 38.)
- Bixo. Open Source Web Mining Toolkit - Bixo, September 2019. (Cited on page 41.)
- D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Latent Dirichlet Allocation*, 3:993–1022, 2003. Publisher: JMLR.org. (Cited on page 145.)
- A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic Clustering of the Web. *Syntactic Clustering of the Web*, 29(8-13): 1157–1166, September 1997. (Cited on page 138.)
- M. Bron, K. Balog, and M. De Rijke. Example Based Entity Search in the Web of Data. In P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz,

- editors, *Advances in Information Retrieval*, pages 392–403, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. (Cited on page 34.)
- J. Cho and H. Garcia-Molina. Effective page refresh policies for Web crawlers. *Effective Page Refresh Policies for Web Crawlers*, 28(4): 390–426, December 2003a. (Cited on pages 36, 114, and 115.)
- J. Cho and H. Garcia-Molina. Estimating Frequency of Change. *Estimating Frequency of Change*, 3(3):256–290, August 2003b. (Cited on page 37.)
- M. Coanca and E. Museanu. Electronic Corpora In Translation Bootcat-Bootstrapping Corpora And Terms From The Web. *Electronic Corpora In Translation Bootcat-Bootstrapping Corpora And Terms From The Web*, 4(1):64–75, May 2010. (Cited on page 26.)
- J. Daniels, K. Gregory, and T. M. Cottom. *Digital Sociologies / Edited by Jessie Daniels, Karen Gregory, Tressie McMillan Cottom*. Policy Press Bristol, UK ; Chicago, IL, 2017. (Cited on pages 1 and 12.)
- Z. Davey, F. Schifano, O. Corazza, P. Deluca, and o. B. o. t. P. W. M. Group. E-Psychonauts: Conducting research in online drug forum communities. *E-Psychonauts: Conducting Research in Online Drug Forum Communities*, 21(4):386–394, 2012. (Cited on page 3.)
- E. Di Minin, C. Fink, H. Tenkanen, and T. Hiippala. Machine learning for tracking illegal wildlife trade on social media. *Machine Learning for Tracking Illegal Wildlife Trade on Social Media*, 2(3):406–407, 2018. (Cited on page 132.)
- DuckDuckGo. DuckDuckGo Instant Answer API, September 2019. (Cited on page 170.)
- P. Essiembre. Norconex, July 2019. (Cited on page 144.)
- M. Ester, H.-P. Kriegel, and M. Schubert. Accurate and Efficient Crawling for Relevant Websites. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases*, volume 30 of VLDB '04, pages 396–407. VLDB Endowment, 2004. event-place: Toronto, Canada. (Cited on pages 39, 74, 168, and 170.)
- B. Everitt. *The Analysis of Contingency Tables*. Chapman and Hall, London, New York, 1992. Open Library ID: OL4538108M. (Cited on page 24.)
- B. Faltings, P. Pu, and P. Viappiani. Preference-based Search using Example-Critiquing with Suggestions. *Preference-Based Search Using Example-Critiquing With Suggestions*, 27(1):465–503, 2006. (Cited on page 33.)
- H. Fang, T. Tao, and C. Zhai. A Formal Study of Information Retrieval Heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information*

- Retrieval*, SIGIR '04, pages 49–56, New York, NY, USA, 2004. ACM. event-place: Sheffield, United Kingdom. (Cited on page 30.)
- C. Fantinuoli. Specialized Corpora from the Web and Terms Extraction for Simultaneous Interpreters. In *Working papers on the Web as Corpus*, Gedit, Bologna, 2005. (Cited on page 26.)
- A. Ferraresi. Google and beyond : Web-as-corpus methodologies for translators. *Google and Beyond : Web-As-Corpus Methodologies for Translators*, 7, 2009. (Cited on pages 13 and 28.)
- J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A Quantitative Approach to Understanding Online Antisemitism. *A Quantitative Approach to Understanding Online Antisemitism*, abs/1809.01644, September 2018. (Cited on page 4.)
- D. Fiser, N. Ljubesic, S. Vintar, and S. Pollak. Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26, Portland, Oregon, June 2011. Association for Computational Linguistics. (Cited on page 26.)
- D. Ford, C. Grimes, and E. Tassone. Keeping a Search Engine Index Fresh: Risk and optimality in estimating refresh rates for web pages. In *Proceedings of INTERFACE 2008*, page 14, 2008. (Cited on page 37.)
- W. Francis and H. Kucera. A Standard Corpus of Present-Day Edited American English, for use with Digital Computers (Brown). Technical report, Brown University, Providence, Rhode Island., 1964. (Cited on page 24.)
- Q. Gao and S. Vogel. Corpus Expansion for Statistical Machine Translation with Semantic Role Label Substitution Rules. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 294–298, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. (Cited on page 26.)
- E. R. Guevara. NoWaC: A large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, Los Angeles, 2010. Association for Computational Linguistics. event-place: NAACL-HLT, Los Angeles. (Cited on page 38.)
- D. Hamid and S. Hassan. Web Pages Retrieval by Using Proposed Focused Crawler. *Web Pages Retrieval by Using Proposed Focused Crawler*, 19:154–164, 2016. (Cited on page 41.)
- H. B. Hashemi and A. Shakery. Mining a Persian-English Comparable Corpus for Cross-language Information Retrieval. *Mining a*

- Persian-English Comparable Corpus for Cross-Language Information Retrieval*, 50(2):384–398, March 2014. (Cited on page 29.)
- M. A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, volume 2, pages 23–28, Nantes, 1992. (Cited on page 23.)
- J. Hedley. Jsoup Java HTML Parser, with best of DOM, CSS, and jquery, July 2019. (Cited on page 106.)
- J. Hernandez-Castro and D. L. Roberts. Automatic detection of potentially illegal online sales of elephant ivory via data mining. *Automatic Detection of Potentially Illegal Online Sales of Elephant Ivory via Data Mining*, 1(10):1–11, July 2015. (Cited on page 131.)
- A. Hinsley and D. L. Roberts. The wild origin dilemma. *The Wild Origin Dilemma*, 217:203 – 206, 2018. (Cited on pages 129 and 131.)
- A. Imani, A. Vakili, A. Montazer, and A. Shakery. An Axiomatic Study of Query Terms Order in Ad-Hoc Retrieval. In L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, editors, *Advances in Information Retrieval*, pages 196–202, Cham, 2019. Springer International Publishing. (Cited on page 30.)
- G. Initiative. Global Initiative - A Network to Counter Networks, July 2019. (Cited on page 56.)
- Insight. The Linked Open Data Cloud, September 2019. (Cited on page 34.)
- M. Jakubíček, A. Kilgarrieff, V. Kovář, P. Rychlý, and V. Suchomel. The TenTen Corpus Family. *The TenTen Corpus Family*, 2013. (Cited on pages 6 and 26.)
- D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Recommender Systems: An Introduction*. Cambridge University Press, New York, NY, USA, 1st edition, 2010. (Cited on page 33.)
- M.-M. Kenning and M. McCarthy. *The Routledge Handbook of Corpus Linguistics*. Routledge, 2010. (Cited on page 26.)
- A. Kilgarrieff. WebBootCaT: A web tool for instant corpora. In *Proceeding of the EuraLex Conference*, volume 9, pages 123–132, Italy, September 2006. Edizioni dell’Orso s.r.l. (Cited on pages 26 and 27.)
- A. Kilgarrieff, P. Rychly, P. Smrz, and D. Tugwell. The Sketch Engine, April 2019. (Cited on page 26.)
- C. Kohlschütter. Boilerpipe, August 2019. original-date: 2014-12-01Too:18:23Z. (Cited on page 103.)

- C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, page 441, New York, New York, USA, 2010. ACM Press. (Cited on pages 103 and 104.)
- Y. Krishnamurthy, K. Pham, A. Santos, and J. Freire. Interactive Exploration for Domain Discovery on the Web. In *ACM KDD Workshop on Interactive Data Exploration and Analytics (IDEA)*, pages 64–71, San Francisco, CA, USA, 2016. ACM. (Cited on pages 9, 48, 50, 57, 81, 88, and 171.)
- M. Kunder. WorldWideWebSize.com - The size of the World Wide Web (The Internet), July 2019. (Cited on page 1.)
- D. Lange, C. Böhm, and F. Naumann. Extracting Structured Information from Wikipedia Articles to Populate Infoboxes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, pages 1661–1664, New York, NY, USA, 2010. ACM. event-place: Toronto, ON, Canada. (Cited on page 29.)
- R. Leaman, L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang, and G. Gonzalez. Towards Internet-age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-related Social Networks. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP '10*, pages 117–125, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. event-place: Uppsala, Sweden. (Cited on page 3.)
- I. Leturia and I. S. Vicente. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet. In *Proceedings of the 5th Web As a Corpus*, 2009. (Cited on pages 23 and 29.)
- Y. Li, Y. Wang, and J. Du. E-Ffc: An enhanced form-focused crawler for domain-specific deep web databases. *E-Ffc: An Enhanced Form-Focused Crawler for Domain-Specific Deep Web Databases*, 40(1): 159–184, 2012. (Cited on page 41.)
- B. Libraries. The British National Corpus, 2007. (Cited on pages 6, 27, and 39.)
- J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining Quality Phrases from Massive Text Corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data - SIGMOD '15*, pages 1729–1744, Melbourne, Victoria, Australia, 2015a. ACM Press. (Cited on page 47.)
- L. Liu, T. Peng, and W. Zuo. Topical Web Crawling for Domain-Specific Resource Discovery Enhanced by Selectively using Link-Context. *Topical Web Crawling for Domain-Specific*

- Resource Discovery Enhanced by Selectively Using Link-Context*, 12(2): 196–204, 2015b. (Cited on page 38.)
- Z. Liu. Optimal robot scheduling for Web search engines. Technical report, 1998. (Cited on pages 36 and 37.)
- H. N. Lucas D. Introna. Shaping the Web: Why the Politics of Search Engines Matters. *Shaping the Web: Why the Politics of Search Engines Matters*, 16(3):169–185, 2000. (Cited on page 13.)
- N. Marres. *Digital Sociology: The Reinvention of Social Research*. Wiley, 2017. (Cited on pages 12 and 13.)
- N. Marres and D. Moats. Mapping Controversies with Social Media: The Case for Symmetry. *Mapping Controversies With Social Media: The Case for Symmetry*, 1(2), 2015. (Cited on page 4.)
- D. McCarthy. DANTE : A New Resource for Research at the Syntax-Semantics Interface. In *Proceedings of Interdisciplinary Workshop on Verbs*, Pisa, Italy, 2010. (Cited on page 27.)
- O. Medelyan, S. Schulz, J. Paetzold, M. Poprat, and K. G. Markó. Language Specific and Topic Focused Web Crawling. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 865–868, Genoa, Italy, 2006. (Cited on pages 37 and 41.)
- L. Medrouk, A. Pappa, and J. Hallou. Review Web Pages Collector Tool for Thematic Corpus Creation. In *CILC2016. 8th International Conference on Corpus Linguistics*, volume 1, pages 274–264, 2016. (Cited on pages 8, 38, 39, and 103.)
- U. of Ljubljana. Orange – Data Mining Data Mining Fruitful and Fun, July 2019. (Cited on page 131.)
- A. Oliveira. Ontology supported system for searching evidence of wild animals trafficking in social network posts. *Ontology Supported System for Searching Evidence of Wild Animals Trafficking in Social Network Posts*, 6:16, 2014. (Cited on pages 2, 5, 43, 44, and 132.)
- C. Olston and M. Najork. Web Crawling. *Web Crawling*, 4(3):175–246, March 2010. (Cited on pages 35 and 36.)
- C. Olston and S. Pandey. Recrawl scheduling based on information longevity. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 437, Beijing, China, 2008. ACM Press. (Cited on pages 36 and 37.)
- L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, November 1999. (Cited on pages 13 and 38.)

- S. Pandey and C. Olston. User-centric Web Crawling. In *Proceedings of the 14th International Conference on World Wide Web, WWW '05*, pages 401–411, New York, NY, USA, 2005. ACM. event-place: Chiba, Japan. (Cited on page 37.)
- N. G. Patel, C. Rorres, D. O. Joly, J. S. Brownstein, R. Boston, M. Z. Levy, and G. Smith. Quantitative methods of identifying the key nodes in the illegal wildlife trade network. *Quantitative Methods of Identifying the Key Nodes in the Illegal Wildlife Trade Network*, 112(26): 7948–7953, 2015. (Cited on page 130.)
- K. Pham, A. Santos, and J. Freire. Learning to Discover Domain-Specific Web Content. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18*, pages 432–440, New York, NY, USA, 2018. ACM. event-place: Marina Del Rey, CA, USA. (Cited on page 37.)
- K. Pham, A. S. R. Santos, and J. Freire. Bootstrapping Domain-Specific Content Discovery on the Web. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, volume abs/1902.09667, pages 1476–1486, San Francisco, CA, USA, 2019. Association for Computing Machinery, Inc. (Cited on pages 2, 31, 32, 33, 38, and 58.)
- M. Prensky. Digital Natives, Digital Immigrants Part 1. *Digital Natives, Digital Immigrants Part 1*, 9(5):1–6, 2001. (Cited on page 1.)
- P. H. Z. Pu and P. Kumar. Evaluating Example-based Search Tools. In *Proceedings of the 5th ACM Conference on Electronic Commerce, EC '04*, pages 208–217, New York, NY, USA, 2004. ACM. event-place: New York, NY, USA. (Cited on page 33.)
- A. PVS, D. McCarthy, D. Glennon, and J. P. And. Domain Specific Corpora from the Web. In R. V. Fjeld and J. M. Torjusen, editors, *Proceedings of the 15th EURALEX International Congress*, pages 336–342, Oslo, Norway, 2012. Department of Linguistics and Scandinavian Studies, University of Oslo. (Cited on page 27.)
- D. Qiu, L. Barbosa, X. L. Dong, Y. Shen, and D. Srivastava. Dexter: Large-scale Discovery and Extraction of Product Specifications on the Web. *Dexter: Large-Scale Discovery and Extraction of Product Specifications on the Web*, 8(13):2194–2205, September 2015. (Cited on pages 39, 74, 81, 100, and 169.)
- X. Qiu, C. Huang, and X. Huang. Automatic Corpus Expansion for Chinese Word Segmentation by Exploiting the Redundancy of Web Information. In *the 25th International Conference on Computational Linguistics*, pages 1154–1164, Dublin, 2014. (Cited on pages 26 and 27.)
- R. Rahimi, A. Shakery, J. Dadashkarimi, M. Ariannezhad, M. Dehghani, and H. N. Esfahani. Building a multi-domain

- comparable corpus using a learning to rank method†. *Building a Multi-Domain Comparable Corpus Using a Learning to Rank Method†*, 22(4):627–653, July 2016. (Cited on pages 6 and 29.)
- S. V. K. Raja and F. K. Akorli. A Survey of Web Crawler Algorithms. *A Survey of Web Crawler Algorithms*, 8(6):1694–1814, November 2011. (Cited on page 36.)
- R. Rapp, S. Sharoff, and P. Zweigenbaum. Recent advances in machine translation using comparable corpora. *Recent Advances in Machine Translation Using Comparable Corpora*, 22(4):501–516, July 2016. (Cited on pages 6 and 26.)
- S. Remus and C. Biemann. Domain-Specific Corpus Expansion with Focused Webcrawling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3607–3611, Portorož, Slovenia, 2016. European Language Resources Association (ELRA). (Cited on pages 9, 39, 40, 41, and 74.)
- A. Reynolds, B. Reilly, and A. Ellis. ELECTORAL SYSTEM DESIGN : The New International IDEA Handbook. Technical report, IDEA, June 2005. (Cited on pages 69, 70, and 71.)
- A. D. Robertson. *Characterising Semantically Coherent Classes of Text Through Feature Discovery*. PhD Thesis, University of Sussex, 2019. (Cited on pages 26, 126, and 138.)
- R. Rogers. The googlization question: Towards the inculpable engine? In *Deep search: the politics of search beyond Google*, pages 173–184. StudienVerlag: Innsbruck, 2009. (Cited on page 13.)
- R. Rogers. *Digital Methods*. The MIT Press, 1 edition, 2013. (Cited on page 14.)
- G. E. Rosen and K. F. Smith. Summarizing the Evidence on the International Trade in Illegal Wildlife. *Summarizing the Evidence on the International Trade in Illegal Wildlife*, 7(1):24–32, August 2010. (Cited on page 132.)
- E. Sandhaus. The New York Times Annotated Corpus - Linguistic Data Consortium, October 2008. (Cited on pages 34 and 81.)
- M. Santini, A. Jönsson, W. Strandqvist, G. Cederblad, M. Nyström, M. Alirezaie, L. Lind, E. Blomqvist, M. Lindén, and A. Kristoffersson. Designing an Extensible Domain-Specific Web Corpus for “Layfication”: A Case Study in eCare at Home. In *Designing an Extensible Domain-Specific Web Corpus for “Layfication”: A Case Study in eCare at Home*, pages 98–155. 2019. (Cited on pages 3 and 26.)
- A. Santos. ACHE Crawler, August 2019. (Cited on page 169.)

- K. P. Scannell. The Crúbadán Project: Corpus building for under-resourced languages. In *WAC-3*, Louvain-la-Neuve, Belgium, 2007. (Cited on pages 6 and 26.)
- R. Schäfer and F. Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey, 2012. European Languages Resources Association (ELRA). (Cited on page 39.)
- B. Settles. Closing the Loop: Fast, Interactive Semi-Supervised Annotation With Queries on Features and Instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. (Cited on page 76.)
- J. Shang, J. Liu, M. Jiang, X. Ren, C. R. Voss, and J. Han. Automated Phrase Mining from Massive Text Corpora. *Automated Phrase Mining From Massive Text Corpora*, 30(10):1825–1837, October 2018a. (Cited on page 47.)
- J. Shang, Q. Zhu, J. Shen, X. Wang, X. Gu, L. M. Kaplan, T. Harratty, J. Han, Lance, and J. D. Kaplan. AutoNet : Automated Network Construction and Exploration System from Domain-Specific Corpora. In *KDD'18, August, 2018, London, UK*, London, 2018b. (Cited on page 47.)
- Y. Shoham. Learning to Surf: Multiagent Systems for Adaptive Web Page Recommendation. Technical report, Stanford, CA, USA, 1998. (Cited on page 33.)
- C. Sievert and K. Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. (Cited on pages 140 and 171.)
- J. Singh, W. Nejdl, and A. Anand. Expedition: A Time-Aware Exploratory Search System Designed for Scholars. In *SIGIR '16 Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, volume abs/1810.10769, pages 1105–1108, Pisa, Italy, 2018. ACM. (Cited on page 34.)
- J. Smith, J. Bartlett, D. Buck, and M. Honeyman. Online Support: Investigating the role of public online forums in mental health. Technical report, Demos, April 2017. (Cited on page 94.)
- N. Spanoudakis and P. Moraitis. The Agent Systems Methodology (ASEME): A Preliminary Report. In *Proceedings of the 5th European Workshop on Multi-Agent Systems*, Hammamet, Tunisia, 2007. (Cited on page 43.)

- A. Spink, D. Wolfram, J. Jansen, and T. Saracevic. Searching the Web: The Public and Their Queries. *Searching the Web: The Public and Their Queries*, 52:226 – 234, 2001. (Cited on pages 34 and 35.)
- Swhite-Msft. Bing Entity Search API v7 Reference, August 2019. (Cited on page 141.)
- Traffic. TRAFFIC and Tencent sign agreement to tackle illicit wildlife trade through social media networks. *TRAFFIC and Tencent Sign Agreement to Tackle Illicit Wildlife Trade Through Social Media Networks*, November 2015. (Cited on page 132.)
- A. Törnberg and P. Törnberg. Muslims in social media discourse: Combining topic modeling and critical discourse analysis. *Muslims in Social Media Discourse: Combining Topic Modeling and Critical Discourse Analysis*, 13:132 – 142, 2016. (Cited on page 4.)
- I. S. Vicente, X. Saralegi, and R. Agerri. Talaia: A Real time Monitor of Social Media and Digital Press. *Talaia: A Real Time Monitor of Social Media and Digital Press*, abs/1810.00647, 2018. (Cited on pages 2, 5, 44, 46, and 133.)
- K. Vieira, L. Barbosa, A. S. Da Silva, J. Freire, and E. Moura. Finding seeds to bootstrap focused crawlers. *Finding Seeds to Bootstrap Focused Crawlers*, 19(3):449–474, May 2016. (Cited on pages 8 and 39.)
- M. Weisfeld. *The Object-Oriented Thought Process*. Addison-Wesley Professional, 4th edition, 2013. (Cited on page 43.)
- S. Wibberley, D. Weir, and J. Reffin. Language Technology for Agile Social Media Science. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 36–42, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. (Cited on page 59.)
- S. Wibberley, D. Weir, and J. Reffin. Method51 for Mining Insight from Social Media Datasets. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 115–119, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. (Cited on page 59.)
- G. Wittemyer, J. M. Northrup, J. Blanc, I. Douglas-Hamilton, P. Omondi, and K. P. Burnham. Illegal killing for ivory drives global decline in African elephants. *Illegal Killing for Ivory Drives Global Decline in African Elephants*, 111(36):13117–13121, 2014. (Cited on page 5.)
- J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman, and L. Ozsen. Optimal Crawling Strategies for Web Search Engines. In *Proceedings of the 11th International Conference on World Wide Web, WWW '02*,

- pages 136–147, New York, NY, USA, 2002. ACM. event-place: Honolulu, Hawaii, USA. (Cited on page [37](#).)
- M. Woolridge and M. J. Wooldridge. *Introduction to Multiagent Systems*. John Wiley & Sons, Inc., New York, NY, USA, 2001. (Cited on pages [9](#) and [15](#).)
- L. M. Yeo, R. S. McCrea, and D. L. Roberts. A novel application of mark-recapture to examine behaviour associated with the online trade in elephant ivory. In *PeerJ*, volume 5, page e3048, March 2017. (Cited on pages [130](#) and [131](#).)
- X. Yu and W. Jia. Moving Targets: Tracking Online Sales Of Illegal Wildlife Products In China, 2015. (Cited on page [130](#).)