2015-09-30

A Novel and Domain-Specific Document Clustering and Topic Aggregation Toolset for a News Organisation

Claire McMahon

Technological University Dublin

Recommended Citation

McMahon, C. (2015) A Novel and Domain-Specific Document Clustering and Topic Aggregation Toolset for a News Organisation, Masters Dissertation, 2015.

A Novel and Domain-Specific Document Clustering and Topic Aggregation Toolset for a News Organisation

Claire McMahon

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Knowledge Management)

July 2015

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Knowledge Management), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed : Dave McMahon

Date: ***04 July 2015***

ABSTRACT

Large collections of documents are becoming increasingly common in the news gathering industry. A review of the literature shows there is a growing interest in data-driven journalism and specifically that the journalism profession needs better tools to understand and develop actionable knowledge from large document sets. On a daily basis, journalists are tasked with searching a diverse range of document sets including news gathering services, emails, freedom of information requests, court records, government reports, press releases and many other types of generally unstructured documents. Document clustering techniques can help address problems of understanding the ever expanding quantities of documents available to journalists by finding patterns within documents. These patterns can be used to develop useful and actionable knowledge which can contribute to journalism. News articles in particular are fertile ground for document clustering principles. Term weighting schemes assign importance to terms within a document and are central to the study of document clustering methods. This study contributes a review of the dominant and most commonly used term frequency weighting functions put forward in research, establishes the merits and limitations of each approach, and proposes modifications to develop a news-centric document clustering and topic aggregation approach. Experimentation was conducted on a large unstructured collection of newspaper articles from the Irish Times to establish if the newly proposed news-centric term weighting and document similarity approach improves document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach. Whilst the experimentation shows that that the developed approach is promising when compared to the manual document clustering effort undertaken by the three journalist expert users, it also highlights the challenges of natural language processing and document clustering methods in general. The results may suggest that a blended approach of complimenting automated methods with human-level supervision and guidance may yield the best results.

Key words: *Document Clustering, Term Weighting, TF-IDF, Topic Aggregation, Data-Driven Journalism*

ACKNOWLEDGEMENTS

I owe my supervisor Deirdre Lawless a debt of gratitude for the considerable support, advice and guidance throughout this work. Thank you for your calmness and for keeping your head when all about you were losing theirs.

I would like extend a thank you to all of the staff at D.I.T. I have had a very positive experience studying here as both an undergraduate and postgraduate student.

Thank you to all of my colleagues in the Irish Times for their support during this project. A special thank you to George Walker for his invaluable suggestions that helped me to identify an approach to fix the memory leaks in the software and for the very useful books on Information Retrieval. A massive thank you to Luke Cassidy, Paddy Logue and David Labanyi for their participation as expert journalist users and for the time they spent manually clustering a large number of documents. I appreciate these tasks were incorporated into an already hectic schedule.

Thank you to my parents John and Agnes McMahon for their constant support and encouragement over the years.

To my daughter Evie, I owe you lots of one on one Mummy time.

Finally, to Dave for his understanding and support while I worked on this dissertation. I promise to give you your weekends back.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF EQUATIONS

1. INTRODUCTION

1.1 Background

Large document sets have become a significant part of the journalism profession (Gray, Chambers and Bounegru, 2013; Brehmer, Ingram, Stray and Muzner, 2014). On a daily basis, journalists are tasked with searching a diverse range of document sets including documents from open government initiatives, news gathering services, freedom of information requests, WikiLeaks documents, court records, social networking data and many other types of large and generally unstructured document sets (Brehmer et al, 2014). Documents of this nature often contain interesting and newsworthy material (Brehmer et al, 2014) but eliciting actionable knowledge represents a challenge due to the sheer number and size of documents in the collection, the time pressures that journalists are under and the limited number of human resources available. The buzzword "big-data" is often invoked when discussing data-driven journalism. The term aims to describe a volume of digital data that is so overwhelming, traditional editorial and data processing techniques are struggling to adapt. A key challenge for the journalism profession therefore, is addressing the accessibility of large document sets and the extraction of value from them.

Working with and understanding large document sets is a specific concern for the journalism profession because of the watchdog role that journalists provide in a democratic society. Sometimes referred to as the fourth estate or fourth power, news media organisations aim to operate as independent entities, protecting public interests and monitoring the activities of governments (Miragliota, Errington and Barry, 2011). News organisations are some of the chief purveyors of information in society and as such are tasked with making the affairs of the government and powerful institutions more transparent to the public by providing access to independent and verifiable information. Journalists aim to hold powerful interests to account and to inform the public about the practices of institutions in society. Data is playing an increasingly significant role in facilitating news media organisations in carrying out this watchdog role in society. Data is steadily being used to source and investigate news stories, to tell complicated stories, to reveal trends, to verify information independently and assist the journalism profession in making hypothesis (Gray et al, 2013; Ridgway and Smith 2013; Howard, 2014). By using data, the role of the journalist shifts its main focus from the being the first to report a story (Gilmore, 2008) to one of providing a deeper insight into what a certain development may mean (McCandles, 2013). Data, big or otherwise, can help journalists to stimulate debate and set the political agenda through the disclosure of public issues.

Keyword search querying has emerged as one of the most traditional and effective paradigms for information retrieval (Tagarelli, 2011). Han and Chan (2002) however, contend that it suffers from several inadequacies. A search can return many answers; especially if the keywords are common across categories and the semantic meaning of terms diverge in different contexts. Mahalakshmi (2014) describes this as the polysemy problem. Polysemy refers to the coexistence of many possible meanings for a word or phrase. Han and Chan, (2002) illustrate an example of this using the search term "jaguar". In different contexts, the term can refer to an animal, a sports brand, a car, an American football team or a computer. For this reason, Han and Chan, (2002) point out that some search results may be only marginally relevant to the searched topic. They also go further to illustrate that many different terms can be used to describe the same thing and as such, keyword-based searching can miss many highly related pages that do not explicitly contain the keywords nominated by the user.
Mahalakshmi (2014) also touches on these limitations by stating that a news topic of any breadth can contain a large number of documents. Conventional information retrieval systems return long lists of ranked documents and users are forced to sift through the search results to manually identify the documents related to a given topic. Another limitation of keyword based searching is that it requires a preconceived notion of what might be found in the document set. Brehmer et al (2014) capture this limitation by stating that if the search target is not known it may not be possible to formulate an effective query. This approach has significant shortcomings especially in an exploratory and investigative journalism context.

As a very specific example of the limitations of keyword-based searching in a journalism context, consider the publication of the Mahon report in Ireland, also known as the "Tribunal of Inquiry Into Certain Planning Matters and Payments" (2012). The report, published in March 2012, was in response to a public inquiry in

Ireland to inquire into the planning history and ownership of 726 acres of land in north Dublin, and to investigate any payments to politicians or officials in connection with its rezoning. The final report consisted of five volumes totalling 3,270 pages. News media organisations were expected to digest the findings of this report in a very short timeframe so that the main themes could be editorialised. The traditional keyword-based searching approach is limited in this context as the journalist must know what to search for in advance. A journalist may hazard a guess that the term "corruption" would feature in the report, but would have to review all of the pages where this term was found and examine the surrounding text to draw any conclusions. Terms closely related to corruption such as bribery, fraud and profiteering may also have been used in the report so the journalist could miss many highly related pages of the report by searching for the term corruption alone. Without reading all of the pages in the report, which is a highly manual and cumbersome task, it is difficult to establish the major themes and key areas of interest in the report.

Document clustering techniques can help address the problems of understanding large document sets and in finding patterns within documents that can be used to develop useful and actionable knowledge (Yoo and Hu, 2006; Green, 2007). Aggarwal and Zhai, (2012) describe the problem of document clustering of that of applying similarity functions to group a set of documents with common features into related documents or clusters based on their similarity or dissimilarity. Basu and Murphy (2013) state that the objective of the document clustering technique is to automatically generate document clusters in which documents within a cluster are similar and documents in different clusters are dissimilar. Document clustering is classified as an unsupervised learning technique as it aims to discover patterns, hidden structures and relationships in unlabelled data without any preconceived notion about the associations that may be found (Green, 2007). The unsupervised nature of the method may have applications for the journalism profession as per the following suggestions.

- Facilitate automatic topic aggregation by identifying the major topics and key areas of interest in large document sets.
- Assist in the progression of investigative exposés by discovering patterns, hidden structures and relationships.

- Facilitate online readers in generating their own interest verticals by recommending similar articles.
- Aid pattern analysis and knowledge discovery.
- Poll news gathering and wire services for emerging topics.
- Facilitate a 360° view of the information space by providing bidirectional links both backwards and forwards in time between topic articles regardless of when the article is published (Smeaton et al, 1998).
- Promote operational efficiencies and lower costs by reducing or eliminating the manual overhead associated with reading large document sets such as freedom of information requests and government reports.
- Use data visualisation techniques to present news and topics in innovative ways.

Considering the potential applications, it is implied that news articles in particular are fertile ground for document clustering principles. News articles have characteristics which distinguish it from other document collections and these should characteristics should be examined before any document clustering approach is determined.

News articles are time dependent. A core component of every daily newspaper is to give an account of events that have recently occurred or are ongoing at the time of publication. As Smeaton (1998) asserts, news topics have a direct relationship and significance related to the date of publication. Consider the World Cup in South Africa in 2010. The vast majority of the articles relating to this event were published in or before the year 2010. The event has a direct relationship with the date. Emerging stories in June 2015 relating to the FIFA corruption scandal may make reference to the World Cup in South Africa in 2010 but these stories have a lesser topic significance to the World Cup topic albeit they are related.

Most news articles can be described as being semi-structured as whilst they follow a field structure that can help separate the semantic elements of the article, the content itself is largely unstructured. An article is composed from several fields such as the headline, summary and body text. Additional metadata such as the publication date, the writer or writers, the category of the article and the genre of the article are also typically held although the availability of this metadata will vary from publisher to publisher. Whilst it is fair to say that all of the text based fields are unstructured and free format, it should be noted that the headline and summary field of an article aim to assert the content of the news in a very succinct manner and as such terms featured in these specific fields have a distinguishing power and a different degree of importance than terms featured in the body text of the article.

As news stories evolve, new articles are added to the collection and may closely relate to articles already present in the collection. Smeaton et al (1998) cite the high-profile Louise Woodworth (1997) case to illustrate the breadth and scope of a news story: a child is murdered, a criminal investigation is launched, evidence is gathered, a suspect is arrested and charged, the charge is reduced to manslaughter, a trial happens, the jury find the suspect guilty, the judge overrules the jury and sets the accused free, she writes a book, makes a million dollars and lives happily ever after. Each published article represents a snapshot of a particular event at a given point in time. Many news articles do not have a defined beginning or end and topics can evolve overtime.

Now consider how a user may navigate through a news topic information space. If each of these articles were published online, hyperlinks could be manually created linking articles within the topic. One obvious limitation of this approach is that it is manually intensive. Another is that links can only represent the information as available at the time of publish and can only be applied retrospectively. It is not possible to link to a future article before it is published. At opening of the trial users are likely to be interested in earlier stories related to the topic, but the navigation path for these earlier stories will be limited or non-existent. No reciprocal hyperlinks from the earlier stories to the more recent stories are likely to exist unless the entire topic linking is re-indexed manually every time a new article relating to the topic is published. This solution represents a large duplication of effort and is not sustainable with the resource constraints for most media organisations. Smeaton et al (1998) propose document clustering as a potential bi-directional linking solution. Regular document clustering could potentially facilitate a 360° view of the information space by providing bidirectional links both backwards and forwards in time regardless of when the article is published. A cluster relating to this topic should endeavour to include all stories relating to the topic regardless of when the story is published.

Term weighting schemes are central to document clustering methods. Term weighting is a statistical measurement that quantifies the importance of a term within a document and within a document collection. In the literature a variety of term weighting approaches have been advanced by various authors. The dominant and most commonly used term frequency weighting functions are as follows:

- Binary - (Salton and Buckley, 1988)
- Term Frequency - (Luhn, 1957, Salton and Buckley, 1988)
- Logarithmic (Harman, 1992; Kolda, 1997)
- Inverse Document Frequency (Spärck Jones, 1972)
- Term Frequency -Inverse Document Frequency (TF-IDF) (Spärck Jones, 1972, Salton and Buckley, 1988).

The core problem under discussion in this work is how this weight is assigned. This study contributes a review of the dominant and most commonly used term frequency weighting functions in the literature, establishes the merits and limitations of each approach, and proposes modifications that are specific to the domain of news articles. This is discussed further in following Literature Review chapter.

A review of the literature shows there is growing interest in data-driven journalism and the broader use of data for journalism (Holliman, 2011; Bradshaw, 2013; Grey et al , 2013; Ridgway and Smith, 2013; Howard, 2014 and Harding, 2015) . The literature highlights a need for better tools for the journalism profession to understand and develop actionable knowledge from large document sets (Green, 2007; Lorenz, 2013; Gray et al, 2013; Meyer, 2013; Charles, 2014; Harding, 2015). Data-driven journalism is an increasingly strong research theme reflecting the increased role that data plays in the production and dissemination of global news. Gray et al (2013) believe that the growing interest is evidenced by the formation of specialist data-driven journalism teams at the world's leading news organisations including the Guardian, the BBC, the New York Times, the Chicago Tribune, the Texas Tribune and Die Ziet. The success of the early influential examples of data-driven journalism at these leading news organisations is influencing mainstream media to develop stronger data-driven reporting capabilities. The emergence of data-driven journalism is driven partly because data is ever increasing, cheaper to store and more accessible, but it also reflects a state of deepening crisis for the journalism profession (O'Donnell, McKnight and Este, 2012). This crisis arises from a deteriorating readership, a fragmenting advertising market, a global economic downturn and the impact that digital technology and social networking practices have had on the traditional business model (Meyer, 2009; Fuller, McChesney & Nichols, 2010; Grueskin et al, 2011; Cokely, Franklin, 2013; Harding, 2015).

To remain profitable, media organisations need to innovate and create new-value, higher-quality content that seeks to innovate on several key experiences for their readers (Mersey, 2010). Now that information is abundant (Meyer, 2011) and real-time reporting is routine (Gilmore, 2008) journalists need to adapt their strategic practices to survive. As circulation figures continue to decline news organisations need to customise their content for niche audiences (Kaye & Quinn, 2010). One area in which news gathering organisations are adapting their strategic practices is through the provision of data-driven journalism. Data-driven journalism is an umbrella term that describes a cultural shift in journalism in which technology and ever increasing volumes of data are shaping how journalists add value to an overcrowded news market. Data-driven news stories are based upon facts and evidence and are becoming increasingly common as a way of developing unique news content, of stimulating debate and in the progression of investigative exposés. Bradshaw (2013) describes data-driven journalism as "the possibilities that open up when you combine the traditional nose for news and ability to tell a compelling story, with the sheer scale and range of digital information now available". By using data, the role of the journalist shifts its main focus from the being the first to report a story to one of providing a deeper insight into what a certain development may mean (McCandles, 2013).

This additional insight or added-value was the cornerstone of the "Story of Why" brand campaign ran by the Irish Times in 2012. The objective of the brand campaign, was to distinguish the Irish Times as a news organisation with the unique insight, context, explanation and clarity to provide the answer to the important question, why?. A billboard campaign relating to the riots that took place in London in August 2011 as described in Table 1 captures the theme succinctly. Digital technology and the internet have provided the capabilities for readers to find out the where, the when, the what and the who of any given story in real-time. The "Story of Why?" brand campaign aimed to differentiate the Irish Times from all other news sources by promoting the journalistic insight that provides answers to the all important "Why?" .

Question	**Answer**
Where?	London
When?	August 2011
What?	Riots
Who?	David Cameron, the Metropolitan Police, Boris Johnson
Why?	The Irish Times - The Story of Why

Table 1: The Story of Why? brand campaign by The Irish Times 2012

The union of journalism and data aims to provide an additional insight and value-add layer to news stories and represents one way in which journalists are attempting to distinguish themselves and add value in an overcrowded and challenging news market (Rusbridger, 2010; Mersey,2010). The union of journalism and data still has the fundamental objective of providing information and analysis to inform the public about the most significant issues of the day albeit with a new toolset.

1.2 Research Project

This study contributes a review of the main trends and challenges in the journalism professional as presented in the literature. It also provides a review of the dominant and most commonly used term frequency weighting functions put forward in research, establishes the merits and limitations of each approach, and drawing from current research and best practices proposes modifications that are specific to the news domain.

The basic research problem is as follows: Given a large collection of unstructured news documents, can the major topics and key areas of interest discussed in that collection be identified automatically?. Grounded in information retrieval theory and drawing from best practices, current research and a review of current literature, a number of news-centric term weighting and document similarity modifications will be proposed and developed into a document clustering toolset. The objective of the study is to establish if the newly proposed news-centric TF-IDF term weighting and document similarity approach improves document clustering accuracy for news articles when compared to the traditional TF-IDF approach.

1.3 Research Objectives

The following objectives have been identified for this dissertation:

- Review of the main trends and challenges in the journalism professional as presented in the literature.
- Complete a review of the dominant and most commonly used term frequency weighting functions put forward in the research to identify research trends and to establish best practice.
- Drawing from best practices and a review of current literature, propose a number of news-centric term weighting and document similarity modifications and design a suitable toolset.
- Develop the proposed toolset using series of open-source technologies.
- Qualitatively assess and compare the effectiveness of each term weighting approach by assessing the data visualisation diagrams produced through the application of quantitative document clustering and data visualisation methods.
- Qualitatively assess the effectiveness of each term weighting approach by comparing the automated results to a manual document clustering effort undertaken by the domain expert users.
- Qualitatively compare the similarity of a number of documents.
- Evaluate findings informed by the body of research in the area.

1.4 Research Methodologies

Much of this study is based upon existing research technology and information retrieval theory. The novelty lies within how these disparate technologies are used in combination and the focus on using these technologies as a means of developing actionable knowledge for the journalism profession. This study contributes a review of the dominant and most commonly used term frequency weighting functions put forward in research, establishes the merits and limitations of each approach, and proposes modifications that are specific to the domain of news articles. The research approach for this study is to review the current literature and establish best practices from which a number of news-centric TF-IDF and Cosine Similarity modifications will be implemented into a toolset.

This study largely follows a quantitative research approach relying on a numeric and statistical approach to generate and visualise the document clusters. The evaluation of the results however largely follows a qualitative approach. According to most of the document clustering research there is no clear definition of what constitutes as the correct clustering solution for any given document set (Green, 2007). The performance of any clustering algorithm may be judged differently depending on the measure that is used. Steinbach (2000) advocates using several measures to evaluate the results. As such, to evaluate the results, three evaluation approaches will be used.

1. The data visualisation diagrams produced through the application of quantitative document clustering and data visualisation methods will be qualitatively assessed for each term weighting approach. Whilst evaluating these diagrams visually may be considered a subjective approach, the document clusters will be generated using non-subjective and quantitative methods.

2. The quality of each term weighting approach will be assessed by comparing the automated results to a manual document clustering effort undertaken by a group of expert users using the same document sample. These users will have extensive experience and knowledge in the domain of journalism and particularly in the daily operation, categorisation and management of the news as it is published in the Irish Times.

3. The quality of each term weighting approach will be assessed by quantitatively assessing the similarity measure of a subset documents. The comparison of articles is a factorial of the number of input articles and as such a subset will be used in this analysis.

1.5 Scope and Limitations

The scope of this project will focus on the development of a document clustering and topic aggregation approach and toolset for a specific news media organisation. The modifications will be tailored to the domain of news and as such, some of the findings may not be applicable to other domains.

As outlined by De Vries et al. (2012), each document collection is unique and clustering results are not comparable with other representations as the results are specific to that exact document representation. The results outlined in this study will be specific to the content and scale of the document sample used. Different results are likely if a different document sample and timescale or different evaluation measures are used.

Objectively evaluating the quality of document clusters and establishing if the clusters are meaningful is a significant challenge. The results will largely represent a qualitative assessment informed by expert users and as such may be subjective in nature.

1.6 Document Outline

This document is organised into six chapters as follows.

Chapter 1 Introduction

This chapter provided and overview and background of the project area. A summary of trends and challenges in the journalism profession were presented and document clustering was put forward as a potential solution to help address some of these challenges. The attributes that are specific to a document collection of news articles and which distinguish it from other document collections were illustrated as these characteristics will need to be considered carefully when reviewing the literature. The research question was presented along with the research objectives and methodologies to be deployed. Finally, the scope and limitations of the research were outlined.

Chapter 2 Literature Review

The literature review contributes an assessment of the main trends and challenges in the journalism profession and highlights a growing interest in data-driven journalism, and specifically that the journalism profession needs better tools to understand and develop actionable knowledge from large document sets. Document clustering methods are explored as a potential solution to help address some of these challenges.

The literature review also provides a review of the dominant and most commonly used term frequency weighting functions and establishes the merits and limitations of each approach. Drawing from this research, a number of news-centric term weighting and document similarity modifications will be proposed in the toolset design.

Chapter 3 Design

In this chapter, informed by best practice and a review of current literature, the study design and toolset architecture will be discussed. The study design outlines the data sources to be used in the experiment, the feature selection and extraction strategies and the proposed evaluation strategy to assess the results. The toolset design section outlines the required data preparation strategies, the document clustering approach as well as the design, functionality and technical architecture of the toolset being developed.

Chapter 4 Implementation and Results

In this chapter, the implementation of the experimentation is described and the initial findings are presented and analysed.

Chapter 5 Evaluation

The focus of this chapter is on the evaluation of the two term weighting approaches. The objective of this evaluation is to establish if the newly proposed term weighting approach improves document clustering accuracy and topic aggregation capabilities for news articles when compared to traditional term weighting term weighting approach.

Chapter 6 Conclusion

The focus of this chapter is to conclude the study by re-establishing the problem definition and providing an overview of the research objectives met during the study. The contributions to the body of knowledge are outlined and the experimentation, evaluation and limitations of the research are discussed. The chapter outlines the future work for the research before concluding the study.

2 LITERATURE REVIEW

2.1 *Introduction*

The following chapter provides an assessment of the main trends and challenges in the journalism professional and contributes a review of the dominant and most commonly used term frequency weighting functions in the literature. It highlights a growing interest in data-driven journalism, and specifically that the journalism profession needs better tools to understand and develop actionable knowledge from large document sets. Drawing from this research, a number of news-centric term weighting and document similarity modifications will be proposed for the toolset design.

2.2 *Trends & Challenges in the Journalism Profession*

2.2.1 The Journalism Profession under Siege

Gray et al (2013) contend that the journalism profession in under siege. Indeed O'Donnell et al (2012), argue that it can be viewed as being in a state of deepening crisis. This crisis arises from a deteriorating readership, a fragmenting advertising market, a global economic downturn and the impact that digital technology and social networking practices have had on the traditional business model (Meyer, 2009; Fuller, McChesney & Nichols, 2010; Grueskin et al, 2011; Cokely, Franklin, 2013; Harding, 2015). The willingness to pay for news information has declined (Gluck & Roca, 2008) and continues to decline year on year. In the traditional newspaper model information was scarce and journalism efforts were dedicated to gathering and distributing information. As Gilmore (2008) states, before the evolution of the internet, the circulation of world news was almost entirely the domain of newspaper journalists. Gilmore describes this as the "gravy train" era of media during which news was delivered very much as a lecture. Harding (2015) argues that the traditional model of the media as news transmitters and consumers as news receivers has been completely transformed by the internet and the emergence of social networking sites such as Twitter, Facebook and YouTube. Social networking sites are playing an increasingly significant role in the dissemination of worldwide news and information (Kwak, Lee, Park, and Moon, 2010). Research put forward by Newman, (2012) suggests that young internet users are more likely to discover and share news through social networking platforms than through traditional media platforms. Power is moving away from news and media organisations as the gatekeepers of news in the public domain and social networking users are increasingly assuming a more active role as assemblers, editors and even creators of their own news. The emergence of social media has provided the communications gateway by which anyone can become a journalist with little or no cost and in theory with global reach (Gilmore, 2008). Harding (2015) captures this shift with the statement that “anyone with an internet connection and a Twitter account can make the news".

2.2.2 Value-Added Journalism

To remain profitable, Mersey (2010) asserts that media organisations need to innovate and create new-value, higher-quality content that seeks to innovate on several key experiences for their readers. As circulation figures continue to decline Kaye & Quinn (2010) attest that news organisations need to customise their content for niche audiences. Now that information is abundant (Meyer, 2011) and real-time reporting is routine (Gilmore, 2008) journalists need to adapt their strategic practices to survive. Rusbridger (2010), editor-in-chief of the Guardian until May 2015, focuses on the value that data can bring to journalism by advocating that journalists should not merely act as stenographers but need to add value to the story. He goes further to state that the value of a journalist merely distributing information in a crowded internet market is nearing zero.

2.2.3 The Emergence of Data Journalism

One area in which news gathering organisations are adapting their strategic practices is through the provision of data-driven journalism. In an overcrowded news market, data-driven journalism is one of the few areas where the media can add value. Data-driven journalism is an umbrella term that describes a cultural shift in journalism in which technology and ever increasing volumes of data are shaping how journalists tell stories and add value to an overcrowded news market. Bradshaw (2013) describes data-driven journalism as “the new possibilities that open up when you combine the traditional ‘nose for news’ and ability to tell a compelling story, with the sheer scale and range of digital information now available”. By using data, the role of the journalist shifts its main focus from the being the first to report a story to that of providing a deeper insight into what a certain development may mean (McCandles, 2013). For this reason, Holliman (2011) advocates that journalists should see data as an opportunity. Lorenz (2013) also contributes to the value debate by articulating that "gathering, filtering and visualising what is happening beyond what the eye can see has a growing value". Observers such as Howard (2014) argue that data and associated activities will only matter more in the years ahead. Harding (2015) shares this view and advocates that an increasing number of news stories will be found in government data, corporate data and data obtained under the Freedom of Information Act (2003) in the future. Ridgway and Smith (2013) contend that as data-driven news stories are based upon facts and evidence they will become increasingly common as a way of developing unique news content, of stimulating debate and in the progression of investigative exposés. The union of journalism and data still has the fundamental objective of providing information and analysis to inform the public about the most significant issues of the day.

2.2.4 Large Document Sets and the Journalism Profession

Large document sets have become a significant part of the journalism profession. This is part because data is ever increasing, more global, cheaper to store and more accessible, but also reflects the significant shift in how society communicates driven by the rise of social networking practices and the prevalence of ubiquitous devices. On a daily basis, journalists are tasked with searching a diverse range of document sets including documents from open government initiatives, news gathering services, Freedom of Information (Act 2003) requests, WikiLeaks documents, court records, social network data and many other types of large and generally unstructured document sets. Brehmer et al (2014) ascertain that documents of this nature often contain interesting and newsworthy material but eliciting actionable knowledge represents a challenge due to the sheer number and size of documents in the collection. The buzzword "big-data" is often invoked when discussing data-driven journalism. The term aims to describe a volume of digital data that is so overwhelming, traditional editorial and data processing techniques are struggling to adapt. A key challenge for the journalism profession therefore, is addressing the accessibility of large document sets and the extraction of value from them.

The literature shows that there is growing interest in data-driven journalism (Holliman, 2011; Bradshaw, 2013; Grey et al , 2013; Ridgway and Smith, 2013; Howard, 2014 and Harding, 2015) or the broader use of data for journalism and specifically that the journalism profession needs better tools to understand and develop actionable knowledge from large data and document sets (Green, 2007; Lorenz, 2013;
Gray et al, 2013; Meyer, 2013; Charles, 2014; Harding, 2015). Harding (2015) captures this succinctly by asserting that the ability to source, represent and make sense of large volumes of data will be vital for the journalism profession. Green (2007) describes an absence of a comprehensive toolkit for the exploration of unstructured text datasets. Lorenz (2013) describes a need for "sensemakers" who are equipped to dig through big data sets and transform it into something tangible. Howard (2012) states that making sense of big unstructured data sets will be a central goal for the journalism profession in the future. Meyer (2013) asserts that the journalism profession need to "bring sense and structure out of the never ending flow of data" and that the processing of information and data will become a more important concern for the journalism profession. Tim Berners-Lee who as often cited as the inventor of the World Wide Web, advocates that the future of journalism lies in analysing data, "lots of data" (Gray et al, 2013). Harding (2015) describes an increasing scope for the automation of some journalism. O'Murchu (2013) hints at the application of topic aggregation by stating that finding ways to pinpoint key areas of interest will become more important for the journalism profession.

Document clustering is still very nascent within the field of journalism which is part of the motivation for this research. A number of studies have shown the application of document clustering for topic detection (Cai, 2005; Xie and Xing, 2013, Pantel and Lin, 2002; Pantel and Lin, 2002; Liebsher, 2004) and social network data mining (Paltoglou and Thelwall, 2010; Aiello et al, 2013) which have a relevance for the journalism profession. Smeaton et al (1998) clustered an archive of newspaper articles for the purposes of improving document retrieval and browsing.

One notable example of document clustering in a journalistic context is the work completed as part of the "Message Machine project" by ProPublica, a non-profit newsroom that produces investigative journalism in the public interest (Cenaiko, 2012). Over 30,000 political broadcast and fundraising email messages that were sent during the U.S. presidential elections in 2012 were collated. Document clustering and decision tree algorithms were applied to analyse how U.S. political parties tailored their fund raising messages based on the perceived demographics and wealth of the email recipients. The results strongly infer that audience targeting had occurred which was based upon the demographic, age, address, donation history and perceived wealth of the email recipients. Obama for America sent 1703 distinct email message variations to supporters.

The work completed by Brehmer et al (2013) on the Iraqi War Logs represents another excellent example of document clustering in a journalism context. These group of researchers developed a document clustering application specifically for journalists for the automatic detection of keyword patterns in large document collections. Their experiment generated keyword clusters from thousands of U.S. military significant action reports leaked by WikiLeaks during the bloodiest month of the Iraq war in December 2006. Rather than reading thousands of documents or searching the document set using preconceived search terms, a term-weighting and document clustering method was applied to automatically produce clusters of the main keywords used in the document collection. The resulting clusters were dominated by disturbing keywords such as blindfolded, corpse, detonated, shot, handcuffed, mortar, injured and abducted to name but a few. Without reading thousands of documents, the cluster labels instantly paint a picture of the disturbing and serious nature of the events in Iraq at time. Certain keywords may spark the imagination of a journalist looking for a news story. This technique facilitated the quick summarisation of a large, complex and unstructured dataset. This type of information can be used to develop unique news content, in an investigative journalism context and to assist the journalism profession in making hypothesis.

Borglund (2013) used document clustering methods to build an innovative and a self-learning news portal that provides innovative and personalised news to the end-user based on different criteria such as user’s location and feedback.

2.3 Document Clustering

2.3.1 Document Clustering Overview

Document clustering techniques can help address problems of understanding large document sets and in finding patterns within documents that can be used to develop useful and actionable knowledge (Yoo and Hu, 2006; Green, 2007). Aggarwal and Zhai, (2012) describe the problem of document clustering of that of applying similarity functions to group a set of documents with common features into related documents or clusters based on their similarity or dissimilarity. Basu and Murphy (2013) state that the objective of the document clustering technique is to automatically generate document clusters in which documents within a cluster are similar and documents in different clusters are dissimilar. Document clustering is classified as an unsupervised learning technique as it aims to discover patterns, hidden structures and relationships in unlabelled data without any preconceived notion about the associations that may be found (Green, 2007).

2.3.2 Natural Language Processing

Natural language processing is a field of computer science concerned with the application of computational and linguistic techniques to analyse the human language (White, 2004). Since its inception, natural language processing research has been focused on tasks such as information retrieval (Salton and Buckley, 1988), document clustering and topic aggregation (Cai, 2005; Xie and Xing, 2013, Pantel and Lin, 2002; Pantel and Lin, 2002; Liebsher, 2004) and more recently, opinion and sentiment mining (Paltoglou and Thelwall, 2010) and social network data mining (Ifrim, Shi and Brigadir, 2014). Research into natural language processing (NLP) began in earnest in the 1950s. Automation started with Turing's model of algorithmic computation (1950) which proposed what is now more famously known as the "Turing Test" in which a human judge engages in natural language conversation with a machine designed to generate a conversation synonymous with a human conversation. Shannon (1951) applied probabilistic models of Markov processes in an attempt to process language automatically. Chomsky (1957) described by Fox (1998) as the "father of modern linguistics” began advancing research into formal linguistics and Harris’ (1954) compared automated methods for document classification. An early reference to the term "bag of words" approach can be found in this work.

2.3.3 Bag of Words Model

In the bag of words model, each document is represented as matrix of terms with a term weight applied. The vast majority of document clustering approaches are based upon the bag of words and vector space model (Salton, 1989). In the bag of words model, the document is converted into an unordered collection of words such that each document can be represented quantitatively. The bag of word model assumes that the order of the words has no significance and a number of studies are critical of the approach as a result. Several researchers take the position that word order is too important to ignore. Wang, McCallum and Wei, (2007) insist that word order is critical to capturing the meaning of the text and ignoring word order is not appropriate to natural language. Wallach (2006) states that word order is a key component to the topic inference itself. Tirilly et al (2008) describe the bag of words vocabulary building process as coarse and likely to produce noisy words. Consider the term, “The wolf ate the lamb”. In the bag of words approach the term “The lamb ate the wolf” has the same probability and significance. Whilst both sentences have the exact same words, the different word order distinguishes the subject and object and the subsequent meaning of the sentence. Some of the available research therefore would indicate that the bag of words approach has significant limitations for natural language processing.

Other researchers offer a different point of view. Salton (1989) affirms that all of the most successful information retrieval systems ignore the order of words and just use the frequencies of words in documents. Pereira (2000) advocates that Harris (1954) developed what is “arguably the best articulated proposal for a marriage of linguistics and information theory”. Wang et al (2007), although critical of the loss of grammatical context do acknowledge that the bag-of-words model has enjoyed a big success. Schutze (1992) has noted that the bag of words approach works better for nouns than for verbs, but less effective than methods that take other relations into consideration. Pereira (2000) and Gale et al (1992) describe the bag-of-words models as having a stronger global coherence than other language models. Gale et al, (1992) capture a particular form of this coherence using the co-occurrence of the words “stocks”, “bonds” and “bank” in a document. If these words are collocated in the same document it infers that the document is financially themed disambiguating these word occurrences and reducing the likelihood that the “bank” is a river bank, that the “bonds” are chemical bonds, or that the “stocks” are an ancient punishment device. Firth (1957), an early adopter in natural language research famously said “You shall know a word by the company it keeps”. As this study is predominantly concerned with automatic topic aggregation the loss of word order is not as significant as identifying a global coherence. Furthermore, as noted by Schutze (1992), this approach favours noun identification. If the objective of this study is to identify the key areas of interest being discussed, an approach that can identify people places and things is advantageous.

2.3.4 Term Weighting

To compare documents with each other they must be transformed into a mathematical

model so that they can be assessed quantitatively. The combined use of term weighting schemes and the vector space model introduced by (Salton et al., 1975) are central to the study of document clustering methods. Term weights indicate how significant a term is with respect to its context (Salton, Wong, and Yang, 1975). The core problem under consideration in this work is how this weight is assigned and which term weighting scheme is most appropriate for the purpose of clustering a collection of unstructured news articles.

One of the earliest references to term weighting in the literature is presented by Luhn (1957) in which he describes a problem of “searching to find those documents within a collection that have a bearing on a given topic”. Luhn (1957) proposed that the weight of a term that occurs in a document is proportional to frequency of the term in that document. Salton (1970) demonstrated that using weighted term frequencies resulted in superior retrieval performance over un-weighted terms.

2.3.5 The Vector Space Model

The vast majority of document clustering approaches are based upon the bag of words and vector space model (Salton, 1989). In the bag of words model, documents are represented as a weighted document term matrix. To convert the bag of words model into a vector space model, each document vector is normalised by the Euclidean length of the vector. The result is that each document vector has a length of 1 or a unit length. This normalisation is applied so that the cosine similarity measure can be calculated.

2.3.6 Cosine Similarity

Computing the cosine similarity of document vector representations is the standard way of measuring the similarity between documents. The cosine similarity of a pair of documents is the cosine measure of the angle between the two vectors.

$$sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}$$

Equation 1: Cosine Similarity

With each document in the collection represented as term-weighted vectors, cosine similarity can calculated between every pair of documents in the collection. Documents with overlapping terms will score a higher cosine similarity score. Cosine similarity values range from a value of 0 if the document shares no overlapping terms and 1 if the documents are identical.

2.3.7 Document Cluster Representation

Document cluster analysis can be performed in a number of ways. According to According to Jain, Murty and Flynn (1999) the methods can be categorised into two groups; partitioning and hierarchical methods. Partitioning methods divide the document collection into a series of clusters where no overlap is allowed. Each document in the collection has a membership of exactly one cluster with which it is most similar. The advantage of partitioning methods is that they are fast, computationally efficient, robust and easy to understand (Kumar, Ranjan and Dhar, 2012).

Hierarchical document clustering methods are more complex and organise the document collection into a tree or hierarchy structure that produces a nested data structure of cluster and sub-cluster relationships (Jain et al, 1999). Documents can belong to more than one cluster with varying degrees of membership. This type of structure is more informative than the flat clustering and has the advantage of providing a taxonomy or hierarchical index (Kumar et al, 2012). Hierarchical clustering methods are often cited as yielding the best cluster quality (Bai and Manimegalai, 2010; Jadon and Khunteta, 2013, Steinbach et al, 2000).

The advantages of hierarchical clustering come at the cost of lower efficiency. Hierarchical clustering methods suffer from quadratic time complexity requiring a processing time that is quadratic in proportion to the number of dimensions. In contrast, partitioning clustering methods offer linear time complexity requiring a processing time that is directly proportional to the number of input elements. For this reason, partitioning clustering methods are cited in the research as being most suited for clustering large document sets (Boley et al, 1999). Some researchers argue that partitioning clustering methods produce inferior clusters but others have found that variants on classic partitioning clustering methods were as good or better than the hierarchical approaches for a variety of cluster evaluation metrics (Steinbach et al, 2000) .

One major drawback of hierarchical clustering methods is that they do not contain any provision for the reallocation of entities (Steinbach et al, 2000; Bai and Manimegalai, 2010). In other words, the algorithm can never undo what has been done previously and does not facilitate a document collection that is ever-growing. The clustering representation must be started again from first principles when new documents are added to the collection.

2.4 Term Weighting Schemes

A large number of term weighting schemes have been proposed in the literature to define and measure the significance of terms.

The dominant and most commonly used term frequency weighting functions are as follows:

- Binary - (Salton and Buckley, 1988)
- Term Frequency - (Luhn, 1957, Salton and Buckley, 1988)

- Logarithmic (Harman, 1992; Kolda, 1997)
- Inverse Document Frequency (Spärck Jones, 1972)
- Term Frequency -Inverse Document Frequency (TF-IDF) (Spärck Jones, 1972, Salton and Buckley, 1988).

It should be noted that a large number of variants of each of the term weighting schemes cited above have been put forward in the literature but presenting each of these and their subtleties is beyond the scope of this study.

2.4.1 Binary Weighting

Binary term weighting (Salton and Buckley, 1988) which is sometimes referred to as Boolean term weighting assigns a binary weight based on a term's presence or absence in a document. A weight of 1 is assigned if a term is present and 0 if a term is absent. One significant shortcoming of the binary term weight metric according to Kolda (1997) is that it gives every single word in the document equal importance. Bassil (2012) advocates that binary and Boolean term weighting schemes are more appropriate for query weighting than document weighting as binary weights do not distinguish between terms that appear once and terms that appear frequently. Polettini (2004) only describes binary weighting as useful in the context where the number of times a word appears is not being considered.

2.4.2 Term Frequency (TF)

The raw term frequency or "TF" of a term within a document is called term frequency (Salton and Buckley, 1988). TF is a weighting scheme that reflects how important a word is to a document by measuring how frequently a term occurs in a document. The tf of term t in document d is represented by the formula presented in Equation 2.

$$tf_{t,d}$$

Equation 2: Term Frequency

TF promotes the idea that terms that occur frequently in a document are of greater importance than terms occurring less frequently. Put simply by Salton and Buckley (1988), terms or words that have occurred frequently in a text should have a higher weight than other terms. Kolda (1997) contents that the raw TF measure applies too much importance on terms that appear frequently in the document the resulting in a possible discrimination of certain documents by making other documents too important. A word that appears ten times in a document is likely to be more important than if it were to appear once, but it is not likely to be ten times more important. Kolda (1997) does contend however that TF is an efficient approach to weight terms due to its simplicity and efficiency.

2.4.3 Logarithmic Weighting

The logarithm variant sometimes referred to as sub-linear TF scaling is a common modification on the standard TF metric. This modification observers the fact that ten occurrences of a term in a document does not equate to ten times the relevance. In other words, the relevance of a term in a document does not increase proportionally with term frequency. According to Roberston et al (2004), the relationship between a query and a document is typically sub-linear to the number of times a term appears in the document.

2.4.4 Inverse Document Frequency (IDF)

The inverse term frequency function better known as IDF is a statistical interpretation proposed by Spärck Jones (1972). It is widely used in an information retrieval context. The objective of the IDF measure is to diminish or normalise the weight of terms that occur too frequently in the document collection to have any real significance. IDF observes the fact that frequent terms are less informative than rare terms. With the IDF measure, terms are weighted not only by the number of times a term appears in a document but also by the number of times the term appears in the entire document collection. This weighting normalises the effect of terms that occur too often in the document corpus to have any significant meaning. Critics of the IDF approach say that it is not set within any specific theoretical framework although Spärck Jones (1972) herself cited a connection to Zipf's law. There have been a number of attempts to theoretically justify the effectiveness of the IDF measure (Aiwana, 2003). The IDF measure is represented by the mathematical formula in Equation 3.

$$idf_t = \log\frac{N}{df_t}$$

Equation 3: Inverse Document Frequency

The idf of term t is the log of the total number of documents in a collection denoted by N, divided by the number of documents in the collection that contain the given term, denoted by df_t.

2.4.5 Term Frequency Inverse Frequency (TF-IDF)

The work completed by Salton and Buckley (1988), proposed a combination of the TF and IDF proposed by Jones (1972). This is known as the term frequency inverse document frequency (TF-IDF) weighting scheme. TF-IDF is one of the most commonly used term weighting schemes in the field of information retrieval with applications ranging from document classification to search engine ranking (Fu and Chen, 2008). Combining the TF and IDF measures, the importance of a term increases proportionally to the number of times it appears in a document (TF). This is offset by the frequency of the word in the entire document collection (IDF). This technique weights words that are important to a document at the same time as excluding words that are common to all documents. This yields the most significant words for each document and for the document collection as a whole. Karkali, Plachouras and Stefanatos (2012) describe TF-IDF as an effective measure for extracting descriptive terms that describe a document well and in eliminating common words in a document collection.

The classic or natural TF-IDF formula can be represented as follows in Equation 4.

$$tf\text{-}idf_{t,d} = tf_{t,d} \cdot \log\frac{N}{df_t}$$

Equation 4: Term Frequency - Inverse Document Frequency

Despite its popularity, TF-IDF has often been considered a convenient heuristic (Aizawa, 2003). It other words it is an empirical or observational method from a probability point of view. Despite this, Aizawa (2003) argues for the measure stating that the effectiveness of TF-IDF method has been justified through the long history of information retrieval and the lack of superior measures.

2.4.6 Temporal TF-IDF

Some variants on the TF-IDF approach include a temporal or time related factor when weighting terms. Liebsher (2004) uses the example of sorting documents into five categories of entertainment, business, sports, politics and weather. Occurrences of the term “Athens” in the sports category are likely to increase significantly relative to the occurrence of the Olympic games in 2004. Many other articles overtime will include the term “Athens” or include references to sports but they are likely to be less relevant to the topic of the Olympics. Liebsher (2004) advocates that intelligent use of the temporal context is likely to prove useful in the field of text categorisation. Karkali, Plachouras and Stefanatos (2012) use temporal modifications on a variant of the standard TF-IDF to automatically cluster web content that is not only relevant but fresh from a temporal perspective.

2.5 Document Clustering Sample Sizes

The literature exhibits a diverse range of document types and test sample sizes in use in document clustering research. Ifrim, Shi and Brigadir (2014) document clustered 1,084,200 tweets from the U.S. presidential elections in 2012. Whilst the document sample used in this research is very large, the document sizes are small and limited to a maximum of 140 characters. At the other end of the scale, Rajaie and Fakhar clustered 50 documents from the Reuters-21578 corpus in their document clustering research. The types of document samples vary greatly also. Paltoglou and Thelwall (2010) clustered 2000 movie reviews on www.imdb.com to establish the opinion or sentiment of the review. In searching for a news related document clustering example, the research highlights the use a number of publically available datasets in document clustering research. One such data set is the Reuters-21578 corpus which lists news articles that appeared on the Reuters newswire service in 1987. As this data set has a news focus the document sample sizes used in the experimentation has a direct relevance to this study.

- Cai (2005) used 8,067 documents from the Reuters-21578 corpus.

- Xie and Xing (2013) used 7,285 documents from the Reuters-21578 corpus.
- Pantel and Lin (2002) used 2,745 documents from the Reuters-21578 corpus.
- Rajaie and Fakhar (2012) clustered 50 documents from the Reuters-21578 corpus.

At the higher end of the document clustering sample size scale, Reed, Jiao, Potok, Klump, Elmore and Hurson (2006) clustered up to 100,000 documents and the entire works of the novel "War and Peace" by Leo Tolstoy. Interestingly, they observed, that there were only small variations in the document frequencies calculated on document sets of different sizes. In other words, as the document collection increases, the distribution of terms does not increase proportionally.

2.6 Document Clustering Challenges

2.6.1 Natural Language Complexities

Language is complex and has structures that are ill defined (Nothman, 2013). Terms are common across categories and their semantic meaning can diverge in different contexts. Mahalakshmi (2014) describes this as the polysemy problem. Polysemy refers to the coexistence of many possible meanings for a word or phrase. Han and Chan, (2002) illustrate an example of this using the search term "jaguar". In different contexts, the term can refer to an animal, a sports brand name, a car, an American football team or a computer. Language is also littered with synonyms where many words and phrases can mean the same thing.

2.6.2 The Dimensionality Curse

The dimensionality of large document collections is often quite vast; a document collection of any magnitude may contain hundreds of thousands of unique terms each of which represents a dimension in the vector space. The phrase "the curse of dimensionality", coined by Bellman (1957), describes the phenomena that arises when an algorithm does not scale well. A large number of document clustering algorithms require an amount of time or memory that is exponential to the number of features or dimensions in the document set. The performance of the algorithm degrades as the size and dimensionality of the document collection increases because the demands for computational resources grow exponentially. Accumulating research indicates that reducing dimensionality is an essential task for clustering large document sets (Yan et al, 2006).

Consider a single edition of the Irish Times with approximately 200 articles. To document cluster this edition, the vectors for each document pair must be multiplied together to calculate at a cosine similarity score. In mathematical terms, the total combination of articles is a factorial of the number of input articles as illustrated in the following mathematical formula.

$$\frac{n!}{(n-r)!\,(r!)}$$

Equation 5: Combinations

A document collection of any magnitude therefore can easily suffer from the dimensionality curse in terms of the required processing time and system resources to cluster the collection appropriately. Table 2 illustrates how the combination of articles scales significantly in comparison to the number of input articles.

Number of Input Articles (n)	Combination (r)	Combinations of Articles
200	2	19,900
400	2	79,800
1000	2	499,500
5000	2	12,497,500

Table 2: Article combinations per number of input articles

2.6.3 Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of random variables under consideration. This is a key consideration for document clustering to reduce overall complexity, to protect computational resources and to limit the sparseness of the document vectors. Two common strategies for dimensionality reduction are feature selection and feature extraction (Yan et al, 2006). The process of feature selection is that of selecting a subset of the existing features without transforming the existing features. The process of feature selection is that of transforming the existing features into a lower dimensional space.

2.6.4 Document Clustering Evaluation

Techniques are required to evaluate the performance of the document clustering method both from an effectiveness and accuracy perspective. Objectively evaluating the quality of document clusters and establishing if the clusters are meaningful is a significant challenge. De Vries, Geva and Trotman (2012), Rosell (2009) and Rosell, Kann and Litton (2004) all contend that evaluating clustering results is very difficult. This difficulty arises according to Bello and Falcón (2008) as the performance of any document clustering algorithm may be judged differently depending on the measure that is used. Green (2007) shares this sentiment and states there is no clear definition of what constitutes a correct clustering solution for any given dataset. The main reason according to De Vrises et al (2012) is that document clustering is used in many different contexts and as such there are many different ways in which the results can be interpreted. He furthers this to say each that as each document collection is unique the clustering results are not comparable with other as results are specific to that exact document representation. Strehl (2002), offers a more general definition by stating that an accurate clustering should reveal the “natural classes” present in the data. Due to the challenges and difficulties in validating document clustering results, Steinbach et al (2000) advocates using multiple validation evaluation methods simultaneously.

Mahalakshmi (2014) offers a definition of the properties that an efficient document clustering algorithm should satisfy as follows:

1. Relevance - The method needs to group relevant documents together separately from irrelevant documents.
2. Brows able-Summaries - The user, at a glance, needs to be able to establish if the cluster’s content is of interest.
3. Overlap - As a document can have multiple topics, documents should not be confined to only cluster.
4. Snippet-tolerance – The method ought to produce high quality clusters even when it only has access to the snippets returned by the search engines

5. Speed – The speed of the clustering method should be of a magnitude of what a user can complete manually.
6. Incremental – The method should start to process each search result snippet as soon as it is received.

To measure the effectiveness and accuracy of the document clustering results, there are generally two approaches put forward in the literature, external cluster validation and internal cluster validation. One external cluster validation approach involves comparing the document clusters with an externally held, known entity such as the applied category. This approach is sometimes referred to as the ground truth as the results are compared to a ground truth set of category labels. Dudoit and Fridlyand . (2002) describe this external cluster validation process as a measure of agreement between two partitions where the first partition is a priori known clustering structure and the second partition are the results from the document clustering procedure. Another external document clustering approach is through manual analysis of clustering results. One such method involves the comparison of the automatically generated document clusters with categories manually applied by domain expert users. Green (2007) argues that whilst it may be possible in some cases for a domain expert to manually evaluate a clustering solution, this will be unfeasible for larger datasets and may introduce an element of human bias.

As described by (Thalamuthu , Mukhopadhyay, Zheng and Tseng, 2006), to validate clusters internally, results are evaluated quantitatively using attributes and features inherent in the data set without external information. The number and quality of the clusters is judged by the ability to attain document clusters in which documents within a cluster are similar and documents in different clusters are dissimilar.

2.7 Conclusion

A review of the literature contributed a review of the dominant and most commonly used term frequency weighting functions put forward in research, established the merits and limitations of each approach. Drawing from this research, a number of news-centric term weighting and document similarity modifications will be proposed in the toolset design as discussed in the following design chapter.

3 DESIGN

3.1 Introduction

In this chapter, informed by best practice and a review of current literature, the study design and toolset architecture will be discussed. The study design outlines the data sources, document sample, term weighting scheme, document similarity measure and document clustering approach to be used in the experiment. The evaluation strategy will also be presented. The toolset architecture outlines the required data preparation strategies, the technologies selected, the design, functionality and technical architecture of the tool being developed.

3.2 Study Design

3.2.1 Study Objectives

The objective of this study is to develop a novel and news-domain specific document clustering and topic aggregation toolset for a news organisation. To meet this objective, experiments will be conducted on a large unstructured collection of newspaper articles from the Irish Times to establish if a newly proposed news-centric term weighting scheme and document clustering method improves document clustering accuracy and topic aggregation capabilities when compared to traditional term weighting approaches.

3.2.2 Data Sources

Newspaper content produced by the Irish Times will be used and document-clustered for this study. Explicit permission was sought and granted by the Irish Times to use their content for the purposes of this study. The Irish Times is an Irish daily broadsheet newspaper, founded in 1859 and considered by many as a pillar of Irish society. Always independent in terms of ownership and political views, the Irish Times has delivered news, opinion and analysis in Ireland for over 150 years. Internationally famous figures such as Daniel O`Connell, W.B. Yeats, Garrett Fitzgerald, Conor Cruise O`Brien, Tony Blair and Bill Clinton have all contributed to the paper over the years. Observers such as Brown (2015) argue that such is the influence of the Irish

Times in Irish society that the history of the Irish Times is also a history of the Irish people.

3.2.3 Document Sample Size

The Irish Times publishes approximately two hundred articles per day, six days per week. A sample of seven weeks worth of newspaper articles published in the Irish Times will be extracted generating a document sample of approximately 4,500 individual documents to cluster. This sample size represents an approximate average of the document clustering sample sizes discussed in the literature review where the using the Reuters-21578 corpus was used. (Cai 2005, Xie and Xing 2013; Pantel and Lin 2002; Rajaie and Fakhar 2012). The Reuters-21578 document corpus is news related and as such the sample sizes used in these experiments has a direct relevance to this study.

3.2.4 Term Weighting Scheme Selection

The literature contributed a review of the dominant and most commonly used term frequency weighting functions put forward in research. The core problem under consideration in this work is which weighting scheme is most appropriate for the purpose of clustering a collection of news articles. The following section establishes the merits and limitations of each term weighting approach from the perspective of meeting this objective. Informed by this research, a number of news-centric term weighting and document similarity modifications will be proposed.

Almost all of the term weighting approaches put forward in literature consider the following metrics to ascertain the importance of a term for a document.

1. The term frequency within each document. The weight of the term increases proportionally to the number of times the term appears in the document.
2. The length of the document. Normalisation is usually applied to prevent bias towards larger documents.
3. The distribution of the term in the document collection. Normalisation is usually applied to dampen the effect of terms that occur too often in the document collection to have any significant meaning.

In the literature, it is pointed out that a simple term frequency measure places too much emphasis on high frequency terms. A number of term weighting variants have been put forward that attempt to normalise the effects of longer documents but the best scheme is still a matter of debate in the literature. Longer documents, as a result of containing more terms tend to have higher TF values. This can increase the importance of terms in longer documents disproportionately resulting in a bias towards longer documents. Newspaper articles whilst undoubtedly variable in length are confined by print space and the range from the minimum article length to the maximum article length is quite small in the context of most document collections. Newspaper articles are typically written to a pre-defined word count reflecting a pre-defined print template or page size. As a result, newspaper articles have a fairly limited range of document lengths so document length bias is not a concern in this domain. The comedian Jerry Seinfeld captures succinctly this by joking once that “It's amazing that the amount of news that happens in the world everyday always just exactly fits the newspaper”. The limits imposed by print and paper sizes negate the need for specialist document normalisation treatment in this study.

As document length normalisation has been ruled out the remaining factors for consideration are:

1. The term frequency within each document.
2. The distribution of the term in the document collection.

On this basis, the most obvious term weighting measure as a baseline for this project is TF-IDF. A number of modifications to the traditional TF-IDF approach will be proposed from which a news-centric term weighting scheme and document similarity approach will be developed into a toolset.

3.2.5 Optimise Key News Fields

An article is composed from several fields such as the headline, summary and body text. The headline and summary field of an article aim to assert the content of the news item in a very succinct manner and as such terms featured in these specific fields have a distinguishing power. Dor (2003) describes the headline field as a relevance optimiser that acts as textual negotiator between the article and its readers. As such terms appearing in the headline and summary fields of articles should be allocated special weights to convey the fact that these terms have a special communicative function. Optimising this communicative function is a key consideration in the news-centric TF-IDF measure. To meet the objective of optimising the terms that feature in the headline and summary fields, each article text field will be treated as a separate document. That is to say, for the proposed news-centric TF-IDF approach, iterations of the traditional TF-IDF measure will be applied to the headline, summary and body text fields as if they were separate and completely unrelated documents. Each article field will arrive at an independent TF-IDF score. The terms in the headline and summary documents will be compared with the larger body text document, and where the terms overlap, the TF-IDF value in the body text document will be increased by the TF-IDF value allocated to the term in the headline or summary document. This approach observers the fact that terms that appear in the headline or summary fields have a special meaning to the article and should have a higher weight than other terms. The traditional TF-IDF approach cannot distinguish these terms which is the basis for the modification. The effect of this modification will be to increase the TF-IDF value of terms in the body text of the document where the term also features in the headline or summary of the article. As the count of terms in the headline and summary fields are significantly lower than that of the body text, the additional term weights that are applied relatively low. They are however sufficient to increase the TF-IDF weight of the most important terms in the document. The most important terms in the document are determined by the inclusion of the term in the headline or summary field.

3.2.6 Cosine Similarity Measure

Articles that share important overlapping terms with a close proximity in terms of publication date are more likely to be related than articles with overlapping terms far apart from each other on the publication timeline. An example of the Olympics in Athens in 2004 was put forward by Liebsher (2004) to illustrate this. Articles overtime may make references to “Athens” or include references to sports but they less likely to be less relevant to the topic of the Olympics if they are not published within a specific timeframe. The work completed by Liebsher (2004) and Karkali, Plachouras and Stefanatos (2012) made modifications to TF-IDF based upon the temporal closeness of documents in the collection. The modifications made are specific to those domains and not directly transferrable to this domain. The concept however of increasing the importance of the documents based on their temporal closeness has interesting connotations for the domain of news articles. To explore this further and to advance research in this area, a modification will be made to the cosine similarity score that increases the similarity of articles as a factor of the initial cosine similarity score multiplied by a discount factor of the time difference between the two articles expressed in weeks. The following formula represents the proposed additional cosine similarity weight to apply.

$$w(d_1, d_2) = \frac{1}{d_1 - d_2} \cdot \left(\frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|} \right) \cdot 0.2$$

Equation 6: Additional cosine similarity weight

The discount function lessens the proposed increase in cosine similarity weight based on the number of weeks between articles. Articles that share important overlapping terms with a close proximity in terms of publication date will yield a higher cosine similarity score. This will have the effect of pulling similar articles published within a close timeframe together when document clustering algorithm is applied.

Cosine similarity values range from a value of 0 if the document shares no overlapping terms and 1 if the documents are identical. A maximum increase of 20% of this weight will be applied for articles featuring in the same week. The additional cosine similarity weight will be a factor of the initial cosine similarity weight such that any increases are proportional to the original similarity of the documents. Articles should not have their original and natural cosine similarity skewed through the application of this weight. Similarity, this additional temporal weight will be applied to all articles as not to introduce bias. If the articles are not very similar, or not close in terms of the publishing date, or both, the increase to the cosine similarity weight will be negligible. The proposed cosine similarity weight increase of 20% of the original cosine similarity value discounted by the difference in weeks is a proposed heuristic and is not set within any specific theoretical framework. Rather, a common-sense approach has been used to boost the similarity of already similar documents when they are published in close succession. Delivering this increase in weight as a configurable parameter will be a key consideration for the design of the toolset. This is sought so that it can be modified with relative ease for future implementations or for other domains.

3.2.7 Clustering Representation

The most influential factor in selecting between a partitioning or hierarchical clustering representation for this study is that hierarchical clustering methods do not contain any provision for the reallocation of entities. As a news archive is ever growing and documents are regularly added to the collection, this approach would not be suitable. To start the clustering process again from first principles every time a new article is added to the collection would represent a large duplication of effort and would be difficult to manage. As such, a partitioning document clustering representation will be used.

3.2.8 Clustering Method

One such partitioning method is the Louvain modularity maximisation method which is an algorithm for detecting and extracting cluster communities from large networks. The Lovain method was proposed by Blondel, Guillaume and Lambiotte (2008). The algorithm uses the cosine similarity edge weight to identify cluster communities. Some researchers argue that modularity maximisation methods fail to detect smaller clusters in some implementations (Fortunato and Barthélemy, 2007). The Louvian method has been used successfully for many large-scale partitioning problems (Haynes and Perisic, 2009).

3.2.9 Dimensionality Reduction

The literature review clearly outlines the need to reduce the dimensional space under consideration. Dimensionality reduction is the process of reducing the number of random variables under consideration. This is a key consideration for document clustering to reduce overall complexity, to protect computational resources and to limit the sparseness of the document vectors. Two common strategies for dimensionality reduction are feature selection and feature extraction.

The process of feature selection is that of selecting a subset of the existing features without transforming the existing features. The following section outlines the feature selection strategies selected for the toolset.

Content produced by the Irish Times can be very diverse in subject matter but broadly speaking every article has one of six parent categories; news, sport, business, opinion, lifestyle and culture. To produce quality clusters the content will need to have some homogeneity or uniformity in substance from the offset. One concern is that the range of content produced by the Irish Times over a long period of time is too broad to provide topic-focused, distinctive and exclusive clusters. Lifestyle content for example is not typically news or topic driven; rather it delivers leisure related content such as food, recipes, fashion, health, beauty and celebrity news etc. As the focus of this study is on topic aggregation the underlying content will need to somewhat homogeneous from the beginning. This will be achieved by only selecting specific genres of content in the document sample. All Irish Times articles are categorised in internal systems with a genre. In journalism terms, a genre aims to reflect the editorial theme of the article, for example “opinion” or “analysis”. The Irish Times adopt the International Press Telecommunications Council (IPTC) global standards for genre classification. Genres specific to current affairs, opinion, analysis, eyewitness amongst a host of other news related themes will be included in the document collection. Conversely, themes specific to life and style content such as recipes, music reviews, wine reviews, book reviews will be excluded from the document sample as their inclusion is likely to clutter the results and obscure emerging topics in the collection. It should be noted, of the remaining document sample, the content will still be diverse in nature and differ significantly in both content and theme.

A number of filters will be implemented in the toolset to exclude any article where the word count is less than 200 words. These articles, sometimes referred to as “shorts”, “blurbs” or “page furniture” in the newspaper industry, are short snippets of text used to fill space or as a teaser to an article within the edition.

The process of feature selection is that of transforming the existing features into a lower dimensional space. The following section outlines the feature extraction strategies selected for the toolset. Stop words and short function words, such as "the”, "is", "at", "which", "and", "on" are filtered out of the documents before they are processed. These types of words are extremely common and non-subject related and as such offer little value when clustering a large document set. There is no single, agreed upon list of stop words for the English language. MySQL, the world's most popular open source database, supply a list of common stop words that are excluded when executing full-text queries against a MySQL database. In the absence of a standard, the MySQL list of stop words will be used to filter words out before the documents are clustered.

As the focus of this study is on topic aggregation not all word types are appropriate for feature extraction. A topic is the subject matter of something that occurs in a certain place at a certain time. Consider "the five W's" information-gathering technique which is often used in a journalism context. Who did it?; what happened?; where did it take happen?; when did it happen?; why did it happen?. As an observation, the five W's pose the exact same questions featured in the Irish Time's "The Story of Why?" brand campaign described in Table 1. Every word in the English language belongs to one or more of eight word classifications as described in Table 3.

Word Classification	**Definition**	**Example**
Nouns	Nouns are used to name people, places, things, or ideas.	Barack Obama, Washington, The White House, Rover, Dog, Cat.
Pronouns	Pronouns take the place of nouns.	The pronouns in the following examples are underlined. Not only is Barack Obama is the President of the United States of America but he is also a bestselling author.
Verbs	Verbs convey actions or states of being.	was, be, am, is, are, run, write, go, wash.
Adjectives	Adjectives describe, or modify, nouns and pronouns. They come right before the nouns that they are modifying.	The adjectives in the following examples are underlined. blue ocean furry cat wet dog wettest dog Claire's book
Adverbs	Adverbs describe verbs, adjectives, or other adverbs.	We will eat here. Your ocean is extremely blue. The dog ran very slowly.

Prepositions	Prepositions show the relationship between a noun or a pronoun and some other word in the rest of the sentence.	The dog ran through the gate. The cat jumped beside the bird.
Conjunctions	Conjunctions are words that join two or more words, phrases, or clauses.	for, and, nor, but, or, yet, so
Interjections	Interjections are words that show emotion. They are not grammatically related to the rest of the sentence.	Hello Goodbye Thanks Phew Cheers Congratulations

Table 3: Word classifications in the English language

To extract a topic, people, places and things are key to establishing the who, the where and the when. As nouns are used to identify people, places and things including nouns is a key feature extraction consideration. To identify what happened, verbs, convey the actions or the states of being. All other word types add colour and flavour to the language but are not key for topic identification. A series of CSV files containing known pronouns, adjectives, adverbs, prepositions, conjunctions and interjections have been compiled. It should be noted there is no single, agreed upon list of each word type. A number of language resources such as WordNet, a lexical database produced by researchers at Princeton University and a number of linguistic websites have been consulted to compile these lists. A distinct file has been provided for each word type such that the scope of the linguistic filters can be reduced, extended or eliminated with relative ease. It should be noted that a crossover exists between words in the stop words list and language filter files. Whilst an inefficiency exists in potentially attempting to remove the same word twice, this approach supports future researchers in customising the use of the toolset for a different problem-specific task. The result of implementing the feature extraction and selection strategies is that all stopwords, non-noun or non-verb terms will be filtered out of the documents before they are processed by the clustering method.

3.2.10 Document Clustering Validation

As presented in the literary review, objectively evaluating the quality of document clusters and establishing if the clusters are meaningful is a significant challenge. As outlined by De Vries et al. (2012), each document collection is unique and clustering results are not comparable with other representations as the results are specific to that exact document representation. One popular external cluster validation approach involves comparing the document clusters with an externally held, known entity such as the applied category. In the context of this study however, there is no external entity standard with which to consult. Articles have a basic navigation-focused category applied, such as sport and football. The category that is applied is distinct from the sub-set of articles that deal specifically with a topic. Some topic data is manually applied to articles in the Irish Times but its use is limited and inconsistent. Furthermore, topics can emerge in the small initially. A defined label or topic may not exist initially when first set of articles related to the topic are published. It is not possible therefore, to consult an external topic reference to evaluate the results.

Another external validation approach put forward in the research involves the comparison of the automatically generated document clusters with categories that are manually applied by domain expert users. As an external evaluation of the accuracy and effectiveness of the proposed news-centric TF-IDF approach and toolset, three senior staff journalists from the Irish Times have been invited to participate in the study.

David Labanyi - Breaking News Editor

David Labanyi is the Breaking News Editor with The Irish Times and a key member of the Irish Times Newsdesk. Prior to joining the Newsdesk he spent three years writing for the Business desk.

Patrick Logue - Lead Homepage Editor, Assistant News Editor

Patrick Logue is the Lead Homepage Editor and an Assistant News Editor with The Irish Times. He has worked for The Irish Times since 1996 having previously held the positions of Breaking News Editor and Online News Editor.

Luke Cassidy – Digital Production Editor

Luke Cassidy is the Digital Production Editor with The Irish Times. He has worked for The Irish Times since 2000. The Digital Production Editor has responsibility for the production and presentation of quality, innovative and engaging digital content online.

In the literature review, Mahalakshmi (2014) suggested the properties an efficient document clustering algorithm should satisfy. The property of “snippet-tolerance” is specific to search and information retrieval and as such has been ignored for this study. The overlap property in which Mahalakshmi argues that documents should not be confined to only cluster appears is at odds with the partitioning clustering method selected for this project. This property oddly enough, also appears at odds with Mahalakshmi’s own statement that the method should be incremental and process each search result snippet as soon as it is received. One of the limitations of hierarchical clustering methods that allow documents to exist to multiple clusters is that they do not contain any provision for the reallocation of entities. It is only assumed that Mahalakshmi’s (2014) intended the clustering to start again from first principles as new information arrived into the collection. Taking these factors into account, a subset of Mahalakshmi’s (2014) properties for efficient document clustering will be used as the evaluation criteria in this study.

1. Relevance - The method needs to group relevant documents together separately from irrelevant documents.
2. Browse able-Summaries - The user, at a glance, needs to be able to establish if the cluster’s content is of interest.
3. Speed – The speed of the clustering method should be of a magnitude of what a user can achieve manually.

The expert users will be tasked with manually categorising a sample of documents and comparing this categorisation effort to the two automated TF-IDF approaches using Mahalakshmi’s (2014) proposed evaluation criteria.

To compliment the expert user analysis and as a measure of how coherently the documents are grouped together, data visualisation diagrams of the clustered document collections will be assessed for each TF-IDF approach. Visualising the clusters provides an overview of the entire document collection and facilitates exploratory data analysis (Feldman and Sanger 2007). It enables intuitive browsing, and navigation of a document collection (Millar and Peterson, 2007). Whilst viewing these diagrams may be considered an empirical approach with a subjective risk, the data visualisations will be generated with the non-subjective and quantitative methods.

Finally, the quality of each term weighting approach will be assessed by quantitatively assessing the similarity measure of a subset documents. The comparison of articles is a factorial of the number of input articles and as such a subset will be used in this analysis.

3.3 Selection of Technologies

The following section details the technologies selected for the toolset.

3.3.1 PHP

PHP is a popular server-side scripting language created in 1995 and designed for web development but also used as a general-purpose programming language. It is the default scripting language in use at the Irish Times and as such was selected for use in this project.

3.3.2 SQLite

SQLite is an open-source embedded relational database management system that implements a server-less and self-contained transactional SQL database engine. SQLite is contained in a C programming library and reads and writes directly to ordinary disk files without a database server. The C programming language is only a few levels of abstraction away from machine language resulting in less translation. Newer programming languages provide support for garbage collection, dynamic typing and other facilities which provide a layer of abstraction that generally make it easier for the programmer to write the program but require more translation and abstraction away from assembler. As a result the C programming language is generally considered faster and more efficient than newer programming languages. Performance is a key consideration when dealing with large document sets. As SQLite is contained in a C programming library it delivers significant efficiencies with it.

3.3.3 Windows Batch Scripts

Using native Windows command line prompts, a series of Windows batch files will be developed to act as the control and entry point for the toolset. The batch files will control the sequence in which the PHP commands are executed. Windows batch scripts were selected to provide a user control point without requiring a graphical user interface.

3.3.4 Ruby

Ruby is a general-purpose object-oriented programming language that was designed and developed in the mid-1990s by Yukihiro Matsumoto in Japan (Flanagan and Yukihiro, 2008). Ruby was initially considered as the implementation programming language for the project. The very first toolset module to parse the Irish Times' newspaper extracts from XML into CSV form was developed in Ruby. The Ruby syntax however, is semantically very different to other programming languages. Mastering Ruby proved too difficult a challenge to proceed but re-engineering the very first module represented a duplication of effort. This module was only used to prepare the document collection in advance of the document clustering method and as such was not integral to the approach.

3.3.5 Gephi

Gephi, an open source data visualisation and exploration platform provides support for the Louvian method and as such will be used to apply the Louvian method and data visualise the document clusters through the application of cluster layout algorithms.

3.3.6 7Zip

An open source file archive application called 7Zip was installed locally for the purpose of unzipping the newspaper extracts. The 7Zip application provides this functionality directly from the command line which facilitates the development of a Windows batch script to complete this task.

3.4 Data Preparation Strategies

To compare and cluster documents they must be firstly transformed into a mathematical model so that they can be assessed quantitatively. Using words as features within each newspaper article, a common set of procedures will be applied to represent each document quantitatively. The following section outlines the data preparation procedures required to transform the articles into the required mathematical model before the document clustering methods can be applied.

3.4.1 Newspaper Files Extractor

For the purposes of processing the ePaper digital edition of the Irish Times, a newspaper extract is automatically generated on a daily basis that represents newspaper articles and any associated images in a compressed XML format. As newspaper images are high resolution images, the extract is quite large and is compressed in a ZIP format. The first data processing step therefore, is to unzip and extract each newspaper ZIP file. Using the 7Zip archive application, a Windows batch file has been created that unzips the XML file from every ZIP file in a given directory and creates a subdirectory for each newspaper edition. With each newspaper represented as an XML file within an edition subdirectory, the next step in the process is to extract the article data fields into a plain text format. A ruby script called parsexml2text.rb has been developed that loops through each subdirectory and parses any XML files held within. The structured XML fields of article id, publication date, page number, headline, summary and body text are extracted into a master plain text CSV file such that each article is represented as a semi-structured row of pipe separated fields.

3.4.2 Document Collection Importer

The module creates a SQLite3 database, creates an articles table and imports the generated CSV file of articles into the table using a pre-defined SQLite3 import function.

3.4.3 Document Collection Data Cleaning

When processing text for document clustering, a key consideration is to remove out non-ASCII, punctuation, formatting, spacing, and other control characters. This is required to reveal text that can be indexed. Consider single and double quote characters, exclamation marks, commas, brackets, asterisks, extra spaces, forward slashes and ellipsis as a small number of punctuation characters that are likely to be present in the text. If left in place, these characters are likely to skew the term frequency and document clustering results. A third-party PHP, open-source licence library written by Nadeau (2008) at NadeauSoftware.com was utilised for this purpose.

3.4.4 Document Feature Selection & Extraction

The feature selection strategies of stop word and non-verb and non-noun word removal are key dimensionality reduction strategies and are implemented as part of this parsing module. The word lists are held in CSV format to facilitate extending or reducing the scope of these filters for any future use of the toolset. The files are read in to a SQLite3 database table and compared to the document collection. Any overlapping words are removed from the document collection and the reduced article table on the database is updated.

3.4.5 Document Index & Term Frequencies

The generated document collection once stripped of all punctuation characters, stopwords, and non-noun or non-verb words is converted into a "bag of words model" such that each article in the collection is represented as a matrix of terms and term frequencies. For each document the key metrics to be counted are the term frequencies and the document frequencies which counts the number of documents each term features in. In the event that a term does not appear in a specific document, but does appear in the document collection, the term frequency for that specific document is not stored as it is implicitly zero. Reducing the number of dimensions in storage is key to reduce the memory and time required to process the document collection. Once the term frequency and document frequency measures have been established for the entire document collection a series of SQLite3 term frequency and document frequency database tables are created and populated with the results.

3.4.6 TF-IDF Weights Calculator

Reading in the SQLite3 term frequency and document frequency database tables, the TF-IDF value is computed for each term and for each document. The calculated TF-IDF values are added to a SQLite TFIDF table.

3.4.7 TF-IDF Vector Space Model Generator

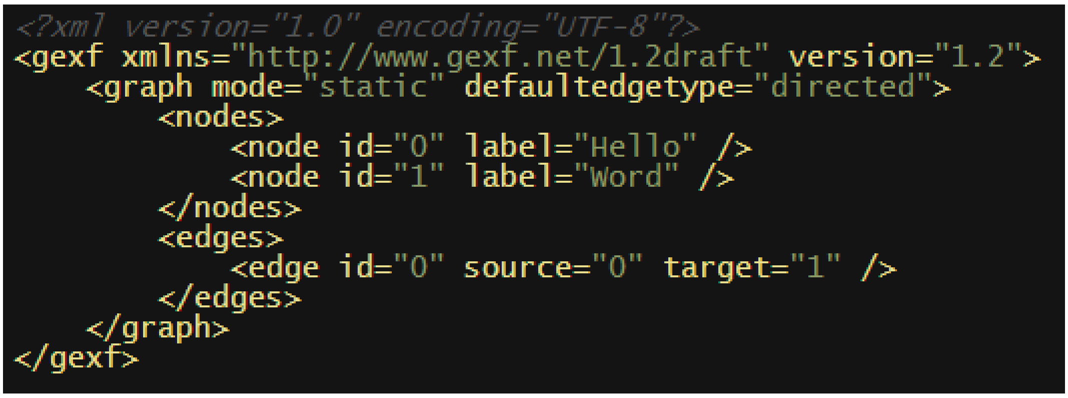
To convert the TF-IDF values into a vector space model, each document vector is normalised by the Euclidean length of the vector. This normalisation is achieved by dividing each TF-IDF value by the total length of the document vector. The result is that each document vector has a length of 1 or a unit length. This normalisation is applied so that the cosine similarity measure can be calculated.

3.4.8 Cosine Similarity Calculator

To document cluster the entire document collection, the TF-IDF vectors for each article must be multiplied together to arrive at a cosine similarity score for each pair of documents. The first step therefore in the cosine similarity calculator module is to calculate all the possible combination of articles in the collection. Once this has been completed, for each pair of articles, the cosine similarity score can be obtained by calculating the dot product between two vectors. A cosine similarity value of 1 indicates that the documents are identical whilst a value of 0 indicates there are no overlapping terms in the documents being compared. This conveniently means the documents with a higher cosines similarity score are the most similar.

3.4.9 GEXF Graph File Generator

In the context of this project, Gephi provides support for implementing the Louvain method and a diverse number of data visualisation and layout algorithms. One of the supported input files for Gephi is a GEXF file. GEXF (Graph Exchange XML Format) is a mark-up language for describing complex intra-related network structures. As such, a module has been built that queries the SQLite3 database for the cosine similarity score for each pair of documents and converts this into a Gephi compatible GEXF file. An example of a basic GEXF file is provided in Figure 1.



```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
    <graph mode="static" defaultedgetype="directed">
        <nodes>
            <node id="0" label="Hello" />
            <node id="1" label="Word" />
        </nodes>
        <edges>
            <edge id="0" source="0" target="1" />
        </edges>
    </graph>
</gexf>
```

Figure 1: Example of a basic GEXF file

PHP provides an in-built DOM parser to generate XML documents. The command line getGexfXML.php utility in the toolset uses the DomDocument library and the cosine similarity score for each document pair to generate a GEXF file. The utility expects a parameter to be provided on the command line that sets the minimum cosine similarity value for the GEXF graph. As presented in the literary review, dimensionality reduction is one of the key considerations for document clustering algorithms. There is no value in including links between articles that have no relationship with each other or that are extremely weakly related to each other. Including these articles on the graph is likely to clutter the graph to such an extent that the performance and accuracy of the document clustering and data visualisation algorithms are affected. The minimum cosine similarity value has been included as a parameter to allow different values to be provided in different contexts and for different document collections or domains. In the context of this experiment a minimum cosine similarity value of 0.2 was provided noting that a cosine similarity value of 0 relates to two articles with no overlapping terms and a value of 1 relates to two identical articles.

3.4.10 Toolset Software Component Diagram

Figure 2 presents a Software Component Diagram for the toolset

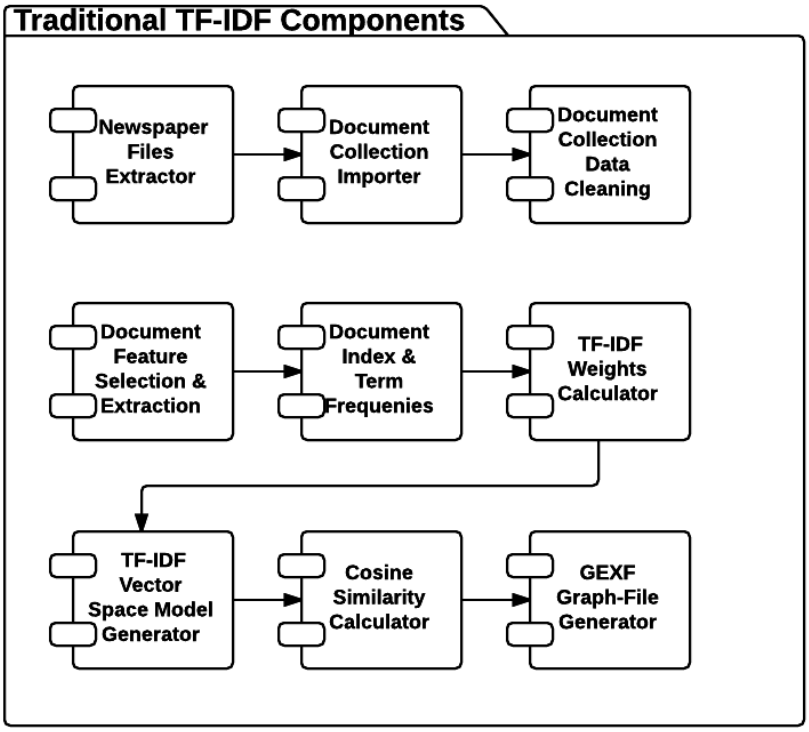

Traditional TF-IDF Components
Newspaper Files Extractor
Document Collection Importer
Document Collection Data Cleaning
Document Feature Selection & Extraction
Document Index & Term Frequenies
TF-IDF Weights Calculator
TF-IDF Vector Space Model Generator
Cosine Similarity Calculator
GEXF Graph-File Generator

Figure 2: Software Component Diagram for the Toolset

3.5 Implementation, Usage and User Tools

The proposed implementation and usage of the framework is to provide journalists with the ability to analyse and compare large unstructured document collections and to automatically and easily identify the major topics and key areas of interest discussed in that collection. Practical uses of the framework from a journalism context may include using the tool to reveal trends, to progress investigative exposés, to develop unique news content, to verify information independently and assist the journalism profession in making hypothesis. The architecture of the framework is modular and extensible and based upon loosely coupled language and data processing libraries. This approach aims to facilitate future researchers in customising the use of the framework allowing for extensions of the framework in the future if required. The toolset also provides a GEXF output function to provide support for future user tools. Including a GEXF export function extends the usefulness of the toolset by providing access to additional data visualisation, analysis and exploration tools.

3.6 Conclusion

Informed by best practice and a review of current literature, this chapter presented the study and toolset design. The scope of this included outlining the data sources to be used in the experiment, the feature selection and extraction strategies and the proposed evaluation strategies to assess the results. The toolset design section outlined the required data preparation strategies, the design, functionality and technical architecture of the tool being developed.

4 IMPLEMENTATION AND RESULTS

4.1 Introduction

Experimentation was conducted on a large unstructured collection of newspaper articles from the Irish Times to establish if the newly proposed news-centric TF-IDF term weighting scheme improves document clustering accuracy and topic aggregation capabilities when compared to the traditional TF-IDF approach. This chapter describes the implementation of this experimentation and presents and analyses the initial findings.

4.2 Document Collection Details

In this experiment, newspaper articles published by the Irish Times during the period of the 11th of April 2015 until the 31st of May 2015 inclusive were extracted. Table 4 describes the data profile of the document collection used in this experiment.

Description	**Count**
Number of documents (before feature selection and extraction)	4,500
Number of documents (after feature selection and extraction)	3,855
Number of document terms	1,884,791
Number of document terms after feature extraction and selection	740,216
Number of unique terms	50,594
Number of document comparisons	7,428,585
Number of document with cosine >0.2	3,708

Table 4: Data profile of the experimental document collection

4.3 Model Building

One facet of the implementation strategy was that of model building. A model of the solution was built initially, tested, trained and refined using one newspaper only. The Irish Times' newspaper from the 17th of April 2015 was used for this purpose. The size of the sample in the initial model was deliberately small as to provide the best possible effectiveness and accuracy of the method before larger document samples were applied.

4.4 Functional Testing

The results generated by the initial model were tested and manually verified against the newspaper content for that publication date. Iteratively, a number of refinements were made to the toolset and the model was run and tested again. The refinements largely related to optimising the feature extraction and selection strategies, expanding the scope of the document cleaning efforts, parsing out punctuation characters from the text and dealing with non-ASCII character corruption. Once the initial model was clustered to a satisfactory level, the document sample was increased.

4.5 Performance Testing

When the large document collection was used, the performance of the toolset was very slow. Whilst the application did not crash at any time, the RAM and CPU resources on the Windows desktop running the toolset were maximised at all times. To improve the performance and speed of the toolset the Windows batch files were converted to Unix shell scripts and the toolset was moved to run on a Unix server instead of a Windows desktop. As significantly more computational resources were available on the Unix server, the performance and speed of the toolset improved drastically. It should be noted however that significant scope still exists to optimise the code further.

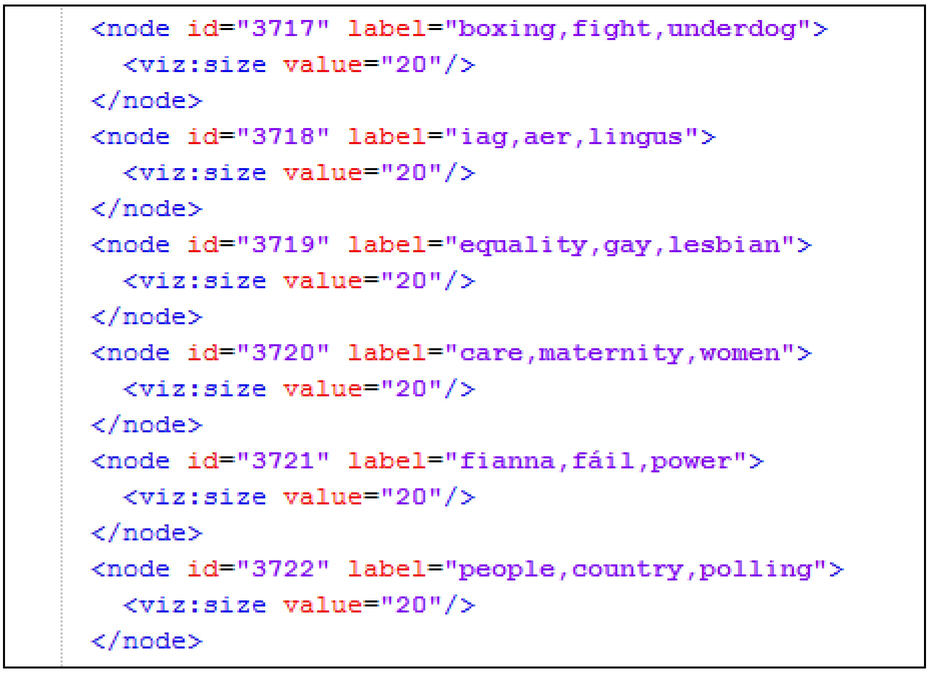
4.6 Implementation

This document collection was imported into the toolset and both the traditional TF-IDF approach and the proposed news-centric TF-IDF approach were applied. For each TF-IDF approach, the toolset generated a cosine similarity matrix for each pair of documents in the collection.

4.6.1 GEXF Graph

The cosine similarity matrix was converted into a GEXF file and imported into Gephi for each TF-IDF approach. The purpose of importing the GEXF file into Gephi is dual purpose, it provides the functionality to cluster the documents through support for the

Louvain method and provides data visualisation capabilities so that the document clusters can be visualised. The label of each node is set to the top three most important terms in the document, as determined by the relevant TF-IDF measurement applied. Figure 3 shows an example of nodes randomly selected from the document collection and the top three terms applied as the node labels.



```
<node id="3717" label="boxing,fight,underdog">
  <viz:size value="20"/>
</node>
<node id="3718" label="iag,aer,lingus">
  <viz:size value="20"/>
</node>
<node id="3719" label="equality,gay,lesbian">
  <viz:size value="20"/>
</node>
<node id="3720" label="care,maternity,women">
  <viz:size value="20"/>
</node>
<node id="3721" label="fianna,fáil,power">
  <viz:size value="20"/>
</node>
<node id="3722" label="people,country,polling">
  <viz:size value="20"/>
</node>
```

Figure 3: Sample GEXF nodes with top terms as labels

It should be noted that the top three terms applied to each document may differ between the traditional TF-IDF approach and the news-centric TF-IDF approach as different weights are applied with each approach. The relationship between documents in the GEXF file is represented as a weighted edge between a source document node and a target document node. For each document in the collection, with a cosine similarity score above or equal to the minimum value provided, an edge is created in the file between the two documents to link them. The cosine similarity value is used as the weight of the edge. The weight defines the degree of friendship between each pair of documents. The edge weight is used by the Louvain method to find partitions or clusters of related documents. It is also used by layout algorithms to data visualise the document clusters by pulling similar documents together and pushing dissimilar documents apart. Figure 4 below illustrates how a sample document identified as document 1118 relates to document 3310 with a cosine similarity weight of 0.30600744163358.

```
<edge id="35017" source="1118" target="3310"
weight="0.30600744163358"/>
```

Figure 4: Document similarity as a weighted edge in a GEXF file

Once imported, Gephi calculates the number of nodes and edges present. The graph initially appears as a rather unimpressive rectangle of grey nodes where each node represents a document as presented in Figure 5.

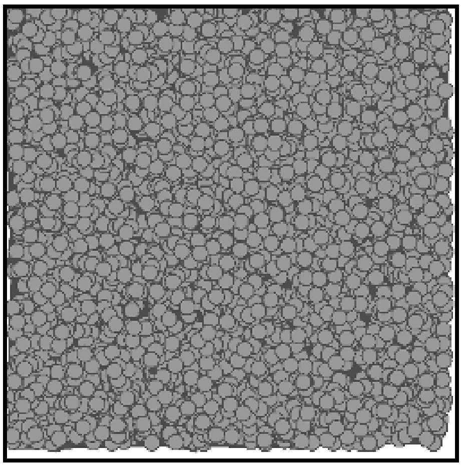
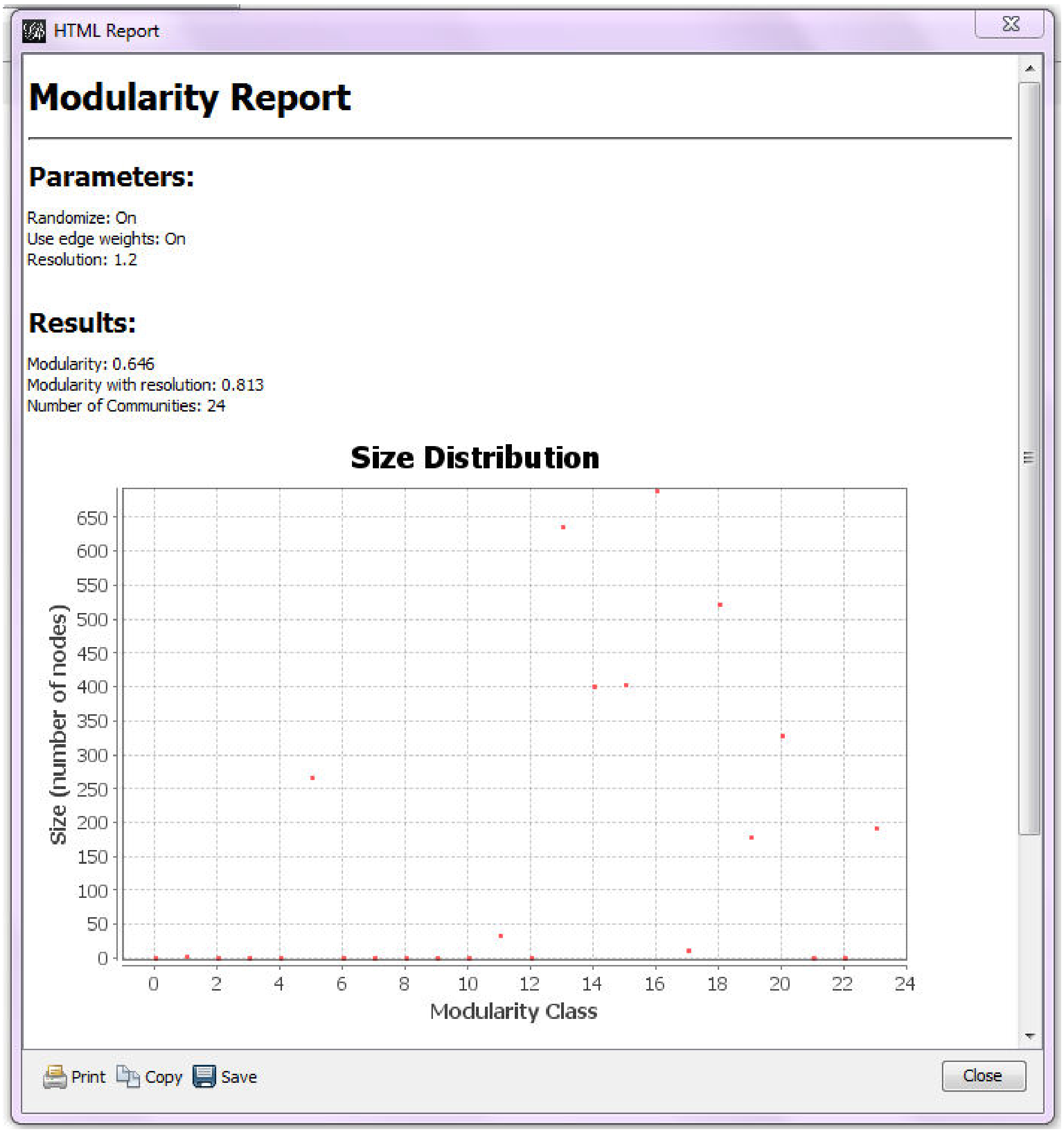


Figure 5: Initial document layout on import into Gephi

4.6.2 Louvain Method

The Louvain method can be executed via the Gephi Statistics panel. The user is given an option to specify the resolution input variable. Lower values yield a higher volume of smaller clusters. Higher values and particularly values greater than 1.0 focus the algorithm in finding larger partitions. A value of 1.2 is used in this experiment. The objective of the study is to find the major themes in the collection, focusing on the term “major” a higher value is set to algorithm to cluster larger communities of content.

The Louvain method loops over all nodes in the graph and puts them in the best community by calculating their modularity. Modularity is a measure between -1 and 1 that compares the density of edges inside communities to edges outside communities. The Louvain method detects small communities first of all by optimising the modularity locally on each node. Each small community is then grouped into one node and the first step is repeated until a maximum of modularity is attained and a hierarchy of communities is produced. It should be noted that nothing ensures that the global maximum of cluster modularity is attained but this problem is not unique to the Louvain method. The randomise option randomly implements the order in which the nodes are being considered for a move to another community. According to the Gephi documentation, running this method with the randomise option can produce a better decomposition. As the order of the nodes is accessed randomly, each instance of the algorithm may yield slight variations in the number of communities found, especially in the detection of smaller ones. Figure 6 illustrates a sample news-centric TF-IDF report.



Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.2

Results:

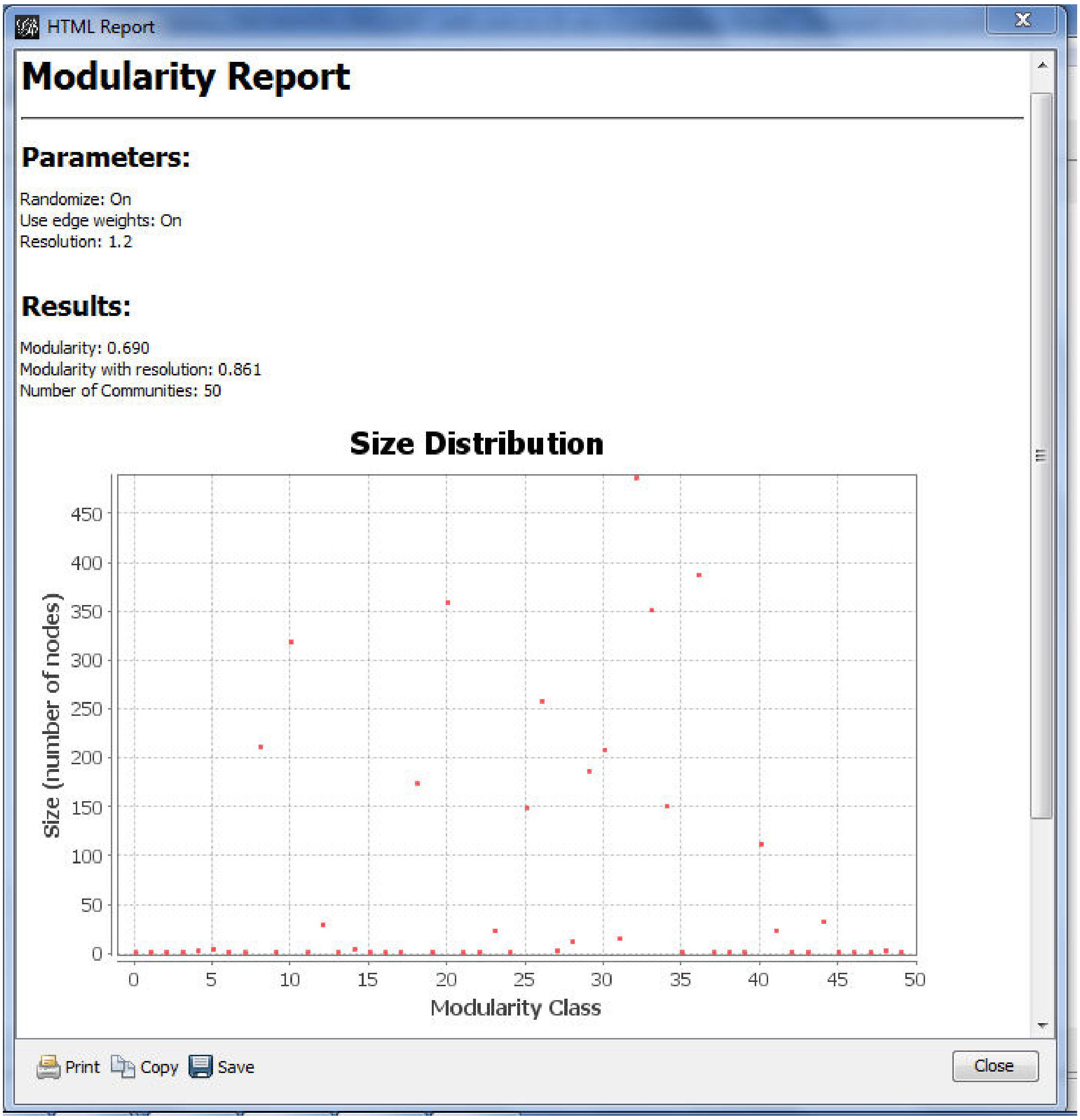
Modularity: 0.646
Modularity with resolution: 0.813
Number of Communities: 24

Figure 6: Sample news-Centric TF-IDF modularity report

Interestingly, a lower number of communities were found with the proposed news-centric TF-IDF approach than found using the traditional TF-IDF approach as presented in Table 5.

Description	Traditional TF-IDF	News-centric TF-IDF
Randomise	True	True
Resolution	1.2	1.2
Modularity	0.690	0.646
Modularity with resolution	0.861	0.813
Number of communities	50	24

Table 5: Modularity statistics for each TF-IDF approach



Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.2

Results:

Modularity: 0.690
Modularity with resolution: 0.861
Number of Communities: 50

Figure 7: Sample traditional TF-IDF modularity report

The number of communities detected using the traditional TF-IDF approach ranged from a minimum of 44 to a maximum of 50. The number of communities detected using the proposed news-centric TF-IDF approach ranged from a minimum of 24 to a maximum of 28. Once the Louvain modularity algorithm has been run and the partitions have been detected by the algorithm, document nodes can be colour coded according to their community cluster. The Gephi application suggests a colour for each detected cluster randomly. As the colour is applied automatically, similar colours can be applied to different clusters which can be misleading especially if the clusters coincidentally appear in close proximity to each other on the graph. Figure 8 illustrates the document nodes colour coded as per their community cluster before a layout algorithm has been applied.

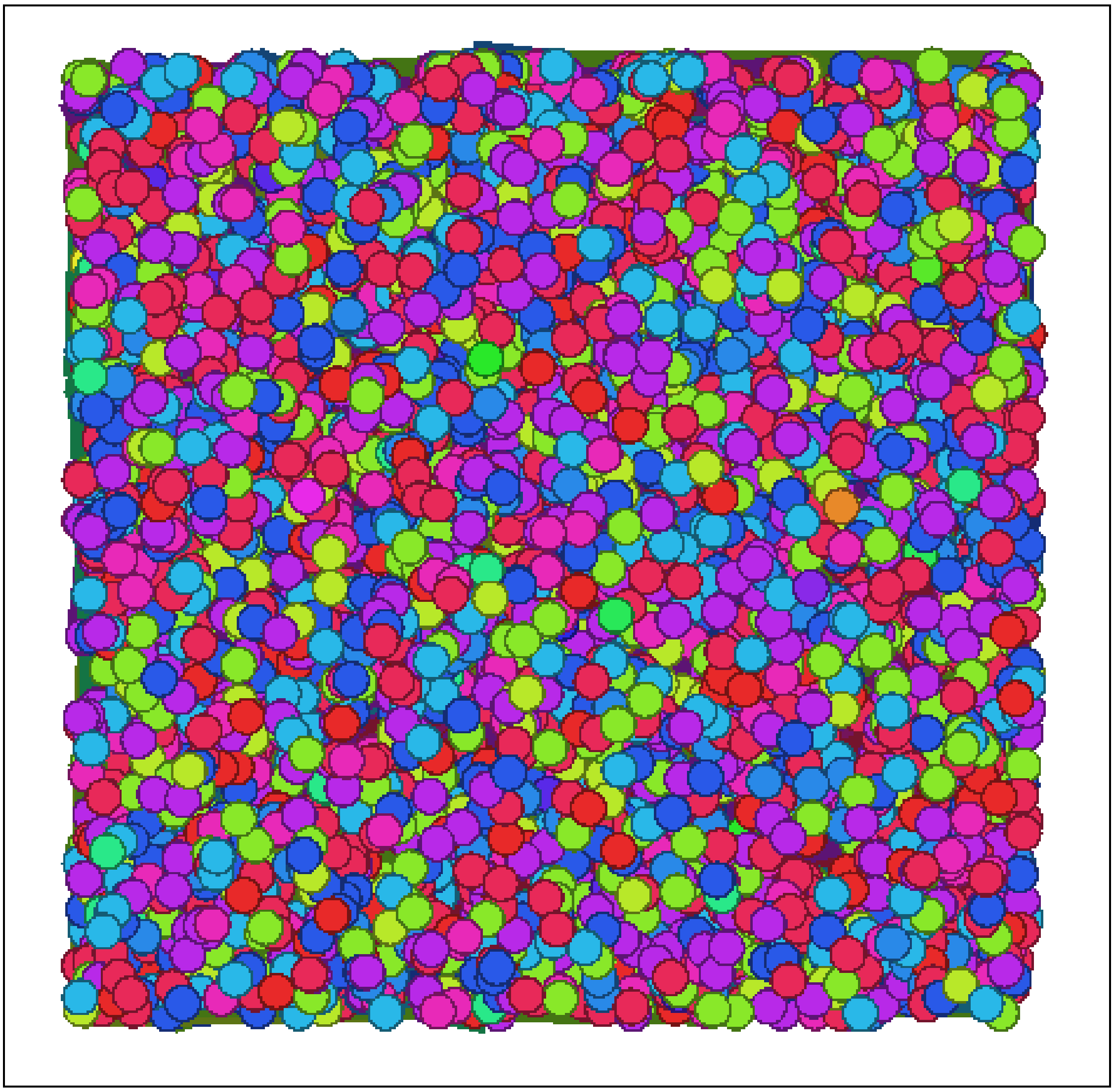


Figure 8: Documents nodes colour-coded per cluster partition

4.6.3 OpenOrd Layout Algorithm

The next step is to visualise the partitioned documents using a layout algorithm. The objective of a layout algorithm is to create a data visualisation of the document clusters. Using the cosine similarity edge weight, the layout algorithm positions document nodes that are connected close to each other and repels document nodes that are not connected. Similar documents are presented in close proximity and dissimilar documents presented far apart from each other. The OpenOrd algorithm proposed by Martin, Brown, Klavans and Klavans (2001) is designed to visualise large undirected weighted graphs and will be used in his experiment. This algorithm is particularly suited for large graphs and has scaled successfully to over 1 million nodes. The OpenOrd layout algorithm encourages larger edge weights to group together.

Figure 9 illustrates the data visualisation of the document collection using the proposed news-centric TF-IDF approach, clustered with the Lovain method of community detection and presented using the OpenOrd layout algorithm. Clusters with fewer than four linked documents have been filtered out.

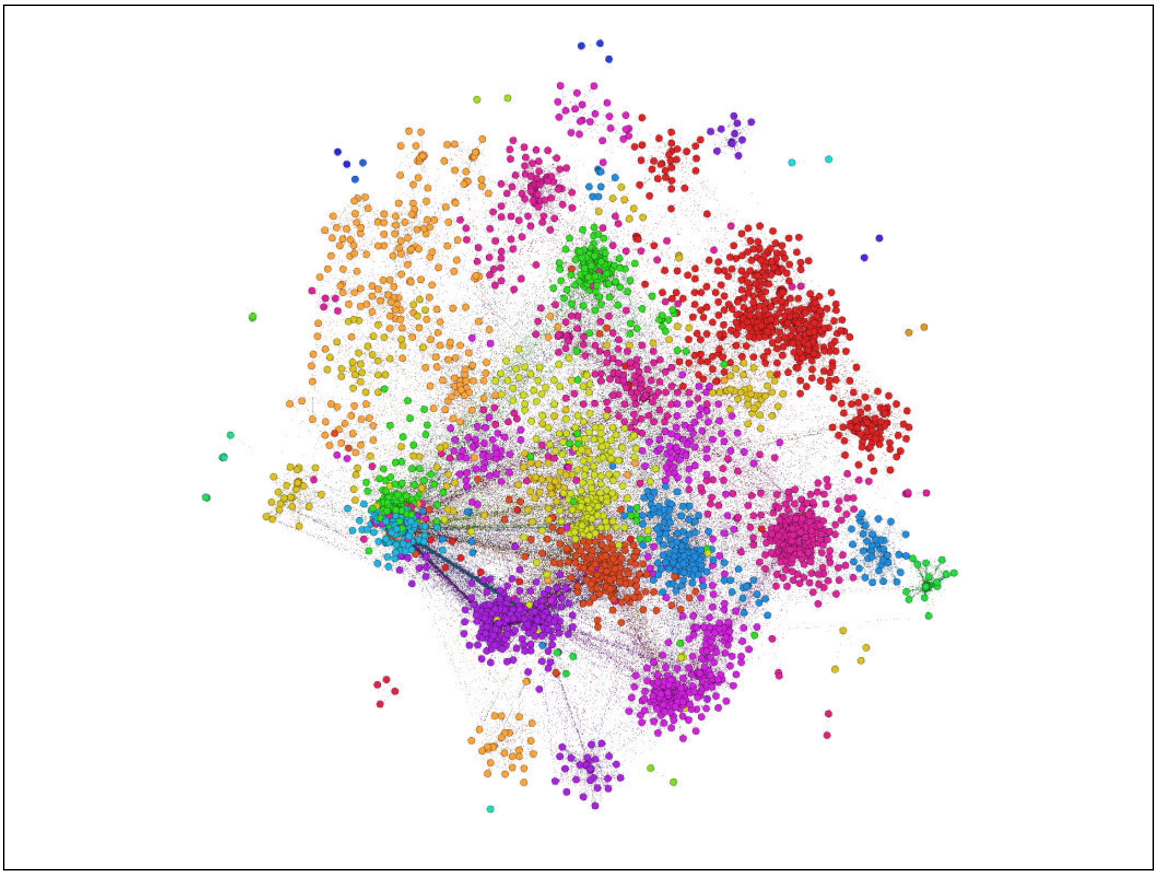


Figure 9: News-centric TF-IDF with Louvain method and OpenOrd layout

Smaller document partitions that contain less than three linked documents are filtered from the graph as they are too small to have any real meaning for the document collection. The remaining partition colours are assessed to ensure that each cluster is presented distinctly. By removing very small partitions it is it is easier to apply a unique and distinctive colour to each remaining document cluster.

4.7 Expert Users

The expert users were provided with a list of articles published on pages one to five in the Irish Times between the dates of the 18th of May 2015 and the 23rd of May 2015. It should be noted that the sample the expert users were asked to evaluate represents a weekly sample of newspaper content. The larger document sample, presented in the visual evaluation section in this chapter, contained over three thousand articles. It is not feasible or practical to ask the expert group to manually categorise this volume of articles. Manually categorising every article would be an onerous and manually intensive task. This is one of the main motivations for attempting to automate the process via a toolset in the first place.

Each user was asked to review the list of articles and where appropriate, manually suggest a topic for each article. The expert users were instructed that providing no topic was also a valid choice. Many articles report a single occurrence event such as a car accident, a press release or a company announcement. A topic by its nature reflects a collection of similarly themed news articles, for example a collection of articles related to the Gay Marriage Referendum or The Rugby World Cup. To qualify for a topic, a suggestion of a minimum of five related articles within the given sample was given as a guideline. The manual categorisation results from each expert user are presented in Tables 6, 7, and 8 below.

Article Headline	**Topic**
Poll bounce for FG as support continues to rise	politics
Fall in offences linked to fewer late bars	Crime
50,000 waiting longer than HSE targets	waiting lists
Boost for those who want Kenny to call early election	Election
Poll details	marriage ref
FG increases lead over SF as it regains confidence of older and rural voters	politics

Gardaí investigate death of student (18)	Crime
Parents of babies who died ‘were lied to’	maternity services
Earlier plan for patient safety body abandoned	Patient safety
Nearly 1,500 cases of mould reported to council	Housing
Yes canvassers say social class is a factor in marriage poll	marriage ref
How to talk to children about issues arising from vote	marriage ref
Bishops united in call for No vote	marriage ref
Citizens happy to share opinions, if not their names	marriage ref
Ronan fears Ireland will be seen as backwards	marriage ref
Royal couple to begin visit at NUIG	Royal visit 2015
US presidential hopeful to fly in for speech	US election 2016
Chinese premier signs two bilateral agreements on farming and visas	Farming / Irish - China trade
Highlights Prince Charles and Camilla’s visit	Royal visit 2015
SF members to join events with prince	Royal visit 2015
Intruders alerted gardaí after discovering two bodies	Rural crime
Majority against taking in fleeing migrants	Migrant crisis
Critical risks at Portlaoise hospital had been raised by medics in 2011	maternity services
Refugees to be given integration courses	Asylum seekers
Playing blame game over Portlaoise is not helping situation	maternity services
EU military mission will target traffickers’ boats	Migrant crisis
Adams and McGuinness to meet Prince Charles during visit to Galway	Royal visit 2015
Tributes to student who died after night out	Drug deaths
Ecstasy can cause variety of adverse health effects	Drug deaths
Review of drug treatment for prisoners	Prison services
Tyrone woman swindled $517,000 from six friends	Julia Holmes
Efforts made to ensure safe prisons, says Minister	Prison services
Bodies may have been at farmhouse for weeks	Julia Holmes
Policing body will be Garda’s ‘critical friend’	Garda Authority
Coalition confident of victory for Yes side in referendum	marriage ref
Adams says regrets expressed on both sides	Royal visit 2015
Fingleton added to list of witnesses to be called up by banking inquiry	Banking inquiry
The risk is that someone will think, ‘This isn’t working, it’s a dud – I’ll take another’	Drug deaths
Man who sold ecstasy to student identified	Drug deaths
Commission convinced it was right to take case	marriage ref
Closely watched test of North’s equality laws	marriage ref
Bakery considers appealing judgment	marriage ref
Arrival in Mullaghmore will bring bittersweet emotions	Royal visit 2015
More sticks than you could shake a prince at	Royal visit 2015
Adams handshake motivated by reconciliation and politics	Royal visit 2015
Prince praises ‘magic’ unique to Ireland	Royal visit 2015
Newgrange to be X-rayed to determine source of granite	?

No vote Carlow-Kilkenny TD ‘broke no rule’	marriage ref
Work on €5m ‘missing link’ of Dublin Bay cycle path begins	Cycling
Independent Newspapers ‘gratuitously identified’ accused	Media
Lack of support for No vote ‘political correctness gone mad’	marriage ref
State plans to introduce elements of family Act	marriage ref
Politicians fail to respond to our fears, says Waters	marriage ref
Archbishop declines to comment on Yes voters	marriage ref
Anti-austerity group backs youth to vote Yes	marriage ref
No selfies, leave the badges at home and remember your ID	marriage ref
Defeat would ‘cost gay children everything’, says ex-president	marriage ref
McAleese says son bullied for being gay	marriage ref
No side ‘trying to change’ referendum’	marriage ref
Referendum over for many ‘special voters’	marriage ref
Yes vote will ‘obliterate prejudice’, says Kenny	marriage ref
Royal couple retrace steps to lay history to rest	Royal visit 2015
All you need to know for tomorrow’s vote	marriage ref
‘No’ supporters open to second referendum	marriage ref
Yes, this vote really is all about the children	marriage ref
‘Healing is possible even when heartache continues’	Royal visit 2015
Prince talks of anguish at death of great-uncle	Royal visit 2015
Charles completes ‘pilgrimage of the heart’ in Sligo	Royal visit 2015
Unions and Minister agree Junior Cert terms	Junior cycle
Student remembered as performer full of life, love and laughter	Drug deaths
Safety fears for children in asylum seekers system	Asylum seekers
O’Brien ‘delighted’ at RTÉ injunction	Denis O'Brien / or Media
‘Jeanie Johnston’ famine ship value sinks from €15 million to €700,000	?
Hurley boxes clever over two rounds	Banking inquiry
Anglo was solvent on night of guarantee, says former Central Bank governor	Banking inquiry
Government expected to stand behind its banks, inquiry told	Banking inquiry
Dunnes Stores closes Wexford supermarket with about 100 staff	Dunnes Stores
Cork has had most weather alerts since 2012	Weather
Most tribunal documents will not be archived	Mahon tribunal
Varadkar marks 30th anniversary of first heart transplant in Ireland	heart transplant
Schools to be allocated extra resource teachers for pupils with special needs	Special needs
€2.3m spent on hotels for families	Homelessness
Parents set up vaccine support group	Cervical cancer
Former support group director on fraud charges	Crime
Bishops consider ‘open’ confession for children	Church reform
Unions to decide on new Junior cycle deal	Junior cycle
Aislinn centre marks sixth anniversary of Ryan report	Ryan report

Six people taken to hospital from club on night student collapsed	Drug deaths
Gardaí confirm identity of man found dead in Limerick house	Rural crime
HSE begins heroin antidote plan	Drug deaths
New frozen berry alert following Swedish deaths	Food safety
Case study It’s freezing. That’s why I have this furry blanket. I share it with my mam	Asylum seekers
waiting lists	Child poverty
Campaigners believe high turnout will carry Yes vote	marriage ref
Obstetricians query inclusion of mothers in group	marriage ref
It’s a long road back to douze points for Ireland	euro vision
When Gerry met Charles – the moving and shaking	Royal visit 2015
Online weekirishtimes.com	?
Threat to Palmyra attacks the core ideals of the West	Isis
Latest agreement allows both sides to save face	?
Unions to ballot on junior cycle reform	Junior cycle
Prince calls on North to free itself from past	Royal visit 2015
Were some voters breaking the law?	marriage ref
Turnout higher than in recent referendums	marriage ref
Gráinne Courtney and Orla Howard	marriage ref
Social media plays major role in Irish poll for first time	marriage ref
Emigrants come home to vote for marriage	marriage ref

Table 6: Expert user 1 David Labanyi topic categorisation results

Article Headline	**Topic**
Poll bounce for FG as support continues to rise	Poll
Fall in offences linked to fewer late bars	
50,000 waiting longer than HSE targets	
Boost for those who want Kenny to call early election	Vote 2016
Poll details	Poll
FG increases lead over SF as it regains confidence of older and rural voters	Vote 2016
Gardaí investigate death of student (18)	
Parents of babies who died ‘were lied to’	
Earlier plan for patient safety body abandoned	
Nearly 1,500 cases of mould reported to council	
Yes canvassers say social class is a factor in marriage poll	Same-sex marriage referendum
How to talk to children about issues arising from vote	Same-sex marriage referendum
Bishops united in call for No vote	Same-sex marriage referendum
Citizens happy to share opinions, if not their names	Same-sex marriage referendum
Ronan fears Ireland will be seen as backwards	Same-sex marriage referendum

Royal couple to begin visit at NUIG	Prince Charles Visit
US presidential hopeful to fly in for speech	
Chinese premier signs two bilateral agreements on farming and visas	
Highlights Prince Charles and Camilla’s visit	Prince Charles Visit
SF members to join events with prince	Prince Charles Visit
Intruders alerted gardaí after discovering two bodies	
Majority against taking in fleeing migrants	Migrant Crisis
Critical risks at Portlaoise hospital had been raised by medics in 2011	
Refugees to be given integration courses	Migrant Crisis
Playing blame game over Portlaoise is not helping situation	
EU military mission will target traffickers’ boats	Migrant Crisis
Adams and McGuinness to meet Prince Charles during visit to Galway	Prince Charles Visit
Tributes to student who died after night out	
Ecstasy can cause variety of adverse health effects	
Review of drug treatment for prisoners	
Tyrone woman swindled $517,000 from six friends	
Efforts made to ensure safe prisons, says Minister	
Bodies may have been at farmhouse for weeks	
Policing body will be Garda’s ‘critical friend’	
Coalition confident of victory for Yes side in referendum	Same-sex marriage referendum
Adams says regrets expressed on both sides	Prince Charles Visit
Fingleton added to list of witnesses to be called up by banking inquiry	Banking Inquiry
The risk is that someone will think, ‘This isn’t working, it’s a dud – I’ll take another’	
Man who sold ecstasy to student identified	
Commission convinced it was right to take case	
Closely watched test of North’s equality laws	Same-sex marriage referendum
Bakery considers appealing judgment	Same-sex marriage referendum
Arrival in Mullaghmore will bring bittersweet emotions	Prince Charles Visit
More sticks than you could shake a prince at	Prince Charles Visit
Adams handshake motivated by reconciliation and politics	Prince Charles Visit
Prince praises ‘magic’ unique to Ireland	Prince Charles Visit
Newgrange to be X-rayed to determine source of granite	
No vote Carlow-Kilkenny TD ‘broke no rule’	Same-sex marriage referendum
Work on €5m ‘missing link’ of Dublin Bay cycle path begins	
Independent Newspapers ‘gratuitously identified’ accused	
Lack of support for No vote ‘political correctness gone mad’	Same-sex marriage referendum

State plans to introduce elements of family Act	Same-sex marriage referendum
Politicians fail to respond to our fears, says Waters	Same-sex marriage referendum
Archbishop declines to comment on Yes voters	Same-sex marriage referendum
Anti-austerity group backs youth to vote Yes	Same-sex marriage referendum
No selfies, leave the badges at home and remember your ID	Same-sex marriage referendum
Defeat would 'cost gay children everything', says ex-president	Same-sex marriage referendum
McAleese says son bullied for being gay	Same-sex marriage referendum
No side 'trying to change' referendum'	Same-sex marriage referendum
Referendum over for many 'special voters'	Same-sex marriage referendum
Yes vote will 'obliterate prejudice', says Kenny	Same-sex marriage referendum
Royal couple retrace steps to lay history to rest	Prince Charles Visit
All you need to know for tomorrow's vote	Same-sex marriage referendum
'No' supporters open to second referendum	Same-sex marriage referendum
Yes, this vote really is all about the children	Same-sex marriage referendum
'Healing is possible even when heartache continues'	Prince Charles Visit
Prince talks of anguish at death of great-uncle	Prince Charles Visit
Charles completes 'pilgrimage of the heart' in Sligo	Prince Charles Visit
Unions and Minister agree Junior Cert terms	Exam Watch
Student remembered as performer full of life, love and laughter	
Safety fears for children in asylum seekers system	
O'Brien 'delighted' at RTÉ injunction	
'Jeanie Johnston' famine ship value sinks from €15 million to €700,000	
Hurley boxes clever over two rounds	Banking Inquiry
Anglo was solvent on night of guarantee, says former Central Bank governor	Banking Inquiry
Government expected to stand behind its banks, inquiry told	Banking Inquiry
Dunnes Stores closes Wexford supermarket with about 100 staff	
Cork has had most weather alerts since 2012	
Most tribunal documents will not be archived	
Varadkar marks 30th anniversary of first heart transplant in Ireland	
Schools to be allocated extra resource teachers for pupils with	

special needs	
€2.3m spent on hotels for families	
Parents set up vaccine support group	
Former support group director on fraud charges	
Bishops consider ‘open’ confession for children	
Unions to decide on new Junior cycle deal	Exam Watch
Aislinn centre marks sixth anniversary of Ryan report	
Six people taken to hospital from club on night student collapsed	
Gardaí confirm identity of man found dead in Limerick house	
HSE begins heroin antidote plan	
New frozen berry alert following Swedish deaths	
Case study It’s freezing. That’s why I have this furry blanket. I share it with my mam	
waiting lists	
Campaigners believe high turnout will carry Yes vote	Same-sex marriage referendum
Obstetricians query inclusion of mothers in group	
It’s a long road back to douze points for Ireland	
When Gerry met Charles – the moving and shaking	Prince Charles Visit
Online weekirishtimes.com	
Threat to Palmyra attacks the core ideals of the West	Syria Crisis
Latest agreement allows both sides to save face	
Unions to ballot on junior cycle reform	Exam Watch
Prince calls on North to free itself from past	Prince Charles Visit
Were some voters breaking the law?	Same-sex marriage referendum
Turnout higher than in recent referendums	Same-sex marriage referendum
Gráinne Courtney and Orla Howard	Same-sex marriage referendum
Social media plays major role in Irish poll for first time	Same-sex marriage referendum
Emigrants come home to vote for marriage	Same-sex marriage referendum

Table 7: Expert user 2 Luke Cassidy topic categorisation results

Article Headline	**Suggested Topic**
Poll bounce for FG as support continues to rise	politics
Fall in offences linked to fewer late bars	
50,000 waiting longer than HSE targets	health
Boost for those who want Kenny to call early election	Election 2016
Poll details	referendum
FG increases lead over SF as it regains confidence of older and rural voters	Election 2016

Gardaí investigate death of student (18)	
Parents of babies who died 'were lied to'	
Earlier plan for patient safety body abandoned	
Nearly 1,500 cases of mould reported to council	
Yes canvassers say social class is a factor in marriage poll	Same-sex marriage referendum
How to talk to children about issues arising from vote	Same-sex marriage referendum
Bishops united in call for No vote	Same-sex marriage referendum
Citizens happy to share opinions, if not their names	Same-sex marriage referendum
Ronan fears Ireland will be seen as backwards	Same-sex marriage referendum
Royal couple to begin visit at NUIG	
US presidential hopeful to fly in for speech	US elections
Chinese premier signs two bilateral agreements on farming and visas	
Highlights Prince Charles and Camilla's visit	
SF members to join events with prince	Royal Visit
Intruders alerted gardaí after discovering two bodies	
Majority against taking in fleeing migrants	Migrant crisis
Critical risks at Portlaoise hospital had been raised by medics in 2011	Portlaoise hospital
Refugees to be given integration courses	
Playing blame game over Portlaoise is not helping situation	Portlaoise hospital
EU military mission will target traffickers' boats	
Adams and McGuinness to meet Prince Charles during visit to Galway	Royal Visit
Tributes to student who died after night out	
Ecstasy can cause variety of adverse health effects	health
Review of drug treatment for prisoners	health
Tyrone woman swindled $517,000 from six friends	
Efforts made to ensure safe prisons, says Minister	
Bodies may have been at farmhouse for weeks	
Policing body will be Garda's 'critical friend'	Garda controversy
Coalition confident of victory for Yes side in referendum	Same-sex marriage referendum
Adams says regrets expressed on both sides	Royal Visit
Fingleton added to list of witnesses to be called up by banking inquiry	Banking Inquiry
The risk is that someone will think, 'This isn't working, it's a dud – I'll take another'	
Man who sold ecstasy to student identified	
Commission convinced it was right to take case	
Closely watched test of North's equality laws	Same-sex marriage referendum

Bakery considers appealing judgment	Same-sex marriage referendum
Arrival in Mullaghmore will bring bittersweet emotions	Royal Visit
More sticks than you could shake a prince at	Royal Visit
Adams handshake motivated by reconciliation and politics	Royal Visit
Prince praises ‘magic’ unique to Ireland	Royal Visit
Newgrange to be X-rayed to determine source of granite	
No vote Carlow-Kilkenny TD ‘broke no rule’	Same-sex marriage referendum
Work on €5m ‘missing link’ of Dublin Bay cycle path begins	
Independent Newspapers ‘gratuitously identified’ accused	
Lack of support for No vote ‘political correctness gone mad’	Same-sex marriage referendum
State plans to introduce elements of family Act	Same-sex marriage referendum
Politicians fail to respond to our fears, says Waters	Same-sex marriage referendum
Archbishop declines to comment on Yes voters	Same-sex marriage referendum
Anti-austerity group backs youth to vote Yes	Same-sex marriage referendum
No selfies, leave the badges at home and remember your ID	Same-sex marriage referendum
Defeat would ‘cost gay children everything’, says ex-president	Same-sex marriage referendum
McAleese says son bullied for being gay	Same-sex marriage referendum
No side ‘trying to change’ referendum’	Same-sex marriage referendum
Referendum over for many ‘special voters’	Same-sex marriage referendum
Yes vote will ‘obliterate prejudice’, says Kenny	Same-sex marriage referendum
Royal couple retrace steps to lay history to rest	Royal Visit
All you need to know for tomorrow’s vote	Same-sex marriage referendum
‘No’ supporters open to second referendum	Same-sex marriage referendum
Yes, this vote really is all about the children	Same-sex marriage referendum
‘Healing is possible even when heartache continues’	Royal Visit
Prince talks of anguish at death of great-uncle	Royal Visit
Charles completes ‘pilgrimage of the heart’ in Sligo	Royal Visit
Unions and Minister agree Junior Cert terms	Exam Watch
Student remembered as performer full of life, love and laughter	
Safety fears for children in asylum seekers system	

O'Brien 'delighted' at RTÉ injunction	Siteserv controversy
'Jeanie Johnston' famine ship value sinks from €15 million to €700,000	
Hurley boxes clever over two rounds	Banking Inquiry
Anglo was solvent on night of guarantee, says former Central Bank governor	Banking Inquiry
Government expected to stand behind its banks, inquiry told	Banking Inquiry
Dunnes Stores closes Wexford supermarket with about 100 staff	
Cork has had most weather alerts since 2012	Weather
Most tribunal documents will not be archived	Mahon tribunal
Varadkar marks 30th anniversary of first heart transplant in Ireland	
Schools to be allocated extra resource teachers for pupils with special needs	
€2.3m spent on hotels for families	
Parents set up vaccine support group	
Former support group director on fraud charges	
Bishops consider 'open' confession for children	Religion
Unions to decide on new Junior cycle deal	Exam Watch
Aislinn centre marks sixth anniversary of Ryan report	Ryan report
Six people taken to hospital from club on night student collapsed	
Gardaí confirm identity of man found dead in Limerick house	
HSE begins heroin antidote plan	health
New frozen berry alert following Swedish deaths	health
Case study It's freezing. That's why I have this furry blanket. I share it with my mam	
waiting lists	
Campaigners believe high turnout will carry Yes vote	Same-sex marriage referendum
Obstetricians query inclusion of mothers in group	
It's a long road back to douze points for Ireland	Eurovision
When Gerry met Charles – the moving and shaking	Royal Visit
Online weekirishtimes.com	
Threat to Palmyra attacks the core ideals of the West	
Latest agreement allows both sides to save face	
Unions to ballot on junior cycle reform	Exam Watch
Prince calls on North to free itself from past	Royal Visit
Were some voters breaking the law?	Same-sex marriage referendum
Turnout higher than in recent referendums	Same-sex marriage referendum
Gráinne Courtney and Orla Howard	
Social media plays major role in Irish poll for first time	Same-sex marriage referendum

Emigrants come home to vote for marriage	Same-sex marriage referendum

Table 8: Expert user 3 Patrick Logue topic categorisation results

Some of the topics nominated by the expert users differ slightly by name but the label clearly refers to the same topic. For example, one user nominated the topic “Same-sex marriage referendum” and another nominated “marriage ref”. Whilst these are different labels, they clearly relate to the same topic. Similarly, stories relating to the visit by Prince Charles were described as “Royal visit 2015” by one expert user and the “Prince Charles Visit” by another. A subjective judgement was applied that these topics were in fact the same.

In some cases, topics were suggested by expert users, where insufficient articles relating to that topic existed in the sample provided. One example of this was the suggested topic of “cervical cancer”. Only one article within the document sample provided related to this topic. As this article had little or no overlapping terms when compared to all other documents in the sample collection it can be argued that it is not topic related, in the context of the sample given. In general, on the larger and more obvious topics all expert users agreed with each other. On broader topics however, the expert users differed in their level of granularity. Articles with a health theme for example, were categorised by one expert user using the generic “health” label whilst another expert user provided sub-topics within an overall health topic such as “maternity services”, “waiting lists” and “patient safety”.

4.8 Visual Results

As a measure of how coherently the documents are grouped together, data visualisations of the clustered larger document collection will be assessed for each approach.

4.8.1 Visual Results of News-Centric TF-IDF Approach

Figure 11 presents a data visualisation of the document collection with the news-centric TF-IDF, Louvain method and OpenOrd layout algorithm applied. Clusters of documents with less than five articles have been filtered out as they are unlikely to be major themes in the document collection. A total of fourteen smaller clusters were filtered out from the graph resulting in fourteen clusters to be evaluated.

As an initial observation, the OpenOrd layout algorithm has detected a number of smaller document clusters within the larger cluster structure detected by the Louvain method. This explains why some of the cluster numbers are repeated in Figure 10. The reasons why smaller document clusters exist within larger community clusters will need to be examined upon exploring the graph further.

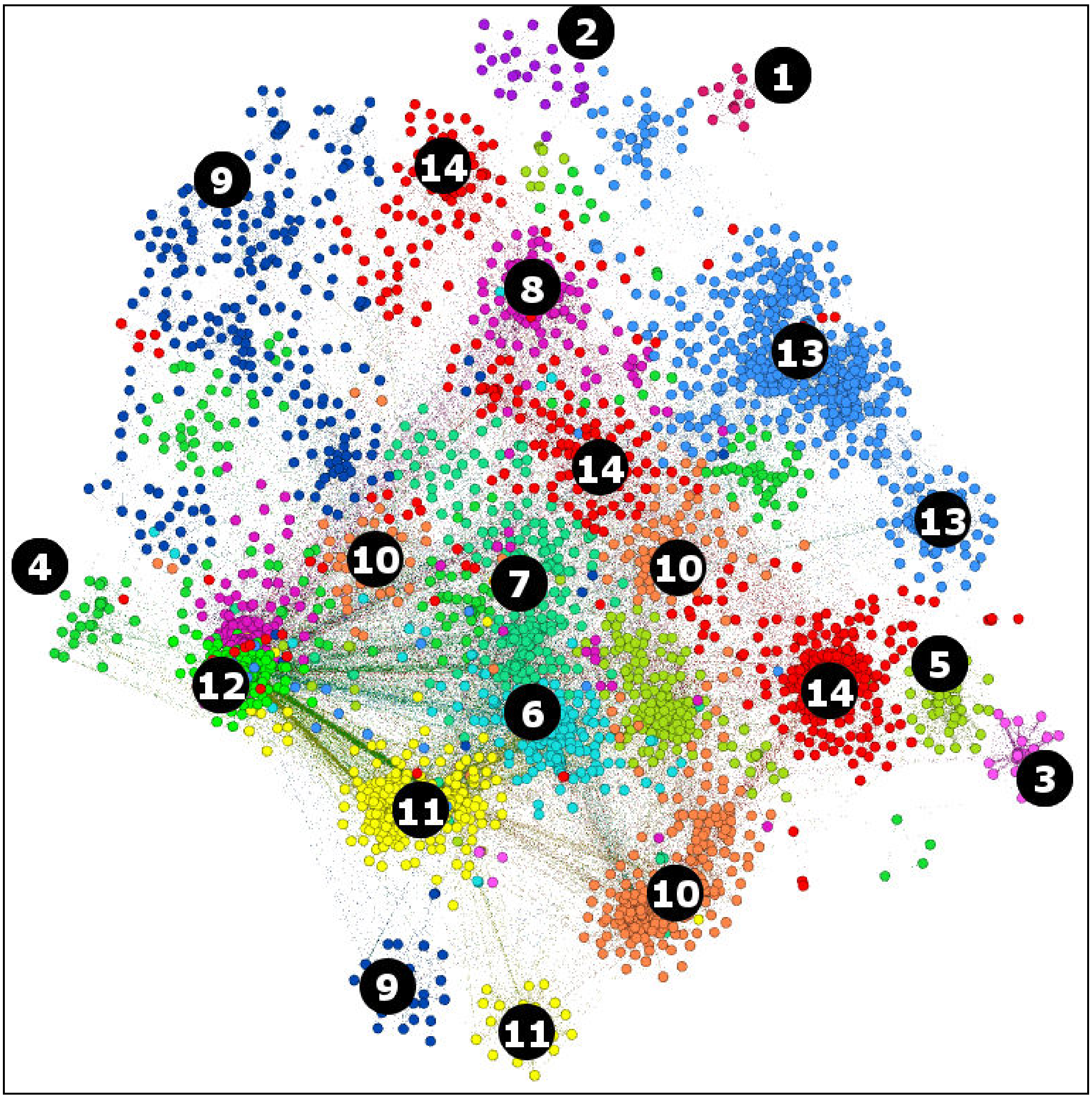


Figure 10: News-centric TF-IDF document collection annotated

Clusters of documents with less than five articles have been filtered out as they are unlikely to be major themes in the document collection. A total of 14 smaller clusters were filtered out from the graph resulting in 14 clusters to be evaluated.

News-centric TF-IDF Cluster 1 & 2

The content of first and second clusters is somewhat underwhelming from a topic aggregation point of view but does highlight that the content is highly related in each cluster which is a promising start for the proposed TF-IDF approach. Cluster number 1 groups a number of golf match results together as presented in Figure 11. The key terms emerging from this cluster are singles, fourball and medal. Similarly, cluster 2 groups a number of horse racing articles together as presented in Figure 12 and the key terms emerging from this cluster are Curragh, Punchestown, Cheltenham and Epsom which are all horse racing courses. This highlights the regular coverage the Irish Times allocate to these sporting events.

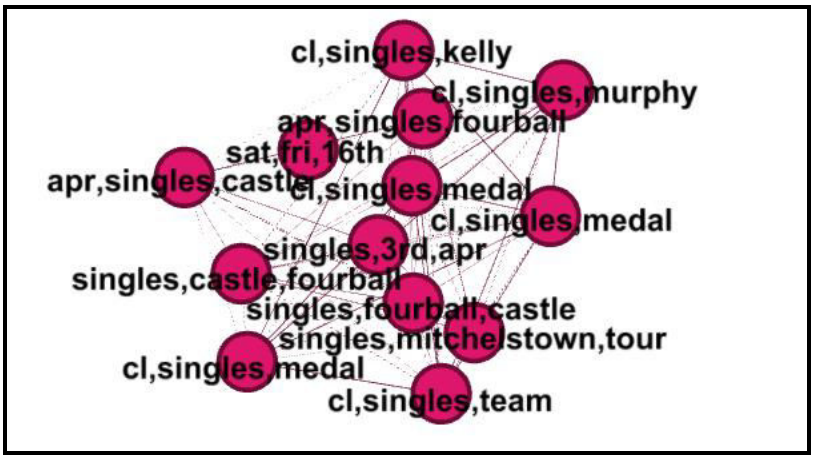


Figure 11: News-centric TF-IDF golf matches

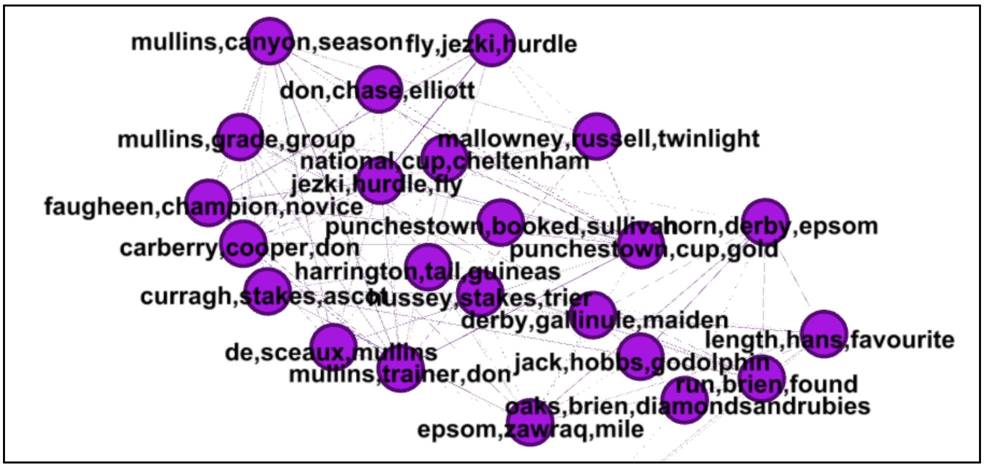


Figure 12: News-centric TF-IDF horse racing

The third cluster as illustrated in Figure 13 presents all of the Irish language articles together in one distinctive cluster. This content is highly related to each other, and not related to other content in the collection. This again shows promise for the proposed document clustering method. This cluster features on the peripheries of the graph as there are little or no overlapping terms with any other articles in the collection. Interestingly, the most common terms emerging from this cluster are the words agus, é and ag. These Gaelic words equate to the English words, “and”, “him” and “at” respectively. All of these words were identified as stopwords in the English language and were removed from the collection early on. The fact these terms are the most commonly used terms in the Irish language articles illustrates how these type of stopwords occur frequently in the text but offer no meaning regarding the content of the article. This further justifies the removal of these types of words from the document collection.

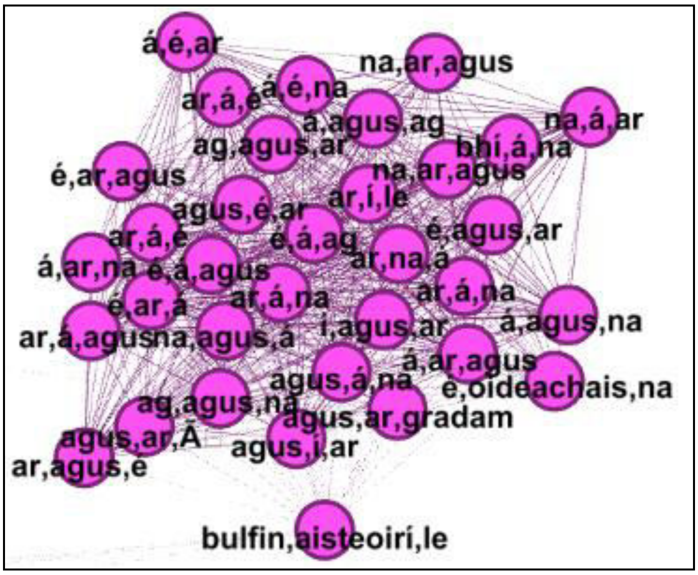


Figure 13: News-centric TF-IDF Irish language

News-centric TF-IDF Cluster 4

The fourth cluster is one of the examples where there Lovain partitioning method identifies one large community or cluster, but the layout algorithm presents clusters of related content within the larger cluster community as illustrated in Figure 14.

The first mini-cluster in fourth cluster covers the visit by Prince Charles to Ireland as illustrated in Figure 15. This was a major news topic during the publishing window. The terms handshake, Gerry and Adams as well as Mountbatten, Prince and Charles emerge from this mini-cluster. Without reading every single article in the collection the key areas of interest of this topic are presented.

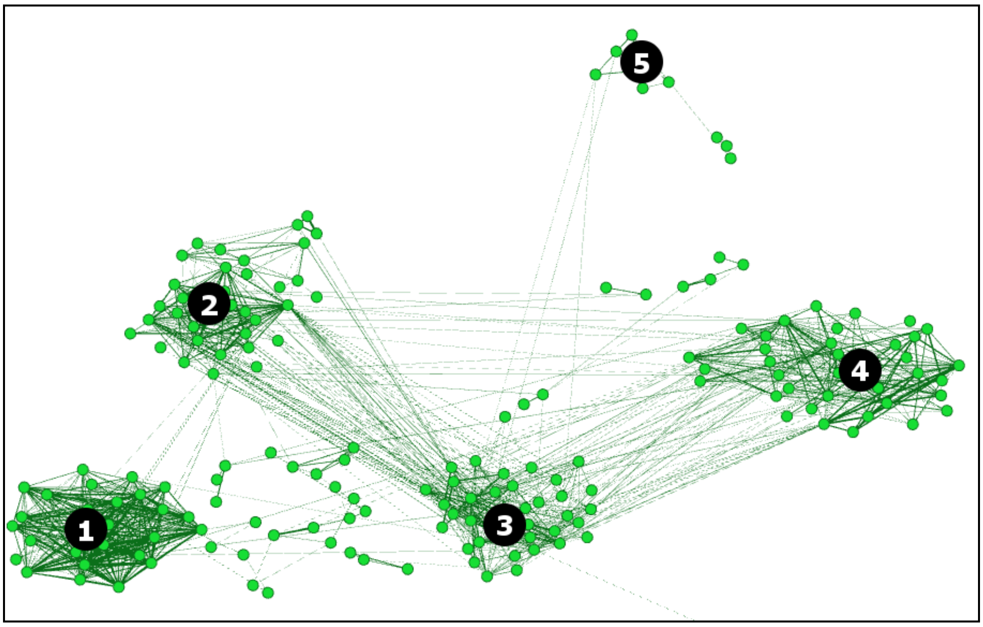


Figure 14: News-centric TF-IDF – Cluster 4

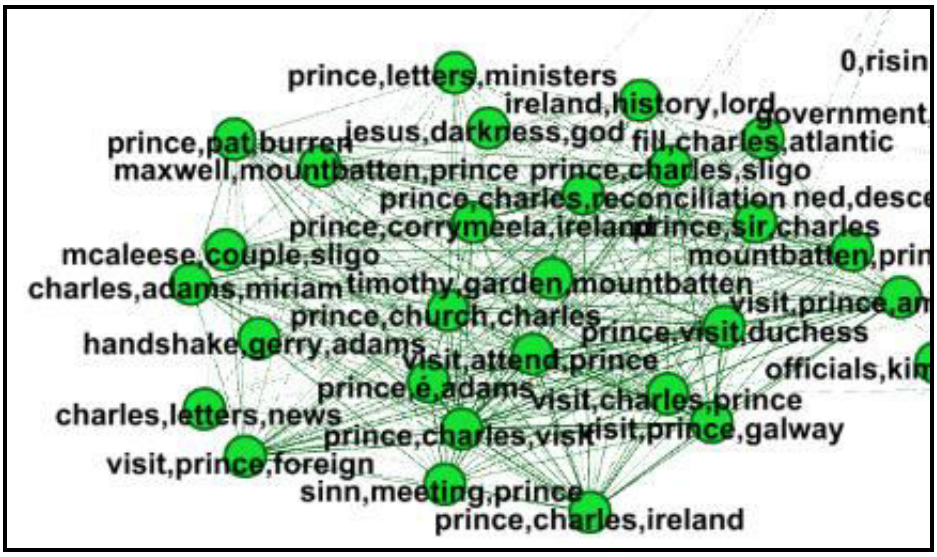


Figure 15: News-centric TF-IDF Prince Charles visits Ireland

The second mini-cluster in the fourth cluster community groups a series of First World War articles together. The terms memorial, war, military, Japan, Berlin and world emerge as the key terms. These articles most likely relate to the current series in the Irish Times that marks the 100th anniversary since the beginning of World War 1. Whilst this is not a brand new news topic, it is a topic of related content in the newspaper and the document clustering method has identified it as such. The relationship between these articles and Prince Charles' visit to Ireland is not instantly clear. The only link that can be found between the two mini-clusters is a relationship between two articles found in the first mini-cluster with the terms visit, attend, prince and Charles with a document in the second mini clusters with the top terms President, visit and forces. The President of Ireland, Michael D. Higgins was the official host of Prince Charles' visit. In the second mini-cluster, Michael D. Higgins represented Ireland in a war memorial service visit to Turkey. The terms president and visit and Higgins overlap between these topics. The Louvain modularity method did not appear to be able to sufficiently distinguish between these topics as it allocated them to one master cluster. The OpenOrd layout algorithm however could distinguish that groups of content existed within this larger cluster albeit that a human can determine that these topics were weakly related. This highlights one of the main limitations with document clustering, links between documents may be drawn because of overlapping terms taken out of context. In this example, the links between Prince Charles' visit and a series of war memorial articles have no significance or relevance. It also highlights how public figures such as the President will play a role in multiple topics related to Irish news. This may create a relationship between articles that would not otherwise exist.

The third mini-cluster in the fourth cluster community appears to have an educatuib and student theme with the terms "students, school and book" emerging. The fourth mini cluster groups articles related to the sale of art and painting together with the terms "art, auction, painting and Picasso" present. The fifth mini-cluster groups articles related to the film festival at Cannes together. Again, the relationship between all of these mini-clusters is not instantly obvious other than the content is generally cultural in theme.

News-centric TF-IDF Cluster 5

The fifth detected cluster in the document collection has also detected mini clusters as presented in Figure 16. This cluster however has a much more coherent theme with the majority content relating to an industrial relations or union issue.

The first mini-cluster as presented in Figure 17 presents the on-going industrial disputes at Dunnes Stores relating to the lack of contracts of employment for part-time workers and the closure of the Gorey store with no notice to staff. The terms "Dunnes, unions, industrial, contracts, pay and workers" emerge in this cluster. Again, without reading every document in this collection, the key areas of interest are disclosed.

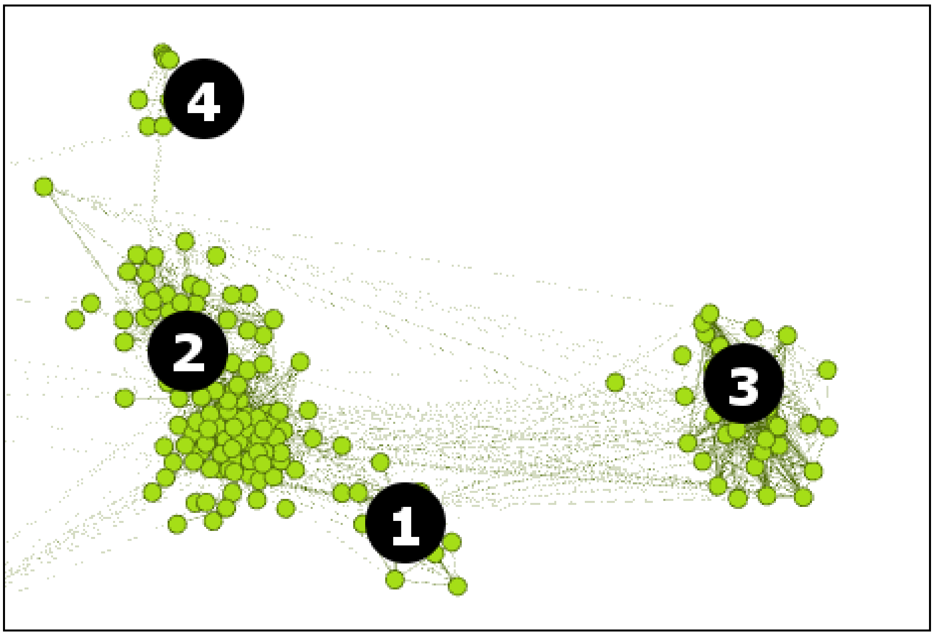


Figure 16: News-centric TF-IDF -- Cluster 5 mini-clusters annotated

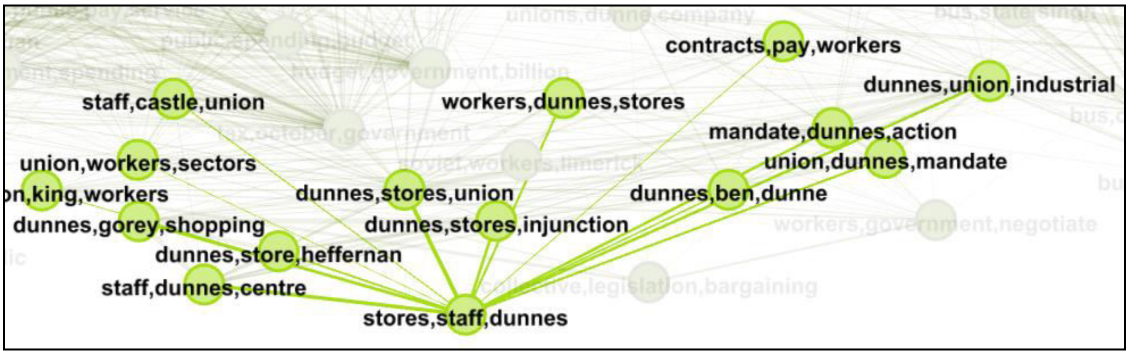


Figure 17: News-centric TF-IDF Dunnes Stores industrial dispute

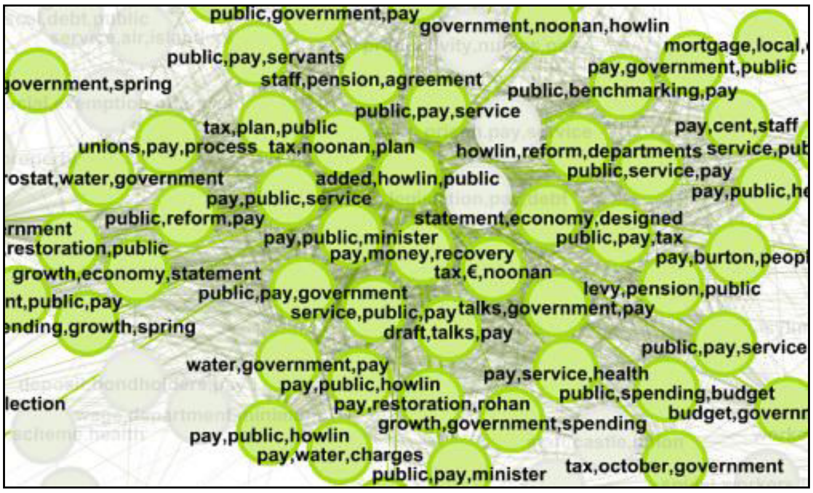


Figure 18: News-centric TF-IDF public pay talks

The second mini-cluster, as presented in Figure 18, groups a series of articles related in one sense or another to the public service pay talks, with the terms public, pay, restoration, spending, budget, Howlin, government and teachers emerging. The third mini-cluster, as illustrated in Figure 19, presents the ongoing industrial dispute at Bus Éireann and Dublin Bus relating to the proposed privatisation of certain bus routes in Dublin.

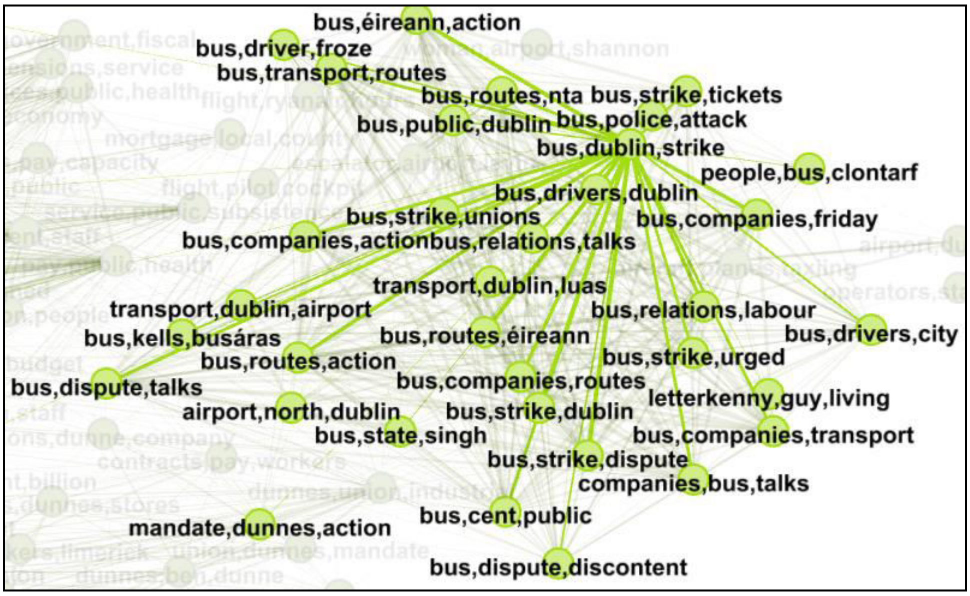


Figure 19 News-centric TF-IDF Bus Éireann & Dublin bus industrial dispute

The final mini-cluster is poorly related to all of the other content. It groups beauty articles together with terms like “skin, foundation, pores and translucent powder” emerging. The only link between these articles and the rest of the cluster appears to be the use of the term water. Within the public pay cluster there are a few articles relating to water charges and protests. The content within the fourth mini-cluster is tightly coupled within the mini-cluster but poorly related to the rest of the cluster.

News-centric TF-IDF Cluster 6

The fifth cluster groups a broad range of content with a relationship to Europe or the E.U. The major themes in this cluster include the Greek debt crisis as illustrated in Figure 20 and the migrant boat crisis across the E.U as illustrated in Figure 21. The terms “Greece, bailout, Euro, European commission, exit, Draghi, Tspiras and Merkel” emerge from the Greek debt crisis cluster. With only a few terms, the cluster instantly paints a picture of the main themes in this crisis. The rest of the cluster includes a very broad range of European related content including among other things the Eurovision song contest. Interestingly, it has picked up on the E.U. membership debate in the U.K. picking up terms British, Europe, referendum and Cameron in a small group of related articles.

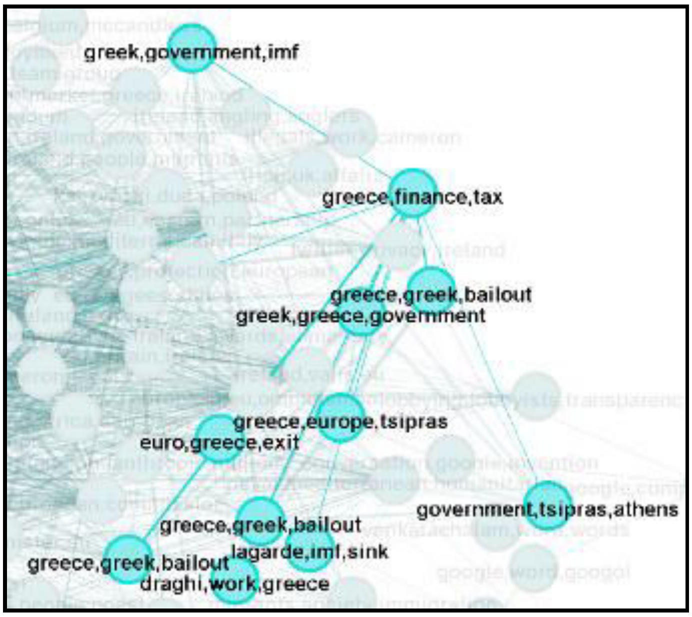


Figure 20: News-centric TF-IDF Greek debt crisis

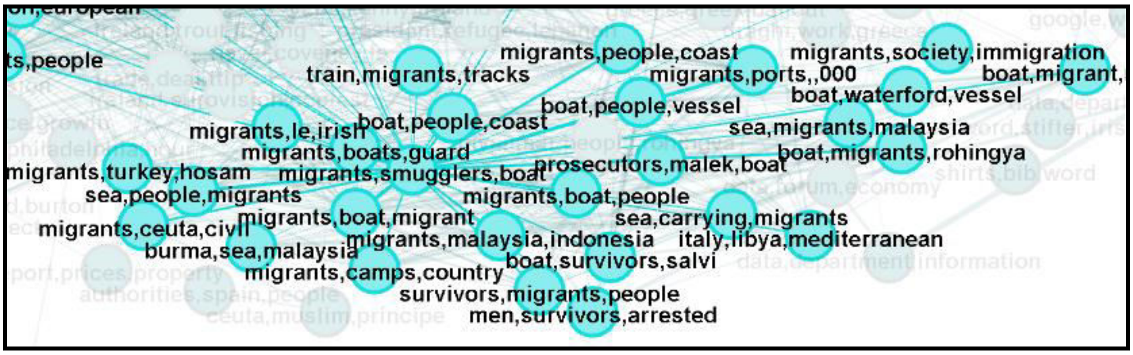


Figure 21: News-centric TF-IDF EU migrant boat crisis

News-centric TF-IDF Cluster 7

Cluster 7 presents an education theme as illustrated in Figure 22. The content is highly coupled with each other in this cluster but no specific topic emerges.

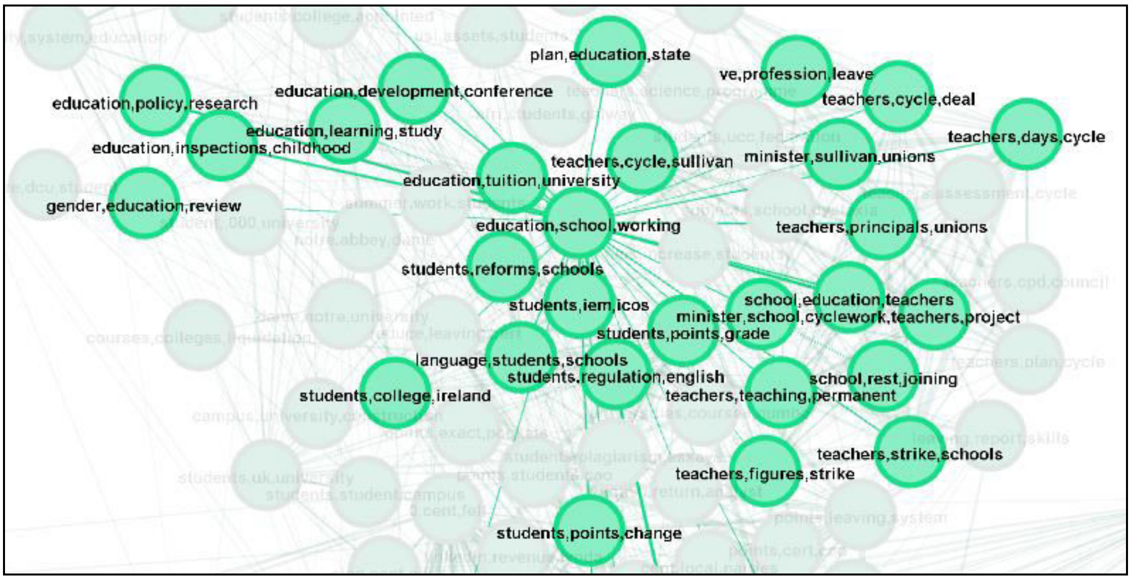


Figure 22: News-centric TF-IDF Education

News-centric TF-IDF Cluster 8

Cluster 8 covers a broad range of health and social issues as illustrated in Figure 23. The layout algorithm has identified five mini-clusters within the larger cluster community covering a broad range of related content relating to alcohol, children, obesity, food and nutrition, parenting, the Health Service Executive (H.S.E.), health budgets and maternity hospitals. There is one mini cluster that is poorly related to the rest of the content. Articles relating to the Guerin report into Garda misconduct are included in this cluster. The link to other health articles appears to be driven by the provision of government reports. The Guerin report has been pulled together on the graph to other health related reports into deaths at certain hospitals.

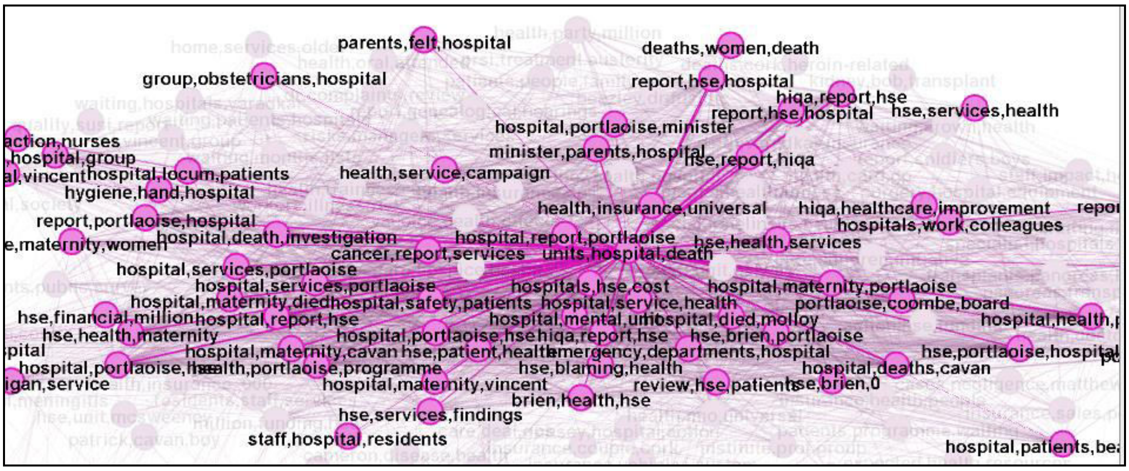


Figure 23: News-centric TF-IDF Health

News-centric TF-IDF Cluster 9

Cluster 9 is focused on World news the content is quite diverse. It ranges from foreign policy, U.S. politics, unrest in the Arab world and the Russian/Ukraine crisis as illustrated in Figures 24 and 25. From a data exploratory perspective, it is interesting to see how this content is linked together

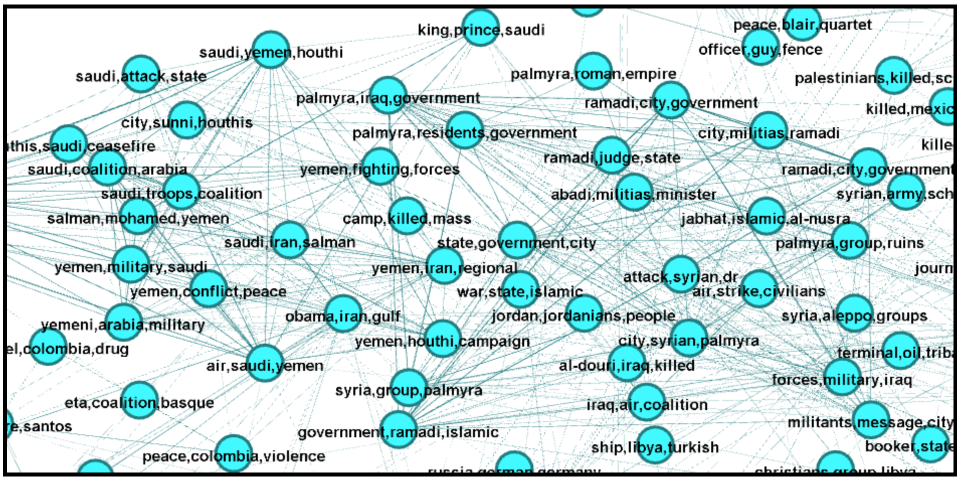


Figure 24: News-centric TF-IDF Unrest in the Arab World

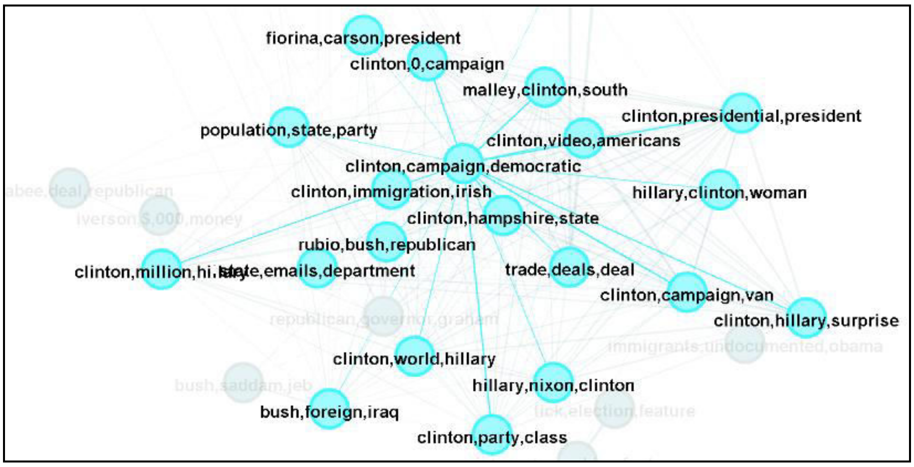


Figure 25: News-centric TF-IDF Hillary Clinton

News-centric TF-IDF Cluster 10

Cluster 10 has grouped content together relating to the banking inquiry, the housing crisis and the earthquake in Nepal. The Louvain method has only detected one cluster but the OpenOrd layout algorithm has presented three distinct clusters within this community. The common theme uniting these three topics is that of housing. The banking inquiry has a deluge of terms relating to mortgages, homes, house prices and property in general as illustrated in Figure 26. The housing crisis refers to the increase of families losing their homes in Ireland. This is driven by two factors, home repossessions and increasing rents in the private rental market.

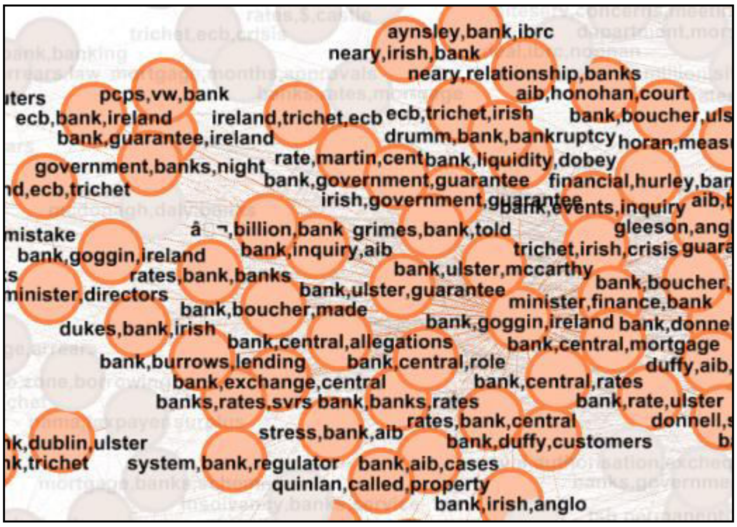


Figure 26: News-centric TF-IDF Banking inquiry

The demand for housing is aggravated by the difficulty in securing mortgages from banks and therefore relationships exist between this content. Finally, one of the major fallouts from the earthquake in Nepal is that of homelessness. Whilst these topics are not related to each other directly the document clustering method has revealed a pattern and drawn parallels between content that may not always be obvious when reviewing the content manually. In some implementations, identifying these types of patters may be useful, especially in an investigative journalism context.

News-centric TF-IDF Cluster 11

Cluster 11 is politics related and contains a large amount of content emphasising the fact that one of the major roles in journalism is that of monitoring the activities of governments. The Louvain method has only detected one cluster but the OpenOrd layout algorithm has presented four distinct clusters within this community relating to Irish politics. U.K. elections, the Aer Lingus sale and Northern Irish politics.

It is interesting to note that the Aer Lingus sale, as illustrated in Figure 27 is grouped within a political cluster. All the other content in this cluster directly relates to political parties and political representatives. The inclusion of Aer Lingus within this content may be partly because the sale relates to the Irish state share in Aer Lingus but some might argue that the operation of Aer Lingus has always been a highly political issue.

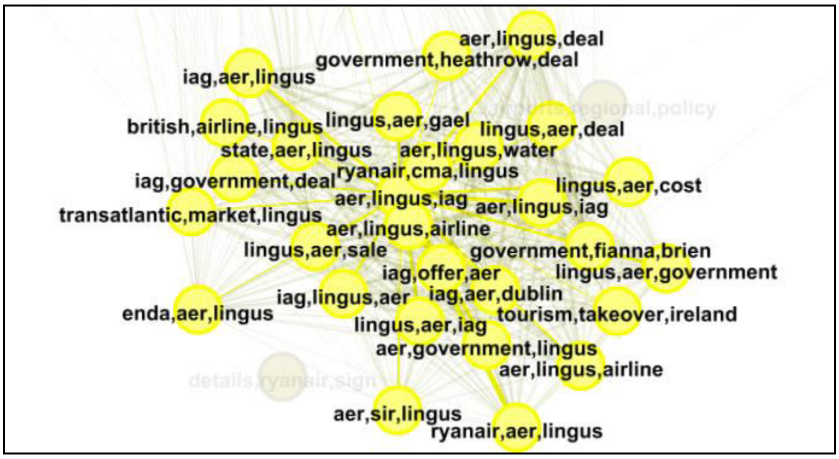


Figure 27: News-centric TF-IDF Aer Lingus sale to I.A.G.

News-centric TF-IDF Cluster 12

Cluster 12 has a very coherent theme aggregating all of the content related to the gay marriage referendum together, as presented in Figure 28. It is interesting to observe that one of the key terms emerging from this cluster is the term "child". This provides the insight, without reading all of the documents in the collection that the issue of gay parenting was a significant component in the public debate. It is also interesting to note the prevalence of the term "sir" in this cluster. These articles relate to letters, written by the general public to the Editor of the Irish Times, Kevin O'Sullivan. A significant number of such articles provides the insight that the gay marriage referendum ignited a significant amount of public debate.

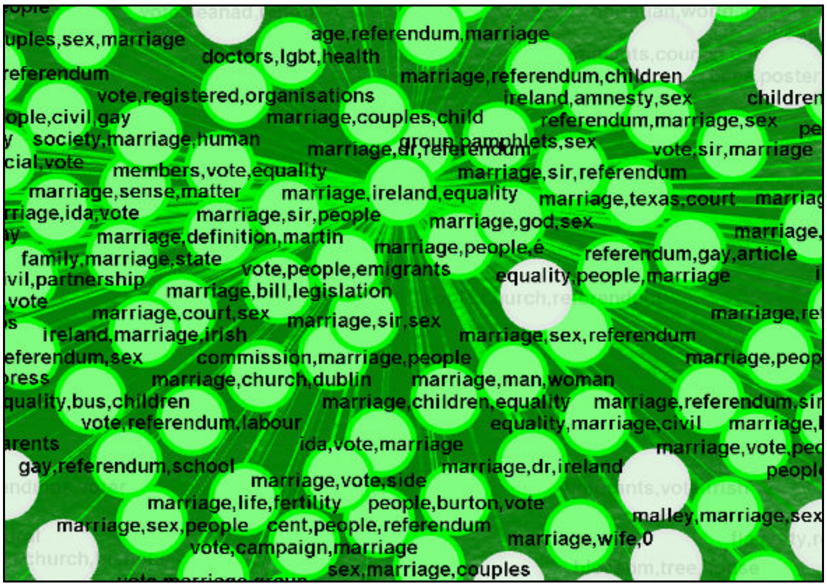


Figure 28: News-centric TF-IDF Gay marriage referendum

News-centric TF-IDF Cluster 13

As the clusters get larger, they also become more general. Cluster 13 groups a large amount of sports content together as illustrated in Figure 29..

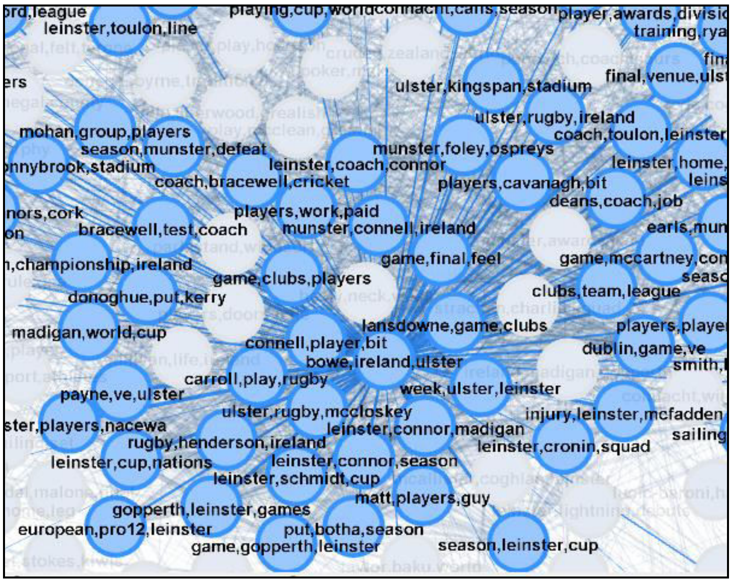


Figure 29: News-centric TF-IDF Rugby content within the Sports cluster

The sports content is not specific to one particular sport, but rather a large cluster with multiple sports content held within. The layout algorithm has not clearly separated out each sport presumably because of terms such as match, player, game and competition are common to a large number of sports. Whilst a lot of this content is not a topic in the true sense, it is at least grouped together coherently on a common sports theme. The only pollution in this cluster relates to the inclusion of some music competition articles. Presumably, the term competition is very prevalent in the sports cluster and has drawn in music articles with overlapping terms relating to competitions. This cluster also shows that sports content is a stable in the Irish Times with a significant breadth of coverage

News-centric TF-IDF Cluster 14

Cluster 14 is the largest cluster in the collection. The coherent theme of this cluster is that of crime and law. The cluster is littered with crime and law related terms such as prison, Gardaí, body, murder, Judge, trial etc. Similar to the sports cluster, the more general the content, the less delineated the individual topics are within the cluster. Some of the notable topics within this cluster include the very tragic murder of Irish nurse Karen Buckley in Glasgow as illustrated in Figure 30. Other notable cases include references to the Graham Dwyer case, the Gail O'Rorke assisted suicide case and the Mark Nash trial.

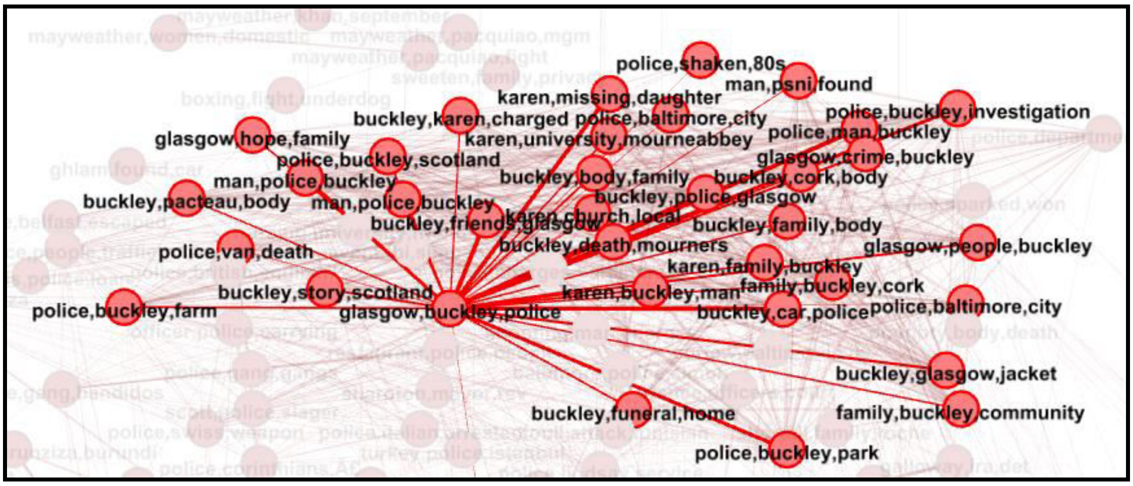


Figure 30: News-centric TF-IDF Karen Buckley

4.8.2 Visual Results of Traditional TF-IDF Approach

Figure 31 presents a data visualisation of the document collection with the traditional TF-IDF, Louvain method and OpenOrd layout algorithm applied. The first observation is that the traditional TF-IDF approach has a smaller number of nodes and edges as presented in Table 9. This may be surprising as the document sample used for each approach is identical. The differences can be attributed to the application of a minimum cosine similarity value as a pre-requisite requirement for generating the GEXF graph file. The news-centric TF-IDF approach found more articles similar to each other presumably because the additional term weights applied to the headline and summary fields increased the cosine similarity value when comparing the document to other documents in the collection where the same increase has been applied. The application of an additional cosine similarity weight based on the time proximity of the articles is also likely to be a factor. The differences that account for these changes to the cosine similarity score warrant further evaluation.

Term Weighting Approach	Number of Nodes	Number of Edges
Traditional TF-IDF	3,592	66,677
News-centric TF-IDF	3,708	91,244

Table 9: Article combinations per number of input articles

The second notable difference is that the traditional TF-IDF approach identifies significantly more clusters than the news-centric TF-IDF approach. A large number of these communities are extremely small with two or three linked articles. Considering the size of the document collection, these articles are not likely reveal any major trends or topics. Similar to the news-centric approach, clusters with less four or less linked documents are filtered out of the graph.

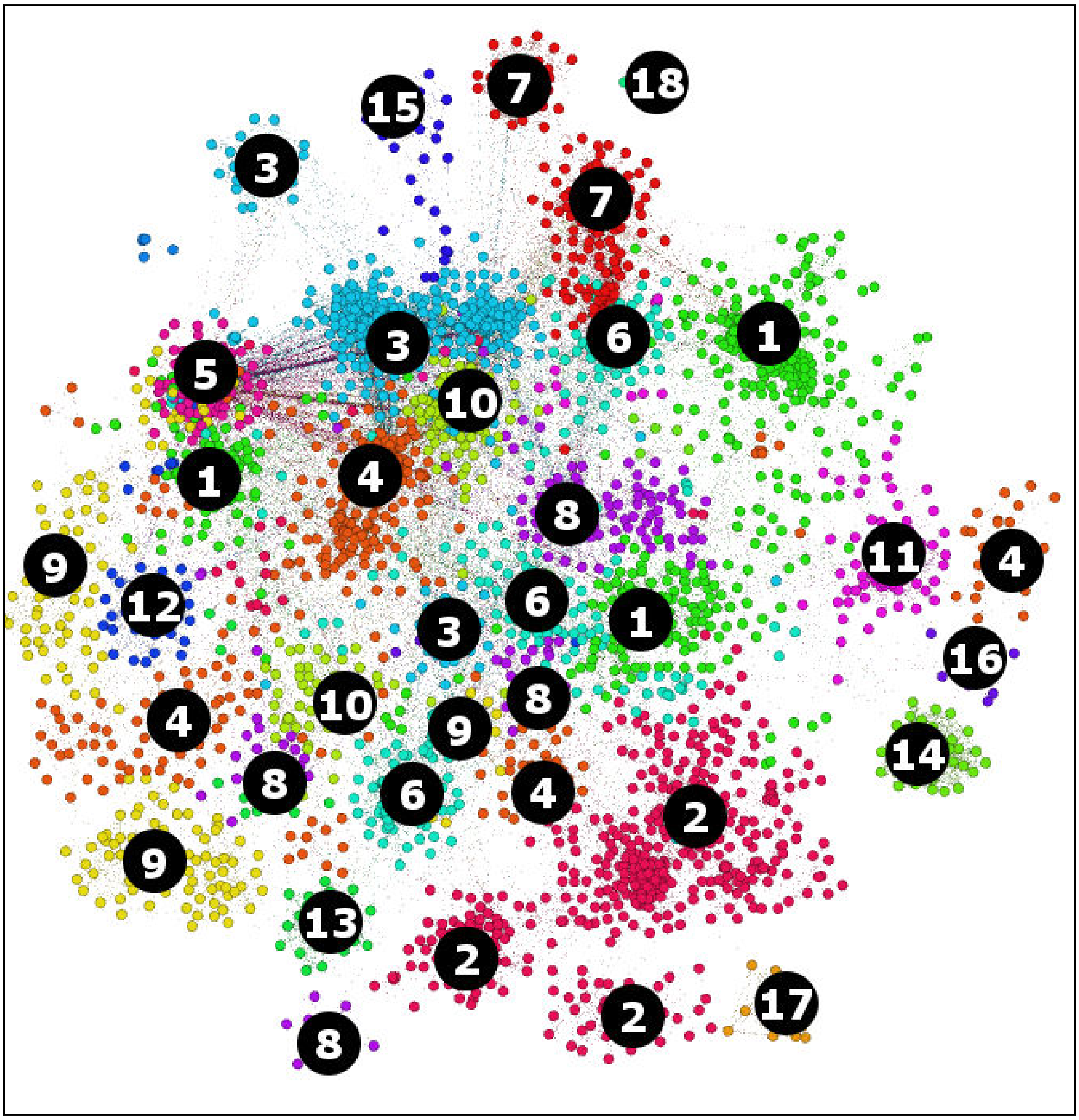


Figure 31: Traditional TF-IDF document collection annotated

A total of 29 communities were filtered from the graph and the remaining 19 clusters were examined. It is interesting to note that the traditional TF-IDF approach generated a larger number of smaller and insignificant clusters the reasons for which are not entirely clear at this point. Figure 31 illustrates the data visualisation of the document collection using the traditional TF-IDF approach, clustered with the Lovain method of community detection and presented using the OpenOrd layout algorithm. Clusters with fewer than four linked documents have been removed.

On first examination of the remaining 19 clusters, it becomes apparent that the OpenOrd layout algorithm has presented a large number clusters within larger cluster communities detected by the Louvain method. The same phenomenon was evident with the news-centric TF-IDF approach but the extent is much more pronounced with the traditional TF-IDF approach. It should be noted that as a different number of clusters are under evaluation for each TF-IDF approach, the cluster numbers are not directly comparable.

Traditional TF-IDF Cluster 13, 14 15, 16 & 17

These clusters present a very mixed bag of content and it is difficult to see what terms have resulted in them being clustered together. Both cluster 17 contain an article with the top three terms, “people, civil, gay” and “gay, homosexual and marriage” respectively as illustrated in Figure 32. These articles on the face of it would appear to be related to the Gay Marriage Referendum. Cluster 17 contains two documents containing golf and sports related terms and one document contains the top terms “English, constitutional and Cameron”. The final document in this cluster contains the top terms “pluralism, disagreement and values”. On reading the article in question, the content of the article relates to the feared destruction of the ancient city of Palmyra in Syria. It is difficult to find anything that relates these documents but clearly there are overlapping terms otherwise they would not have been deemed related.

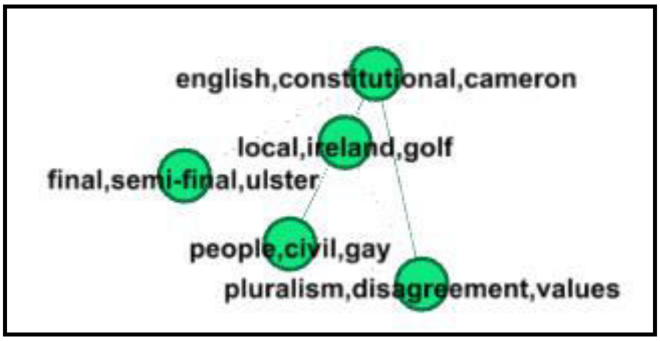


Figure 32: Traditional TF-IDF Cluster 17

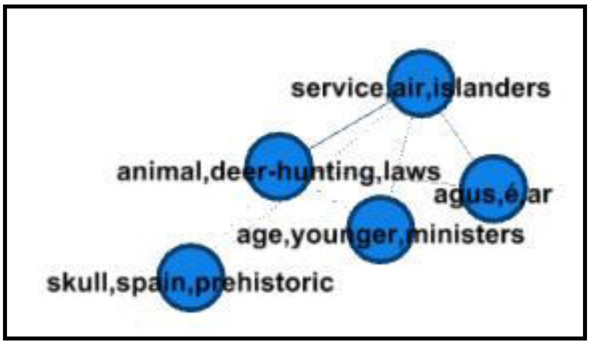


Figure 33: Traditional TF-IDF Cluster 16

Cluster 16 also contains another very mixed bag of seemingly unrelated content. Interesting it contains an Irish language story with the top three terms "agus, é and ar" as illustrated in Figure 33. In the news-centric TF-IDF approach all Irish language articles were clustered together and appeared on the on the peripheries of the graph as the similarity with other documents in the collection was small to non-existent.

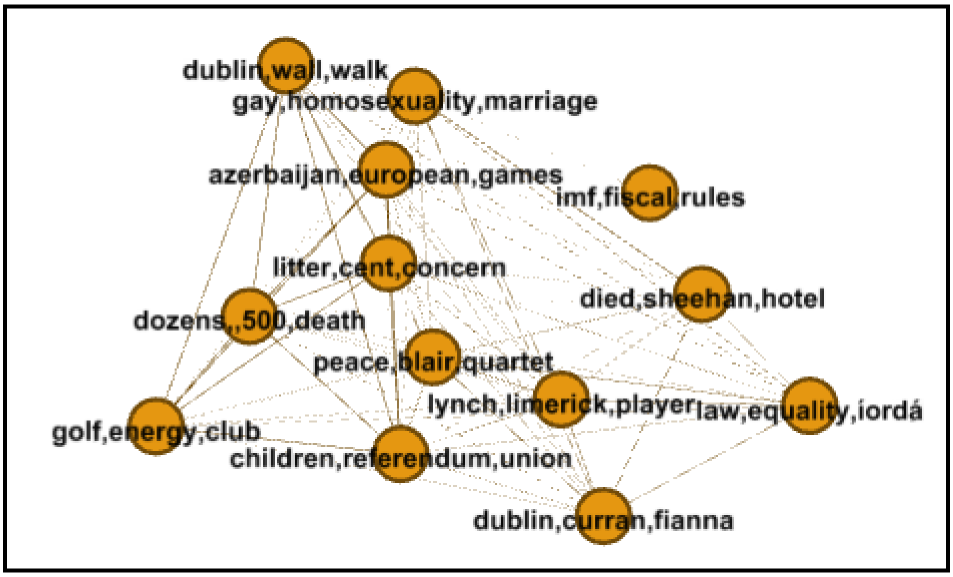


Figure 34: Traditional TF-IDF Cluster 15

Cluster 14 contains content related to gay marriage, the U.K. elections, public pay talks, politics and the banking enquiry as presented in Figure 35. Unfortunately this mixed content trend continues throughout the entire document collection with the traditional TF-IDF approach making any comparisons with the news-centric approach extremely difficult.

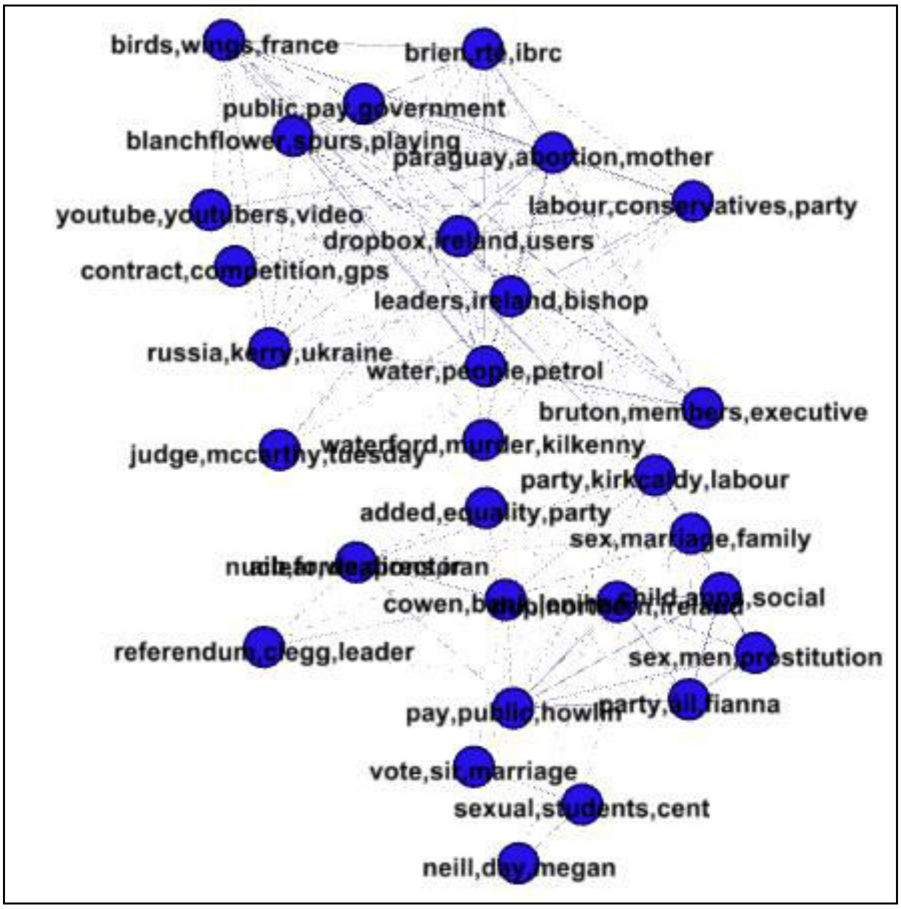


Figure 35: Traditional TF-IDF Cluster 14

As the randomness of the clustering with the traditional TF-IDF approach continued throughout all of the remaining clusters and as such, no further graphical examples will be displayed. Rather, the reasons for this will be examined and discussed in the evaluation chapter.

4.9 Conclusion

This chapter described the implementation of this experiment and presented and analysed the initial findings by visually assessing the data visualisations produced by each TF-IDF approach. The initial findings suggest that the news-centric TF-IDF approach is promising and may have some merit. The main topics identified by the expert users are all present in the significantly larger document sample automatically processed using the news-centric TF-IDF approach. In general, the clusters are all distinctive and the content is grouped coherently together although in all clusters, unrelated content is included to some extent or another. In most cases, it is easy to identify the overlapping terms that result in unrelated topics being pulled together on the graph. With the news-centric TF-IDF approach, the clusters that have performed the best relate to topics where terms do overlap significantly with other topics. More general topics such as health and sport are less distinctive albeit that the content is at least coherently grouped. It should be noted that a significant overlap exists in the terms used in sports which the visualisations clearly show have an impact.

The initial findings from visualising the traditional TF-IDF approach are poor. Topics are consistently spread throughout the clusters for no obvious or discernible reason. It suggests that applying weights to terms based on their term frequency alone is insufficient to distinguish the major themes and key areas of interest in the text. The dimensionality of the document collection is vast, and the most significant terms to describe the story are possibly becoming lost due to the extensive vocabulary present and the same treatment and priority being applied to every word in the collection. Further assessment of the traditional TF-IDF approach will be undertaken in the following, evaluation and analysis chapter.

5 EVALUATION

5.1 Introduction

The focus of this chapter will be on the evaluation of two TF-IDF approaches. The objective of this evaluation is to establish if the newly proposed new-centric TF-IDF approach improves document clustering accuracy and topic aggregation capabilities for news articles when compared to traditional TF-IDF term weighting approach. As presented in the implementation chapter, the expert users were provided with a list of articles published on pages one to five in the Irish Times between the dates of the 18th of May 2015 and the 23rd of May 2015. This smaller document sample represents a week newspaper content. The same sample was processed using the two TF-IDF approaches and the results are compared.

As highlighted in the literary review, evaluating the results of any document clustering solution is difficult because the results may be judged differently depending on the measure that is used and different representations cannot be compared with each other. Steinbach (2000) advocated using several measures when evaluating clustering results. As such, the toolset will be evaluated using three approaches and as follows:

- Data Visualisation Evaluation
- Expert-User External Evaluation
- Cosine Similarity Evaluation

5.2 Visual Evaluation

Data visualisation diagrams were produced for each TF-IDF approach and the Louvain method and OpenOrd layout algorithm were applied. Figure 36 presents the data visualisation diagram for the news-centric TF-IDF approach and Figure 37 presents the data visualisation diagram for the traditional TF-IDF approach . On the face of it, both data visualisation diagrams show distinct clusters presentations that are colour coded by their cluster community.

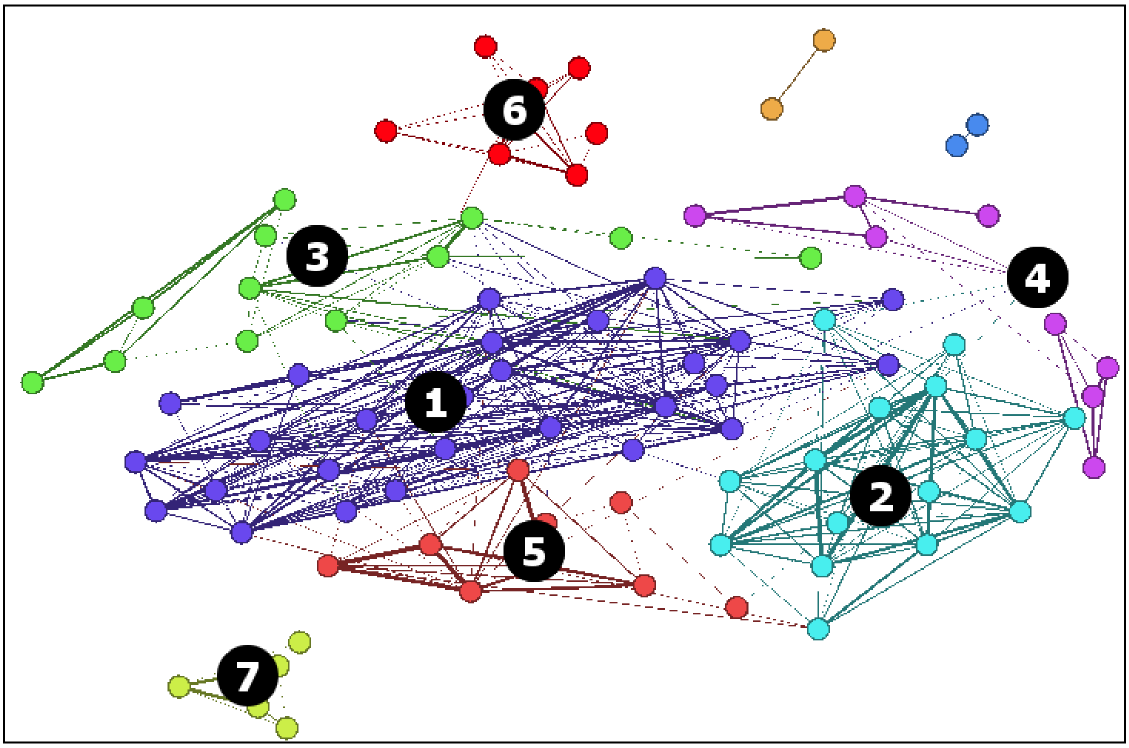


Figure 36: News-centric TF-IDF with Louvain method and OpenOrd layout

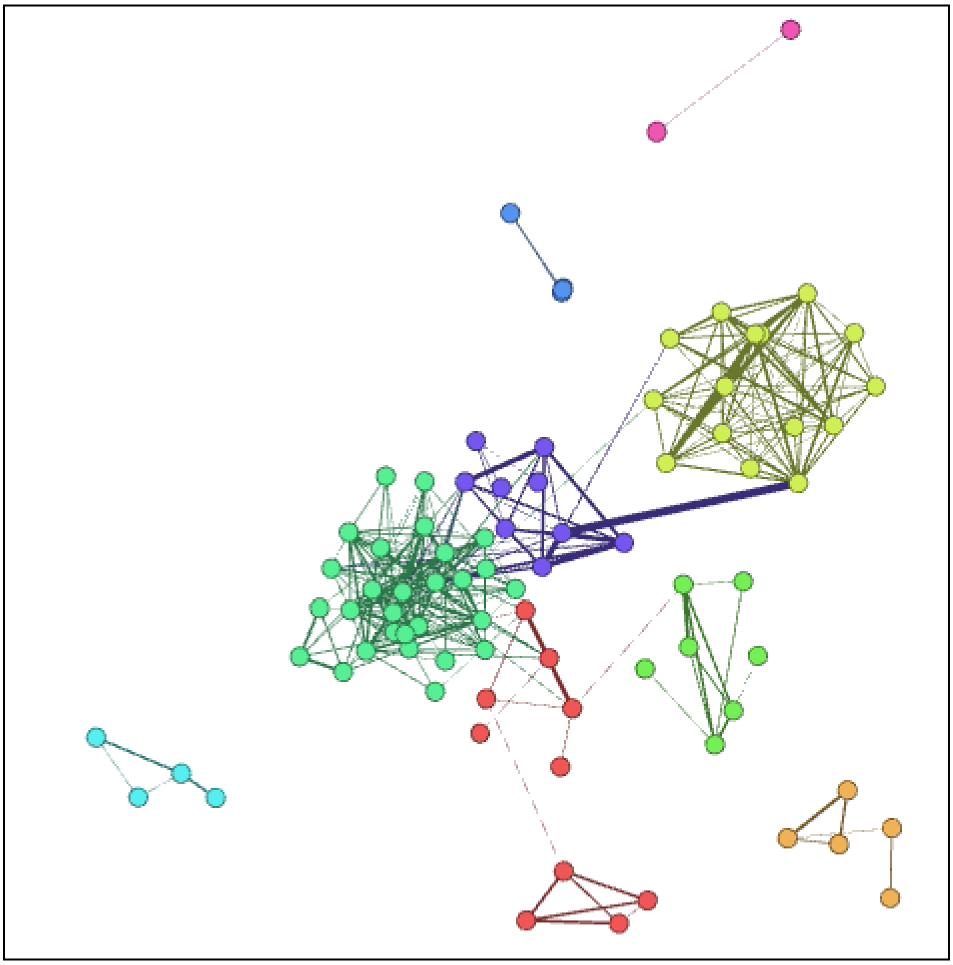


Figure 37: Traditional TF-IDF with Louvain method and OpenOrd layout

Cluster 1 Gay Marriage Referendum

The largest cluster evident when the news-centric TF-IDF approach was applied groups 27 articles together relating to the gay marriage referendum as illustrated in Figure 38. The news-centric TF-IDF approach produced an excellent clustering result for this topic with the smaller document sample. This may be aided by the fact that some of the key terms used in this topic are unique to this topic. The terms “gay, homosexual, marriage and referendum” are not used in any other topic in this document sample. That said, more generic terms such as children, parents, church and religion featured heavily in this debate and these terms will have significant relationships with other documents in the collection.

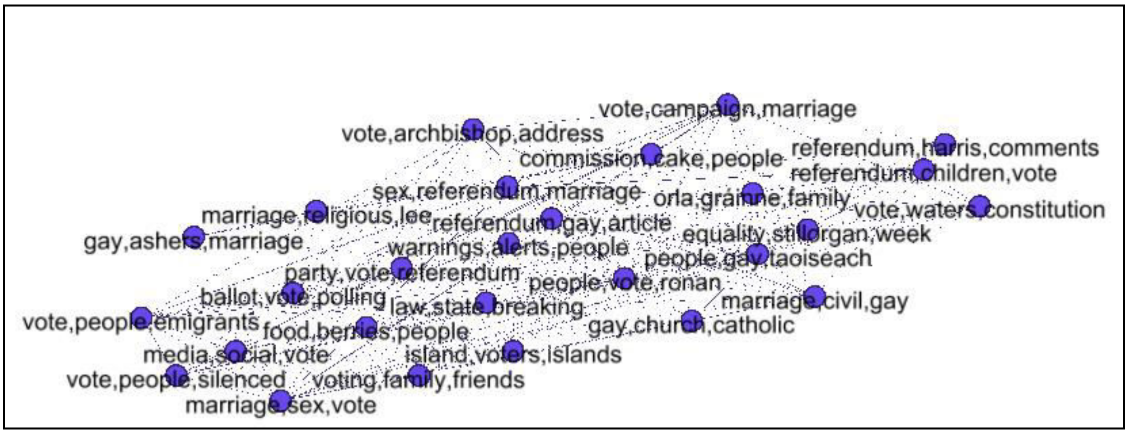


Figure 38: News-centric TF-IDF Gay Marriage Referendum

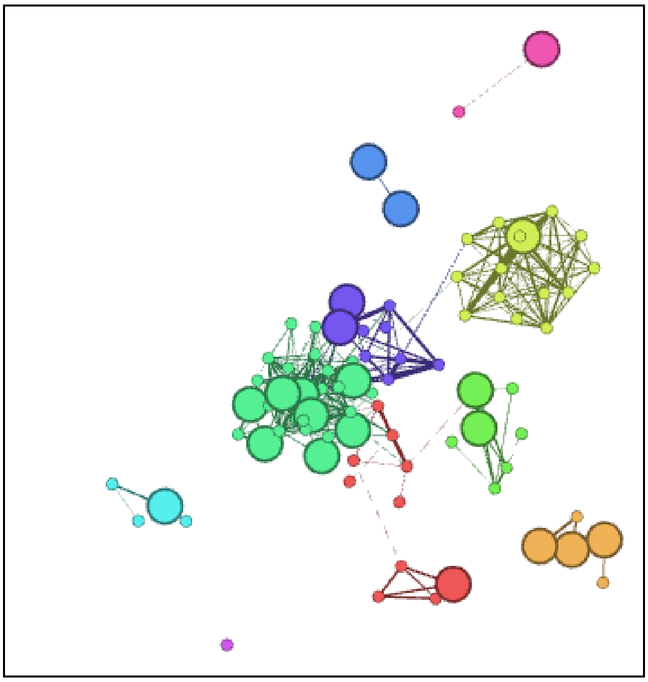


Figure 39: Traditional TF-IDF Gay Marriage articles distributed

The traditional TF-IDF approach was unable to cluster the exact same documents in one distinct cluster. Figure 39 shows how the articles that were grouped in Figure 38 using the news-centric TF-IDF approach are distributed across the graph when the traditional TF-IDF approach was applied. The node sizes of the articles in question have been increased to illustrate the distribution across the graph. It should be noted that only 21 of the total 27 articles clustered by the news-centric TF-IDF approach featured on the traditional TF-IDF graph. A total of 6 articles did not meet the minimum cosine similarity score to be included. On this basis, it is fair to say that increasing the importance of the headline and summary term weights and the application of temporal cosine similarity weighting has had a positive effect in for this cluster.

Cluster 2 Price Charles visits Ireland

The second cluster identified with the news-centric TF-IDF approach relates to Prince Charles' visit to Ireland in May 2015 as illustrated in Figure 40. The news-centric TF-IDF approach produced an excellent clustering result for this topic. This may be aided by the fact that some of the key terms used in this topic are unique to this topic.

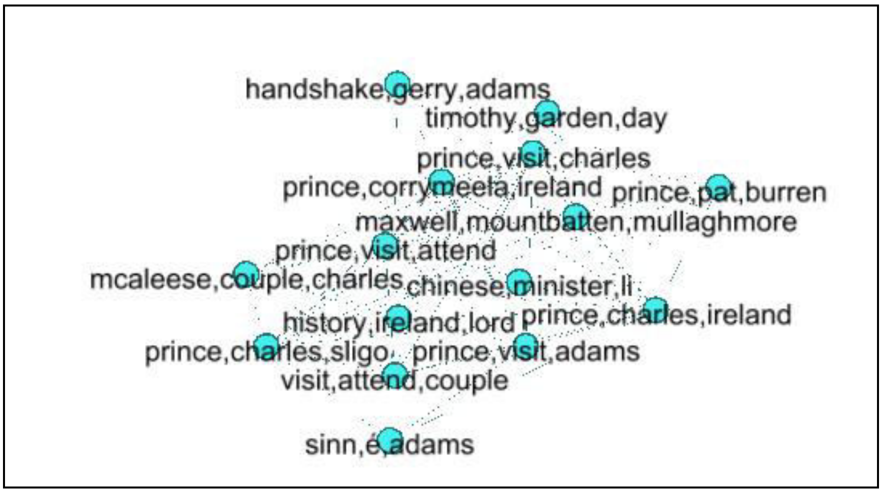


Figure 40: News-centric TF-IDF Prince Charles visits Ireland

Similarly, the traditional TF-IDF approach did not appear to cluster this content particularly well. Figure 41 shows the distribution of the documents featured in Figure 40 distributed across the traditional TF-IDF graph. The size of the nodes have been increased to highlight the articles in question. It should be noted that one article featured using the news-centric TF-IDF approach did not meet the minimum cosine similarity score to be included in the traditional TF-IDF GEXF graph.

Comparing the data visualisation diagrams that were produced against the traditional TF-IDF approach is an extremely difficult task. The topic articles, as determined by the expert user group, are distributed so dispersedly throughout the graph there is no basis for comparison or for qualifying the results. The randomness of the clustering with the traditional TF-IDF approach continued throughout all of the remaining clusters and as such, no further graphical examples will be displayed.

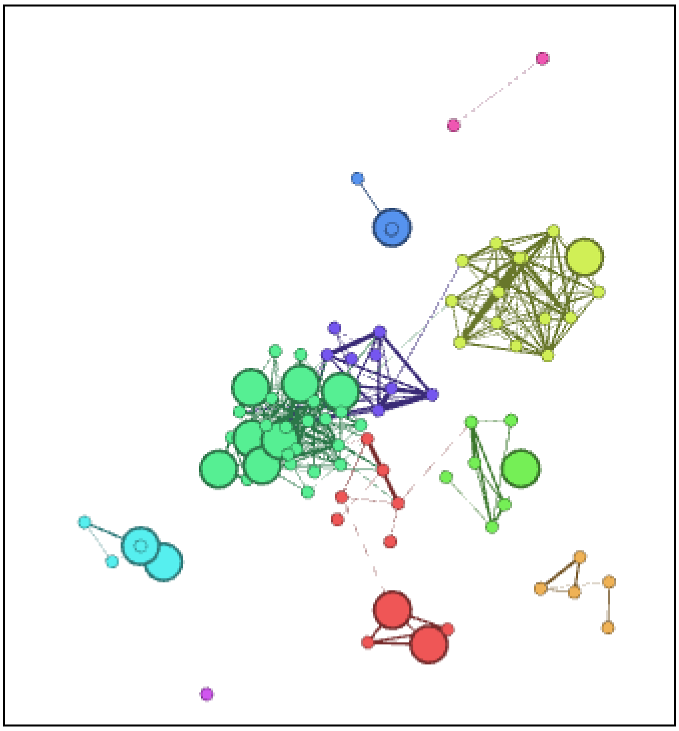


Figure 41: Traditional TF-IDF Prince Charles visits Ireland distributed

Cluster 3 Teachers Pay, Education & Children in Provision

As illustrated in Figure 42, cluster 3 produced a group of articles relating to teachers, pay discussions and education as well as articles relating to children, child poverty and children in direct provision. The term "children" features heavily in all of these articles. Interestingly, the term children features significantly in the gay marriage referendum cluster but the clustering algorithm has detected that this is a different topic. The content has a coherency but from a topic aggregation point of view, the clustering and delineation of topics could be better.

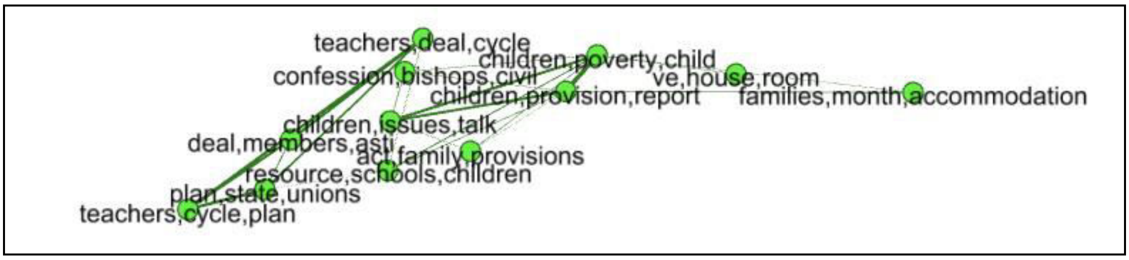


Figure 42: News-centric TF-IDF Teachers pay, children in provision

Cluster 4 Deaths of Anna Hicks and Julia Holmes

Cluster 4 groups two topics together as illustrated in Figure 43. One topic relates to the discovery of two bodies in a farm in Limerick by burglars where it is thought the bodies had lain undiscovered for some time. The other topic relates to the death of a young Dublin teenager from a presumed use of ecstasy. The main terms linking these two unrelated topics together are bodies, discovered and death. This has a coherency to this cluster albeit that the topics are not directly related to each other.

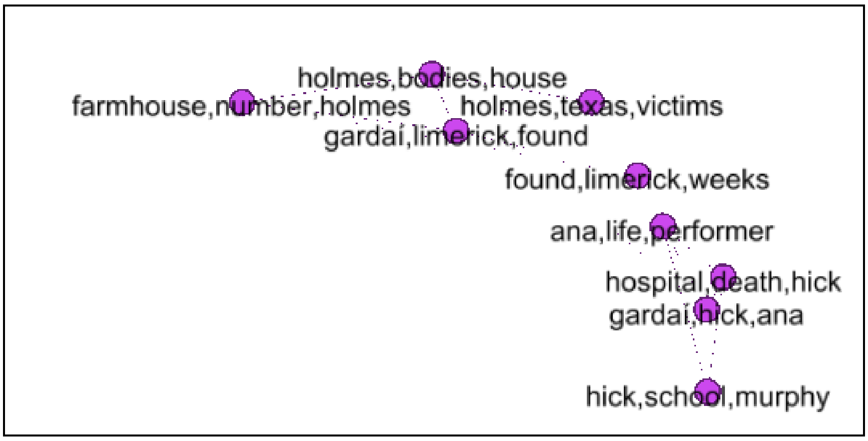


Figure 43: News-centric TF-IDF Deaths of Anna Hick and Julia Holmes

Cluster 5 Health

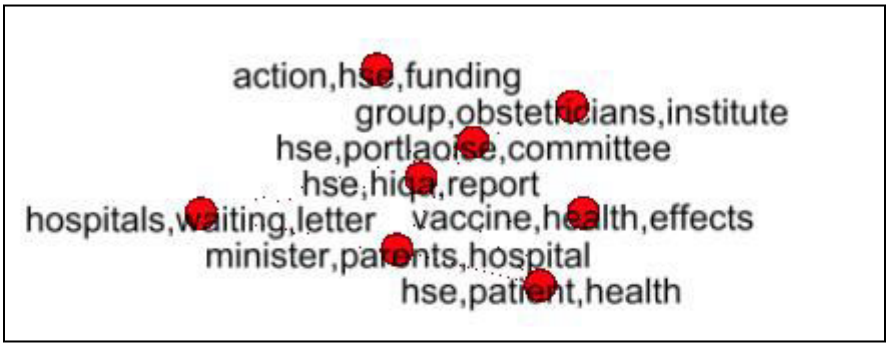


Figure 44: News-centric TF-IDF Health

Cluster 5 groups health related content together as illustrated in Figure 44. No specific topic emerges but the content is coherently grouped.

Cluster 6 General Election

Cluster 6 groups content together related to the next general election as illustrated in Figure 45. An article referring to migrants voting in the election has formed a relationship with an article related to refugees that is not related to this topic. Another article, related to the Eurovision and RTÉ is included here and has no obvious relationship with the topics.

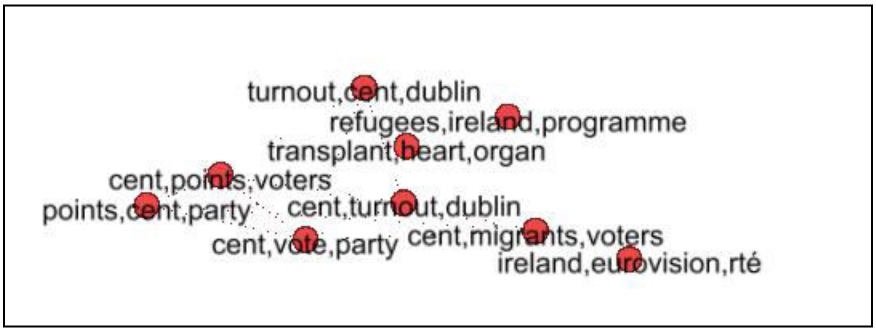


Figure 45: News-centric TF-IDF General Election

Cluster 7 Banking Inquiry

Cluster 7 is relatively small and groups content related to the Irish banking inquiry together. This cluster has performed well. An article related to the sale of SiteServ is included here but this includes write downs made by the former Anglo Irish bank which has a direct relevance to the other content in the cluster.

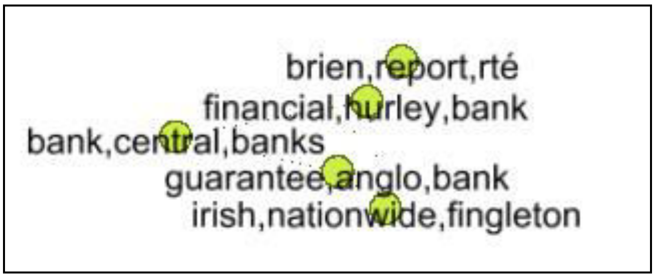


Figure 46: News-centric TF-IDF Banking Inquiry

5.3 Expert User Evaluation

To compare the quality of the news-centric TF-IDF approach to the topic categorisation completed manually by the expert users, the verbal qualifiers very poor, poor, fair, good and excellent were used and applied to a subset of Mahalakshmi's (2014) properties for efficient document clustering as follows:

1. Relevance - The method grouped relevant documents together separately from irrelevant documents.
2. Browse able-Summaries - The user, at a glance, could establish if the cluster's content was of interest.
3. Speed – The speed of the clustering method was of a magnitude of what a user could complete manually.

Cluster 1 Gay Marriage Referendum

One of the articles categorised by expert as "marriage ref" related to a party political opinion poll and did not meet the required minimum cosine similarity score.
Interestingly, the other experts did not include this article and in reviewing it further the content of this article is not in any way related to the gay marriage referendum.
This highlights that experts are not fallible and are prone to making mistakes. Table 10 presents the gay marriage referendum cluster as evaluated by expert users.

	Expert 1	**Expert 2**	**Expert 3**
Relevancy	Excellent	Excellent	Excellent
Browse ability	Excellent	Excellent	Excellent
Speed	Excellent	Excellent	Excellent

Table 10: Gay marriage referendum cluster evaluated by expert users

Cluster 2 Prince Charles visits Ireland

The cluster that grouped articles relating to Prince Charles's visit together was one of the best performing clusters in the document collection. The news-centric TF-IDF approach identified all of the articles manually categorised by all of the experts. One expert did not categorise an article with the headline "Highlights Prince Charles and Camilla's visit" which is surprising but also highlights, again, expert users are not always 100% accurate either. Although not obviously visible in the graph, the Prince

Charles cluster included some content more appropriate for the gay marriage referendum. Table 11 presents the Prince Charles visits Ireland cluster as evaluated by expert users.

	Expert 1	Expert 2	Expert 3
Relevancy	Excellent	Good	Excellent
Browse ability	Excellent	Excellent	Excellent
Speed	Excellent	Excellent	Excellent

Table 11: Prince Charles visit evaluated by expert users

Cluster 3 Teachers Pay, Education & Children in Provision

The prevalence of the term children grouped content together that would not have been related otherwise. Expert 1 nominated the topic "Junior Cycle" for three of the education related stories. Other topics nominated by this user for this content include "Asylum seekers" and "Child poverty". The automated clustering method has picked up on these themes but could have performed better in identifying the different topics within this content. Table 12 presents the teachers pay & children in provision cluster as evaluated by expert users.

	Expert 1	Expert 2	Expert 3
Relevancy	Good	Fair	Good
Browse ability	Good	Fair	Fair
Speed	Excellent	Excellent	Excellent

Table 12: Teachers pay & children in provision evaluated by expert users

Cluster 4 Health

One expert user categorised this content very granularly suggesting topics such as "waiting lists", "maternity services" and "patient safety". It should be noted that a minimum number of five articles were not present within the document sample for these suggested topics. The other experts categorised at a more general "health" level.. Table 12 presents the health cluster as evaluated by expert users.

	Expert 1	Expert 2	Expert 3
Relevancy	Excellent	Excellent	Excellent
Browse ability	Excellent	Excellent	Excellent
Speed	Excellent	Excellent	Excellent

Table 13: Health cluster evaluated by expert users

On the basis that all of this content is health related the expert users scored the results from the automated method highly.

Cluster 6 General Election

The first expert user nominated 6/7 of these stories as either politics or migrant and refugee related so there is some merit in the clustering results albeit they should be at a more granular level. Table 14 presents the general election cluster as evaluated by expert users.

	Expert 1	**Expert 2**	**Expert 3**
Relevancy	Good	Good	Good
Browse ability	Good	Fair	Fair
Speed	Excellent	Excellent	Excellent

Table 14: General election cluster evaluated by expert users

Cluster 7 Banking Inquiry

The banking inquiry cluster performed exceptionally well. All experts suggested the same articles for this topic and the automated news-centric TF-IDF method has clustered all of these articles together in one cluster. Table 15 presents the banking inquiry cluster as evaluated by expert users.

	Expert 1	**Expert 2**	**Expert 3**
Relevancy	Excellent	Excellent	Excellent
Browse ability	Excellent	Excellent	Excellent
Speed	Excellent	Excellent	Excellent

Table 15: Banking inquiry evaluated by expert users

5.4 Cosine Similarity Evaluation

5.4.1 Cosine Similarity Minimum Value

A cosine similarity minimum value of 0.2 was included when generating the GEXF files for each TF-IDF approach. As the document sample is significantly smaller than the sample used in the visualisation evaluation, it might be argued that no cosine similarity minimum should be applied. This was tested and it found that the Louvain method performed very poorly when edges were included on the graph that linked articles below the minimum threshold value. A maximum of three clusters were found using for the news-centric TF-IDF approach when the minimum cosine similarity limit was excluded. The Louvain method uses the edge weight to detect the clusters but appears to perform very badly if any edges are included between documents that are only minimally related to each other.

With the traditional TF-IDF approach, eighteen articles did not meet the cosine similarity minimum measure using this smaller document sample. In contrast, a total of twelve articles did not meet the cosine similarity minimum measure when the news-centric TF-IDF approach was applied. This infers that by increasing the weights of the terms that feature in the headline and summary fields and by increasing the cosine similarity score when articles are both similar and published in close proximity with each other, the process of identifying similar documents in the collection is marginally improved.

5.4.2 Cosine Similarity Comparison

The following section provides a brief comparison of the cosine similarity scores allocated by each TF-IDF approach for the articles categorised by the expert users as "Prince Charles visits Ireland". The Prince Charles topic was selected as there was a good correlation between the articles categorised by the expert user group and by the automated news-centric TF-IDF method. Table 16 summaries the articles in this cluster.

Doc ID	Headline
15	Royal couple to begin visit at NUIG
18	Highlights Prince Charles and Camilla's visit
19	SF members to join events with prince
26	Adams and McGuinness to meet Prince Charles during visit to Galway
35	Adams says regrets expressed on both sides
42	Arrival in Mullaghmore will bring bittersweet emotions
43	More sticks than you could shake a prince at
44	Adams handshake motivated by reconciliation and politics
45	Prince praises 'magic' unique to Ireland
61	Royal couple retrace steps to lay history to rest
96	When Gerry met Charles – the moving and shaking
101	Prince calls on North to free itself from past

Table 16: Articles manually categorised as Prince Charles visits Ireland

Table 17 shows each document pair within this topic and the calculated cosine similarity score using each TF-IDF approach. These results quantitatively present the impact of the news-centric TF-IDF weights and cosine similarity weights on the articles. The cosine similarity scores calculated using the traditional TF-IDF offer an explanation why the document clustering and data visualisations performed so poorly. A large number of the article comparisons do not meet the minimum cosine similarity score. Whilst the nodes may have been included on the graph because of the relationship with other articles in the collection, edges were not drawn between the articles within the Prince Charles topic. The cosine similarity edge weight is integral to the operation of both the Louvain clustering method and the OpenOrd layout algorithm.

Doc ID 1	Doc ID 2	Cosine Similarity Traditional TF-IDF	Cosine Similarity News-Centric TF-IDF
15	18	0.10391358	0.842081
15	19	0.182264564	0.381198
15	26	0.11293883	0.822371
15	35	0.016665365	0.322542
15	42	0.023736369	0.253562
15	43	0.013699916	0.294228
15	44	0.035327061	0.155119
15	45	0.020590128	0.60158
15	61	0.044975692	0.242178
15	96	0.018796129	0.1636
15	101	0.003629196	0.31315
18	19	0.234414657	0.427687
18	26	0.196957298	0.778451
18	35	0.060994505	0.288942
18	42	0.033696609	0.283827
18	43	0.109398744	0.394956
18	44	0.069218801	0.176438
18	45	0.003044807	0.533553
18	61	0.050990012	0.204223
18	96	0.036133702	0.206853
18	101	0.01932029	0.45887
19	26	0.210347327	0.689882
19	35	0.071917818	0.516746
19	42	0.029077491	0.19849
19	43	0.070181993	0.316169
19	44	0.179498308	0.573017
19	45	0.015286841	0.50655

19	61	0.062609154	0.130062
19	96	0.033142918	0.314288
19	101	0.02627083	0.399926
26	35	0.054031168	0.560036
26	42	0.030782494	0.359711
26	43	0.071066905	0.435283
26	44	0.057267344	0.381747
26	45	0.004450382	0.713375
26	61	0.071288183	0.265266
26	96	0.020313123	0.32166
26	101	0.009413042	0.447987
35	42	0.083975679	0.317137
35	43	0.117844318	0.275776
35	44	0.079645261	0.576028
35	45	0.011902738	0.53266
35	61	0.058932292	0.267289
35	96	0.123868562	0.403148
35	101	0.105737447	0.324004
42	43	0.043314808	0.328266
42	44	0.062827321	0.158563
42	45	0.01898735	0.289077
42	61	0.033574749	0.418262
42	96	0.037142142	0.166745
42	101	0.060240458	0.299845
43	44	0.065808875	0.186466
43	45	0.053229065	0.429007
43	61	0.05471529	0.200152
43	96	0.054307885	0.271415
43	101	0.029801978	0.359687
44	45	0.080740146	0.384973
44	61	0.041641346	0.158004
44	96	0.06142111	0.35585
44	101	0.037356851	0.213381
45	61	0.025698033	0.279578
45	96	0.042958632	0.333641
45	101	0.066356347	0.401462
61	96	0.04431142	0.140895
61	101	0.018118043	0.256393
96	101	0.431595308	0.216483

Table 17: Cosine similarity comparison for Prince Charles visits Ireland

5.5 Conclusion

Comparing the data visualisation diagrams that were produced by each TF-IDF approach was an extremely difficult task. The related articles, as determined by the expert user group, were distributed so dispersedly throughout the graph there was no basis for comparison or for qualifying the results. In examining the cosine similarity scores that were calculated using the traditional TF-IDF it became more evident why the document clustering and data visualisations performed so poorly. The traditional TF-IDF approach does not appear to have sufficient distinguishing power between the general terms in the article and the key terms that could identify the key areas of interest discussed in that article. This mirrors what Strehl (2007) argument that only a small number of words and terms in a document have a disguisable power.

Whilst the expert user analysis provided a grounding with which the proposed news-centric TF-IDF approach could be compared, it should be acknowledged that not all expert users were in agreement at all times. They differed significantly in the granularity with which they suggested the topics. In some cases the expert user categorisation was inaccurate. This highlights that human categorisation of content is not infallible. Experts can and do make mistakes and are subjective in their views.

The experimentation shows that that the developed news-centric approach is promising when compared to the manual document clustering effort undertaken by the three journalist expert users. A number of clusters performed very well and in some cases almost as well as the expert users. It should be noted however, that the best performing clusters appeared to contain terms that are somewhat unique to the topic with little or no overlap with other documents in the collection.

Clusters on more general topics such as health and sport performed fairly but not optimally and the limitations of natural language processing and document clustering become apparent here. Sport is littered with terms that are common across categories such as “team, match, player and competition”. Health contains a large number of synonyms. Even the term health itself can be substituted with a diverse range of terms such as fitness, wellness, well-being, healthiness, welfare and vitality to name but a few.

Where the automated method scored very highly was in the category of speed. The automated method completed the task in a fraction of the time it took the expert users to categorise the content. The expert users spent between one and two hours manually categorising the content. The automated method completed the task within three minutes using the same document sample. The automated method has the potential to deliver operational efficiencies by crunching large document sets in a fraction of the time it takes a human resource to complete the same task.

Based on the evaluations, the findings indicate that applying weights to terms based on their term frequency alone is insufficient to distinguish the major themes and key areas of interest in the text. Conversely, the findings indicate that boosting the weights of the headline and summary terms, and increasing the similarity of already similar documents when they are published in close proximity together can improve topic aggregation and document clustering when applied to news articles.

In addressing the research question therefore, the newly proposed news-centric term weighting and cosine similarity scheme did improve document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach.

Whilst the experimentation shows that that the developed approach is promising when compared to the manual document clustering effort undertaken by the three journalist expert users, it also highlights the challenges of natural language processing and document clustering methods in general. The results suggest that a blended approach of complimenting automated methods with human-level expertise may yield the best overall results.

6 CONCLUSION

6.1 Introduction

Large document sets have become a significant part of the journalism profession. A review of the literature showed that there is a growing interest in data-driven journalism and specifically that the journalism profession needs better tools to understand and develop actionable knowledge from large document sets. Motivated by this problem and informed by best practice and a review of current literature, this study developed a novel and domain-specific document clustering and topic aggregation toolset for a specific news organisation.

6.2 Problem Definition & Research Overview

This study contributed a review of the main trends and challenges in the journalism professional as presented in the literature. It also provided a review of term frequency weighting functions put forward in research and established the merits and limitations of each approach. Informed by this research, a number of news-centric term weighting and document similarity modifications were proposed and developed into a toolset.

The basic research problem was as follows: Given a large collection of unstructured news documents, could the major topics and key areas of interest discussed in that collection be identified automatically and did the proposed term weighting and document similarity approach improve document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach.

6.3 Contributions to the Body of Knowledge

The study contributed to the Body of Knowledge by meeting the following objectives as follows:

- Reviewed the main trends and challenges in the journalism professional as presented in the literature.

- Completed a review of term frequency weighting functions and document clustering literature to identify research trends and to establish best practice.
- Informed by best practice and a review of the current literature a number of news-centric term weighting and document similarity modifications were proposed and developed into a toolset.
- Developed the proposed toolset using series of free and open-source technologies.
- Qualitatively assessed the data visualisation diagrams produced for each TF-IDF approach following the application of quantitative document clustering and data visualisation methods.
- The effectiveness of the toolset was qualitatively assessed by comparing the manual document clustering efforts undertaken by the domain expert users to the automated approach.
- Evaluated the findings informed by the body of research in the area.

6.4 Experimentation, Evaluation and Limitation

Experimentation was conducted on a large unstructured collection of newspaper articles from the Irish Times. The document sample was processed using a traditional TF-IDF approach and again, using the newly proposed news-centric TF-IDF approach. The objective of the experimentation was to establish if the newly proposed news-centric term weighting and cosine similarity scheme improved document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach.

As suggested in the research, objectively evaluating the quality of document clusters and establishing if the clusters were meaningful was a significant challenge. The evaluation approaches that were implemented in the study were as follows:

- Data Visualisation Evaluation
- Expert-User External Evaluation
- Cosine Similarity Evaluation

Comparing the data visualisation diagrams that were produced by each TF-IDF approach was an extremely difficult task. The related articles, as determined by the expert user group, were distributed so dispersedly throughout the graph there was no basis for comparison or for qualifying the results. In evaluating the cosine similarity scores calculated using the traditional TF-IDF approach it became more evident why the document clustering and data visualisations performed so poorly. The traditional TF-IDF approach did not appear to have sufficient distinguishing power between the general terms in the article and the key terms that could identify the key areas of interest discussed in that article.

The experimentation showed that that the developed news-centric TF-IDF approach is promising when compared to the manual document clustering effort undertaken by the three journalist expert users. A number of clusters performed very well and in some cases almost as well as the expert users. It should be noted however, that the best performing clusters appeared to contain terms that are somewhat unique to the topic with little or no overlap with other terms in the collection. Clusters on more general topics such as health and sport performed fairly but not optimally and the limitations of natural language processing and document clustering become apparent here.

Optimising the performance of the toolset would be a key consideration if the document sample was to be increased significantly. The toolset was developed on a standard personal-use Windows laptop. The toolset was tweaked to provide support to run this on a Unix server. Even at this however, significant scope exists to optimise the code. In retrospect, PHP may have not been the best implementation language to use. Python offers specific support for some TF-IDF and vector model functions, is efficient and fast and may be a better choice for any future implementations.

Drawing conclusions from the results, it would infer that to produce an effective document clustering result, it is key that the distinguishing terms in each article are identified. The dimensionality of the document collection is too large to treat each term with equal importance. Boosting the importance of headline and summary terms, and increasing the similarity of already similar articles when they are published in close proximity to each other appears to have had a positive effect of the effectiveness of the document clustering. In addressing the reasearch question, the newly proposed news-centric term weighting and cosine similarity scheme did improve document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach.

6.5 Future Work & Research

This research could be improved by extending the scope of the natural language data preparation strategies. Introducing functionality to handle synonyms and identifying a strategy for the polysemy problem has the potential to further reduce the dimensionality of the document set under consideration. Another improvement could be delivered by proving functionality to support unigram and bigram frequencies. Bigrams frequencies identify interesting and not easily explainable patterns of word usage that habitually appear together like “weapons of mass destruction”, “abortion debate” and “European union”. Words that occur as a bigram statistically more often than their individual frequencies suggest that the word order is important and should be maintained. Further semantic preservation techniques such word stemming and entity extraction algorithms could be deployed to preserve as much of the intended semantic meaning as possible.

Evaluating the document clustering results was a significant challenge. As noted in the literary review, one of the main challenges stems from the fact that was each document collection is unique; it is not possible therefore to compare the clustering results with other document representations. In retrospect, this study may have been better served if one of the publically available news data sets such as the Reuters-21578 corpus was used. This would have allowed for a direct comparison with other research and the results of the study may have had a more quantitative grounding. The extent and severity of the evaluation challenges were not apparent at the time when the Irish Times document corpus was selected. For future implementations, the use of a more standard document collection would be advisable so that the results can be more easily evaluated.

The proposed term weighting and document similarity modifications and are not set within any specific theoretical framework. Rather, a common-sense approach has been used to boost the importance certain terms and features within news documents. A future implementation may benefit from a more theoretical approach.

The architecture of the toolset is modular and extensible and based upon loosely coupled language and data processing libraries. This approach aims to facilitate future researchers in customising the use of the toolset allowing for extensions of the toolset in the future if required. The toolset also provides a GEXF output function to provide support for future user tools. Including a GEXF export function extends the usefulness of the toolset by providing access to additional data visualisation, analysis and exploration tools.

6.6 Conclusion

In addressing the research question, the newly proposed news-centric term weighting and cosine similarity scheme did improve document clustering accuracy and topic aggregation capabilities for news articles when compared to the traditional term weighting approach. Whilst the experimentation shows that that the developed approach is promising, it also highlights the challenges of natural language processing and document clustering methods in general. It may not be possible to ever fully automate this process given the complexities of natural language processing but the automated method may go some of the way. The results may suggest that a blended approach of complimenting automated methods with human-level supervision and guidance may yield the best results.

BIBLIOGRAPHY

Aggarwal, C. C. & Zhai, C. (2012) 'A Survey of Text Clustering Algorithms', in Charu C. Aggarwal & ChengXiang Zhai (eds.) *Mining Text Data*. Springer US. pp. 77–128.

Aiello, L. M. et al. (2013) Sensing Trending Topics in Twitter. *Multimedia, IEEE Transactions*. Volume 15 (Issue: 6), pp: 1268–1282.

Aizawa, A. (2003) An information-theoretic perspective of tf—idf measures. *Journal Information Processing and Management: an International Journal*. 39 (1), .

Allen, JF 1987, *Natural language understanding*, Benjamin/Cummings, Redwood City, Calif.

Azzopardi, J. & Staff, C. (2012) Incremental Clustering of News Reports. *Algorithms*. 364–378.

Bassil, Y. (2012) Hybrid Information Retrieval Model For Web Images. *arXiv:1204.0182 [cs]*.

Basu, T. & Murthy, C. A. (2013) CUES: A New Hierarchical Approach for Document Clustering. *Journal of Pattern Recognition Research*. 8 (1), 66–84.

Bellman, R. (1957) 16399. *Dynamic programming*. Princeton University Press.

Bello, R. & Falcón, R. (2008) 00014. *Granular Computing: At the Junction of Rough Sets and Fuzzy Sets*. Springer Science & Business Media.

Bing-Quan, L. et al. (n.d.) *IEEE Xplore Abstract - Finding main topics in blogosphere using document clustering based on topic model.*

Blondel, V. D. et al. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2008 (10), P10008.

Boley, D. et al. (1999) *Partitioning-based clustering for web document categorization. Decision Support Systems.*

Brehmer, M. et al. (2014) Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists. *Visualization and Computer Graphics, IEEE Transactions*. 20 (12), 2271–2280.

Cai, D. et al. (2005) Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*. 17 (12), 1624–1637.

Cairncross, F 2001, *The death of distance: How the communications revolution will change our lives*, Harvard Business Press

Cenaiko, A. M. (2012) *Advertising Democracy: Audience Segmentation and Targeted Emails in the 2012 US Presidential Election*. University of Ottawa.

Chomsky, N 1957, *Syntactic structures*, The Hague.

Clifton, C & Cooley, R 1999, 'TopCat: Data mining for topic identification in a text corpus', in *Principles of Data Mining and Knowledge Discovery*, Springer, pp. 174–183

Cokley, J 2013, 'Journalism at the Speed of Bytes: Australian newspapers in the 21st century', *Digital Journalism*, vol. 1, no. 2, pp. 287–289.

Cunningham, H. et al. (2011) *DROPS - Challenges in Document Mining (Dagstuhl Seminar 11171)*

De Vries, C. M. et al. (2012) Document Clustering Evaluation: Divergence from a Random Baseline. *arXiv:1208.5654 [cs].*

Dhillon, I. et al. (2001) Efficient Clustering of Very Large Document Collections. *Data Mining for Scientific and Engineering Applications*. Volume 2 (Massive Computing), pp 357–381.

Dor, D. (2003) On newspaper headlines as relevance optimizers. *Journal of Pragmatics*. 35 (35), 695–721.

Dudoit, S. & Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*. 3 (7), RESEARCH0036.

Feldman, R. & Sanger, J. (2007) 01142. *The Text Mining Handbook*. Cambridge University

Feldman, R & Hirsh, H 1996, 'Mining Associations in Text in the Presence of Background Knowledge.', in *KDD*,pp. 343–346

Firth, JR 1957, 'Papers in Linguistics', *London: Oxford University Press.*

Flanagan, D. & Matsumoto, Y. (2008) 00229. *The Ruby Programming Language*. First. O'Reilly.

Fortunato, S. & Barthélemy, M. (2007) Resolution limit in community detection. *Proceedings of the National Academy of Sciences*. 104 (1), 36–41.

Fox, M 1998, 'A Changed Noam Chomsky Simplifies - New York Times', accessed February 6, 2014, from <http://www.nytimes.com/1998/12/05/arts/a-changed-noam-chomsky-simplifies.html>.

Franklin, B 2013, *The Future of Journalism*, Routledge

Fu, X. & Chen, M. (2008) 'Exploring the Stability of IDF Term Weighting', in Hang Li et al. (eds.) *Information Retrieval Technology*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 10–21.

Fuller, J 2010, *What is happening to news: The information explosion and the crisis in journalism*, University of Chicago Press

Gale, WA, Church, KW & Yarowsky, D 1992, 'One sense per discourse', in *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, pp. 233–237

Gilmore, D 2008, *We the Media: Grassroots Journalism By the People, For the People*, OReilly Media.

Gluck, M & Roca, M 2008, 'The future of television : advertising, technology and the pursuit of audiences'

Gray, J. et al. (n.d.) 00102. *The Data Journalism Handbook*. O'Reilly Media.

Green, D. (2007) *A State-of-the-Art Toolkit for Document Clustering*. PhD Thesis thesis. University of Dublin, Trinity College.

Grueskin, B, Seave, A & Graves, L 2011, *The story so far: What we know about the business of digital journalism*, Columbia University Press

Han, J. & Chang, K. (2002) Data Mining for Web Intelligence. *IEEE Computer Society Press*. 35 (11), PP 64–70.

Harding, J. (2015) The Future of News BBC Rreport

Harris, Z 1954, 'Distributional structure', *Oxford University Press*.

Holliman, R 2011, 'Telling science stories in an evolving digital media ecosystem: from communication to conversation and confrontation', *Journal of Science Communication*, vol. 10, no. 4, pp. 1–4.

Howard, A 2012, 'In the age of big data, data journalism has profound importance for society - Data'

Ifrim, G. et al. (2014) '*Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering*', in 2014

Jadon, C. & Khunteta, A. (2013) A New Approach of Document Clustering. *International Journal of Advanced Research in Computer Science and Software Engineering*. Volume 3 (Issue 4), .

Jain, A. K. et al. (1999) Data Clustering: A Review. *ACM Comput. Surv.* 31 (3), 264–323.

Jones, KS 1972, 'A statistical interpretation of term specificity and its application in retrieval', *Journal of documentation*, vol. 28, no. 1, pp. 11–21.

Karkali, M. et al. (2012) Keeping Keywords Fresh: A BM25 Variation for Personalized Keyword Extraction. *Proceedings of the 2nd Temporal Web Analytics Workshop*.

Kaye, J & Quinn, S 2010, *Funding journalism in the digital age: business models, strategies, issues and trends*, Peter Lang

Kolda, T. G. (1997) *Limited-Memory Matrix Methods with Applications*.

Kumar, R. et al. (2012) 'A Fast and Effective Partitioning Algorithm for Document Clustering', in Rajkumar Kannan & Frederic Andres (eds.) *Data Engineering and Management*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 264–271.

Kwak, H, Lee, C, Park, H & Moon, S 2010, 'What is Twitter, a social network or a news media?', in *Proceedings of the 19th international conference on World wide web - WWW '10*, Raleigh, North Carolina, USA, p. 591

Lewis, S. & Westlund, O. (2015) BIG DATA AND JOURNALISM Epistemology, Expertise, E conomics, and E thics. *Digital Journalism*. (Journalism in an Era of Big Data: Cases, Concepts, and Critiques), .

Liebscher, R. (n.d.) Temporal context: applications and implications for computational linguistics. *Proceeding ACLstudent '04 Proceedings of the ACL 2004 workshop on Student research Article No. 43*. 2004.

Lorenz, M 2013, 'Focusing on the future: Visualise the "tsunami of data" | Masters of Media'

Mahalakshmi, G. (2014) A Comparative Analysis on Web Clustering Algorithms. *International Journal of Engineering Research and Development*. Volume 10 (3), PP. 79–86.

Mahon, A. (2012) *Tribunal of Inquiry into Certain Planning Matters and Payments Bill 2004*.

McCandles, D. (2013) *The Data Journalism Handbook*. O'Reilly Media.

Mersey, RD 2010, *Can Journalism be Saved?: Rediscovering America's Appetite for News*, ABC-CLIO.

Meyer, P. (2009) 00000. *The Vanishing Newspaper [2nd Ed]: Saving Journalism in the Information Age*. Second Edition edition. Columbia: University of Missouri.

Meyer, P 2011, 'Nieman Reports | Precision Journalism and Narrative Journalism: Toward a Unified Field Theory'

Millar, J. & Peterson, G. (2009) *Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps.*

Miragliotta, N. et al. (2009) 00002. *The Australian Political System in Action*. 1 edition. South Melbourne, Vic: Oxford University Press.

Nadeau, D. (2007) *PHP tip: How to strip punctuation characters from a web page | Nadeau Software* [online]. Available from: http://nadeausoftware.com/articles/2007/9/php_tip_how_strip_punctuation_characters_web_page (Accessed 1 March 2015).

Newman, N. (2012) *The Reuters Institute's Digital News Report*.

Nothman, J. (2013) *Grounding event references in news*. University of Sydney.

O'Donnell, P. et al. (2012) Journalism at the Speed of Bytes: Australian newspapers in the 21st century. *Sydney: Media, Entertainment and Arts Alliance*.

Pantel, P. & Lin, D. (2002) 'Document Clustering with Committees', in *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '02. 2002 New York, NY, USA: ACM. pp. 199–206.

Pereira, F 2000, 'Formal grammar and information theory: together again?', *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1769, pp. 1239–1253.

Polettini, N. (2004) *The Vector Space Model in Information Retrieval-Term Weighting Problem.*

Rajaie, J. & Fakhar, B. (2012) A Novel Method for Document Clustering using Ant-Fuzzy Algorithm. *The Journal of Mathematics and Computer Science (JMCS).* 4 (2), .

Reed, J. W. et al. (2006) 'TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams', in *Proceedings of the 5th International Conference on Machine Learning and Applications*. ICMLA '06. 2006 Washington, DC, USA: IEEE Computer Society. pp. 258–263.

Ridgway, J. & Smith, A. (203AD) 'Open Data, Official Statistics and STATISTICS EDUCATION THREATS, AND OPPORTUNITIES FOR COLLABORATION', in *Statistics Education for Progress*. 203AD Macao: .

Robertson, S 2004, 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of documentation*, vol. 60, no. 5, pp. 503–520.

Rosell, M. et al. (2004) Comparing Comparisons: Document Clustering Evaluation Using Two Manual Classifications. *ICON*.

Rosell, M. (2009) *Text Clustering Exploration – Swedish Text Representation and Clustering Results Unraveled.*

Salton, G 1988, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*, Addison-Wesley, Reading, Mass.

Schutze, H 1992, 'Dimensions of meaning', in *Supercomputing'92. Proceedings*, IEEE, pp. 787–796

Shannon, CE 1951, 'Prediction and entropy of printed English', *Bell system technical journal*, vol. 30, no. 1, pp. 50–64.

Shawn Martin, W. M. B. (2011) OpenOrd: An Open-Source Toolbox for Large Graph Layout. *Proc SPIE*. 7868786806.

Singh, L, Scheuermann, P & Chen, B 1997, 'Generating association rules from semi-structured documents using an extended concept hierarchy', in *Proceedings of the sixth international conference on Information and knowledge management*, ACM, pp. 193–200

Singhal, A. et al. (1996) 'Pivoted Document Length Normalization', in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '96. 1996 New York, NY, USA: ACM. pp. 21–29.

Smeaton, A. et al. (1998) An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts. *British Computer Society*. Proceeding IRSG'98 Proceedings of the 20th Annual BCS-IRSG conference on Information Retrieval ResearchPages 10–10.

Steinbach, M. et al. (2000) 'A comparison of document clustering techniques', in *In KDD Workshop on Text Mining*. 2000

Strehl, A. (2002) *Relationship-based Clustering and Cluster Ensembles for High-*

dimensional Data Mining.

Tagarelli, A. (2011) *XML Data Mining: Models, Methods, and Applications*. 1 edition. Hershey, Pa: IGI Global.

Thalamuthu, A. et al. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*. 22 (19), 2405–2412.

Tirilly, P, Claveau, V & Gros, P 2008, 'Language Modeling for Bag-of-visual Words Image Categorization', in *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, ACM, New York, NY, USA, pp. 249–258

Turing, AM 1950, 'Computing machinery and intelligence', *Mind*, vol. 59, no. 236, pp. 433–460.

Van der Haak, B, Parks, M & Castells, M 2012, 'The Future of Journalism: Networked Journalism', *International Journal of Communication*, vol. 6, pp. 2923–2938.

Wallach, HM 2006, 'Topic modeling: beyond bag-of-words', in *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 977–984

White, B 2014, 'Jumping NLP Curves: A Review of Natural Language Processing Research', *IEEE Computational Intelligence Magazine*, vol. 9, p. 2.

Wang, X. et al. (2007) 'Topical n-grams: Phrase and topic discovery, with an application to information retrieval', in *In Proceedings of the 7th IEEE International Conference on Data Mining*. 2007

Xie, P. & Xing, E. P. (2013) Integrating Document Clustering and Topic Modeling. *arXiv:1309.6874 [cs, stat]*.

Yang, Y. & Liu, X. (1999) 'A Re-examination of Text Categorization Methods', in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. 1999 New York, NY, USA: ACM. pp. 42–49.

Yan, J. et al. (2006) Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. *IEEE Transactions on Knowledge and Data Engineering*. 18 (3), 320–333.

Yoo, I. & Hu, X. (2006) 'Clustering Large Collection of Biomedical Literature Based on Ontology-Enriched Bipartite Graph Representation and Mutual Refinement Strategy', in Wee-Keong Ng et al. (eds.) *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science. Springer Berlin Heidelberg. pp. 303–312.