THE APPLICATION OF TEXT MINING AND DATA VISUALIZATION TECHNIQUES TO TEXTUAL CORPUS EXPLORATION

THESIS

Jeffrey R. Smith Jr, Captain, USAF

AFIT-ENS-MS-18-M-163

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENS-MS-18-M-163

THE APPLICATION OF TEXT MINING AND DATA VISUALIZATION TECHNIQUES TO TEXTUAL CORPUS EXPLORATION

THESIS

Presented to the Faculty

Department of Operational Sciences

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Jeffrey R. Smith, BS

Captain, USAF

March 2018

AFIT-ENS-MS-18-M-163

THE APPLICATION OF TEXT MINING AND DATA VISUALIZATION TECHNIQUES TO TEXTUAL CORPUS EXPLORATION

Jeffrey R. Smith, BS

Captain, USAF

Committee Membership:

Dr. C. M. Smith
Chair

Dr. B. C. Boehmke
Member

Dedicated to my wife and family.

Abstract

Unstructured data in the digital universe is growing rapidly and shows no evidence of slowing anytime soon. With the acceleration of growth in digital data being generated and stored on the World Wide Web, the prospect of information overload is much more prevalent now than it has been in the past. This potential for overload may present users with issues such as difficulty finding and parsing relevant information, personalizing information to their needs, and using information to gain deeper insight. These difficulties also present a threat to industry leaders trying to maintain their share of the market and to military commanders defending this nation. As a preemptive analytic measure, organizations across many industries have begun implementing text mining techniques to analyze such large sources of unstructured data. This effort is not only being done to enhance their ability to retrieve and decipher known information, but to discover previously unknown information hidden within this unstructured data.

Utilizing various text mining techniques such as n-gram analysis, document and term frequency analysis, correlation analysis, and topic modeling methodologies, this research seeks to develop a tool to allow analysts to maneuver effectively and efficiently through large corpuses of potentially unknown textual data. Additionally, this research explores two notional data exploration scenarios through a large corpus of text data, each exhibiting unique navigation methods analysts may elect to take. Research concludes with the validation of inferential results obtained through each corpus's exploration scenario.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Christopher Smith, for his guidance and support through the course of this thesis effort. I would also like to thank my committee member Dr. Bradley Boehmke for his dedication to my understanding of the R language, and Dr. Brandon Greenwell and Mr. Andrew McCarthy for guiding me through the code development process.

Jeffrey R. Smith Jr, Capt

Table of Contents

page

List of Figures

List of Tables

THE APPLICATION OF TEXT MINING AND DATA VISUALIZATION TECHNIQUES TO TEXTUAL CORPUS EXPLORATION

I. Introduction

1.1 Background

There has never been a time in the history of the digital universe where information has been more readily available online. The explosive growth in digital data being generated and stored on the World Wide Web, has created a problem of information overload much more prevalent than it has been in the past. A study produced by IDC suggests that the size of the digital universe is doubling every two years[1]. Furthermore, the expected growth in this data presents users with issues such as difficulty finding and parsing relevant information, personalizing information to their needs, and using information to gain deeper insight into their data. These difficulties present a threat to industry leaders trying to maintain their share of the market and to military commanders defending this nation. To address the ever-increasing size of data created in this new information age, organizations across many industries have begun focusing their efforts on methods to analyze large sources of data. This effort is not only being done to

enhance their ability to retrieve known information, but to discover previously unknown information.

The copious amounts of information being created daily make it necessary to implement methodologies and applications to extract relevant knowledge and provide meaningful insight. Originally this solution presented itself in the form of data mining, also referred to as Knowledge Discovery from Data (KDD) [2]. KDD is a method to automate the extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories, or data streams [3]. The goal of the data mining process is to extract raw information from a large data set and transform it into understandable and useable information. These tasks are performed through the consideration of the applications of statistical and machine-learning methodologies to discover novel relations in large relation databases [2]. Data mining is an extremely powerful tool when applied to information that is already highly organized and easily searchable. This type of information is often referred to as structured data. While these KDD techniques work for structured data they are not as useful for data which is not highly organized such as digital media files, both audio and video, word processor documents, and text files. Such files as these are typically referred to as unstructured data. This data is not organized in a pre-defined manner as typically found in structured data. Unstructured data is typically text heavy, yet also contain multiple other data types such as numbers and dates. According to a 2011 study by IDC, unstructured data will account for more than 90% of the digital universe in the next decade [4]. Considering such a significant amount of data being unstructured and the high level of disorder inherent in these data sets it can be necessary to implement a method to

scour through this data and transform it into understandable information that can be easily used and analyzed. Text mining, also referred to as Knowledge Discovery from Text (KDT), is a specialized variant of data mining that facilitates the analysis of unstructured data [5]. It focuses on discovering unknown information and patterns, and extracting interesting, non-trivial information from large amounts of unstructured data. Text mining has continued to gain increasing attention in recent years. This attention stems from the large amounts of unstructured data that are being created from an ever-increasing number of sources, including social networks, data bases, and the world wide web to name a few. While the term text mining encompasses multiple topics, more specified applications maintain the primary goal of analyzing and discovering any interesting patterns, including trends and outliers, in text data. [6].

1.2 Research Goal and Objectives

Important aspects of text mining and the subsequent analysis of textual data reside in the mining techniques utilized, such as entity extraction, clustering, and visualization techniques such as network graphs. While analysts could read documentation and extract non-trivial information, this form of analysis can be extremely time and resource intensive. Employing these techniques, this research will focus on the development of a textual corpus exploration software package that would allow analysts to not only sift through unstructured data at an accelerated pace, but understand the data quickly and clearly.

The overall objective of this thesis research is to create a text mining R software package [7], using R version 3.4.3, that will quickly and effectively scour massive

corpuses of semi-structured data in the form of news articles. News articles are an excellent proxy for many other document types which may be of a proprietary nature, or whose content contains information that is not publicly available.

Three specific objectives of this thesis research include the following:

- Develop a methodology to allow the analyst to effectively and efficiently explore large textual data corpuses through the utilization of various exploratory text mining techniques
- Develop a robust user defined data manipulation feature that allows the analyst, applying their subject matter expertise to the context of the textual corpus, the ability to create, merge, separate, and delete terms from the corpus. Additionally, manipulated data will then be interpreted in subsequent analysis.
- Develop visualization methods as outputs to various text mining techniques used throughout textual corpus analysis.

The overall application of this package will identify and parse relevant information from each news article to be analyzed both individually and together as a single body of knowledge.

Additionally, applying statistical machine learning techniques in concert with a human-in-the-loop (HITL) factor, this package will allow analysts to be able to generate timely and useful data for decision makers. The algorithms applied to this R software package will focus on text mining techniques such as n-gram analysis, term correlation

analysis, and networks, and include information visualization methods to provide simple yet powerful graphics to aide analysis of these data sets.

1.3 Limitations

Data for this research was collected through accessing Global Database of Events, Language and Tone (GDELT) Project API. GDELT is one of the largest and most comprehensive open database of the world's news media consisting of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to the present [8]. GDELT's creator, Kalev H. Leetaru, has been studying the web and building systems to interact with and understand the impact it has made on society for more than two decades. The GDELT Project was born from the want to better understand global human society and the connection between communications and society's behavior [8]. This research relied on data generation using the GDELT API and only links housed within this database were captured.

This database was accessed to extract news article links via the development of a web scraping algorithm. Due to the unique architecture of each online news source and the direct increase in algorithmic complexity and time with each additional news source, it was decided to limit the focus of data collection to eight specific, and recognized, and "trusted" news sources. These sources include Reuters, BBC, CNN, Fox News, CBS News, USA Today, Washington Post, New York Times. Data was also limited to the article URLs captured from each of these sources via the GDELT database, meaning that this research would only have access to data contained in the GDELT database.

Research was also limited in the range of time the news articles were acquired. The method used to generate the URLs for each news source restricted the user of generating data 85 days prior to the date of retrieval. During URL data generation, the earliest date applicable for data collection was April 18, 2017. An upper bound date of August 31, 2017 was determined as a sufficient end date for data generation. All news articles, from the eight news sources, generated for this research were collected between the date ranges of April 18, 2017 to August 31,2017. The method used to generate this body of data will be discussed in Chapter III.

1.4 Assumptions

The creation of a textual corpus exploration R software package will rely heavily on the combination of multiple, previously existing, R packages. The main underlying assumptions in the use of these packages are that they will continually be monitored and updated as newer versions of R are released and that each package works as intended by their creators.

Additionally, it is assumed all news article data generated from the above-mentioned news sources are accurate accounts of historical events (i.e., no fake news). This assumption focuses on the database in which data collection is performed. It is also assumed that the sample of news articles generated during the established period represents a sufficient size to provide meaningful analysis and draw insightful conclusions from. Potential bias among the various news sources was assumed to nullify between the various news sources. Therefore, potential idiosyncratic biases were not considered in the selection of news sources.

II. Literature Review

2.1 Chapter Overview

Unstructured data accounts for more than approximately 90% of the digital universe and is projected to continue growing in the coming years [4]. There is speculation that the volume of unstructured data is growing at a rate of 62% per year and that by 2022, 93% of all data in the digital universe will be unstructured [9]. This expected explosion in unstructured data growth, along with novel methods to capture, organize, analyze, and act on this new information, has become a hot topic across a variety of industries. This chapter provides a summary of the analytical framework that industry leaders are currently using to analyze unstructured data as well as a summary of the previous work done relevant to the research presented in this thesis and is organized as follows. The first section attempts to provide a synopsis of the history of text mining, the methodology used to explore text based unstructured data, and explains how text mining differs from other text analysis fields such as Natural Language Processing (NLP) and Information Retrieval (IR). Next, definitions of common terminology and concepts found in text mining will be presented. Following this, business cases exhibiting how text mining is currently being utilized in both the medical field and in industry will be presented to help the reader better understand the usefulness of this analysis method. This section continues with a discussion of various text mining methods including IR, entity extraction, and named entity recognition (NER) approaches. Discussion then turns to an

explanation of corpus exploration. This section concludes with an examination of analysis methods used for word relationships.

2.2 Background Information

This research attempts to merge the distinct, yet similar, academic fields of Operations Research (OR), Natural Language Processing (NLP) and computer science to produce a robust, but easy to use, R package for the analysis of large textual data corpuses. While these fields maintain distinct representations in term of their respective knowledge base, many analytical concepts and vocabulary are shared between them. This section focuses on bridging any terminology gaps between the academic fields with concentration on the R package deliverable for this thesis. This section begins by defining some common terminology and concepts used in text mining, as well as this study, to provide a baseline understanding of more technical notions of text mining. Next is a brief discussion of a typical text mining pipeline used in various text analysis applications and an explanation of associated terminology and processes.

2.2.1 Terminology and Concepts

The following are explanations of terminology and concepts found throughout the reviewed literature and the research conducted for this thesis. Most definitions are provided through Dr. S. Vijayarani et al. [10] and Andreas Hotho et al. [11].

- Natural Language Processing (NLP) – a computational research avenue which explores methods in which computers can be used to understand and manipulate natural language

- Information Retrieval (IR) – the ability to retrieve information from a number of text-based documents
- Filtering – comprises the various methods of removing words from text documents. Words removed typically will bear little to no content information to the context of the text.
- Stemming - a method of filtering used to reduce the number of words accurately recognized in a text document. This procedure will remove the suffix of words to match the stems with the purpose of saving time and memory
- Lemmatization - a method of filtering used to reduce the number of words accurately recognized in a text document by attempting to map verb forms to the infinite tense and nouns to singular form.
- Stop Words – extremely common words found throughout different documentation which are usually of the form of articles, prepositions, and pro-nouns, etc. This words typically do not provide meaning to the documents.
- Tokenization – text preprocessing method of splitting continuous word streams by removing all punctuation marks and replacing all other non-text characters by single white spaces

2.3 Overview of Text Mining

Text mining is a method of uncovering hidden and extracting useful information with the purpose of assisting researchers who may be overwhelmed with vast amounts of textual data. The procedures and techniques involved in these processes differ greatly than that of the concept of Information Retrieval (IR). IR, in the academic setting, can be

described as the application of statistical techniques to search, index, and locate index specific material in large volumes of unstructured data [12].

While the information generally used for traditional data mining are typically housed in database systems, text data is typically managed via a search engine [13]. This method of managing text data allows users to implement information retrieval techniques to access specific topics. This method does not however, focus on analyzing text data to discover patterns and trends within the text and extract useful and high-quality information. While text mining may provide useful methods to fill these analytical voids, it is not a simple task.

It has been acknowledged that the bulk of information growth has become one of the most formidable communication issues to arise in past century [14]. Although an interest in the ability to automatically analyze textual data has existed for decades, the initial focus of research into data and gaining analytical insight has been dominated by the field of data mining and the use of highly structured data formats. This was due to the technological limitations placed on information management systems and the early creation and reliance on relational databases. Relational databases are a means of storing information in tables and grew to popularity in the 1970's. These databases utilize a predefined schema which store information in tuples (rows) and columns, making it a favorable method of storing numerical data. The structured nature of relational databases not only allowed data to be stored effectively, it also facilitated an efficient and effective means of data retrieval for use in business analysis. Data mining is an extremely powerful

tool when applied to information that is already highly organized and easily searchable such as this.

The goal of the data mining process is to extract raw information from a large data set and transform it into understandable and useable information. These tasks are performed through the consideration of the applications of statistical and machine-learning methodologies to discover novel relations in large relation databases [2]. While these KDD techniques work for structured data they are not useful for data that does not maintain such a highly organized structure such as is the case with unstructured data. Unstructured data is data that does not fit the structured schema of relational databases such as text heavy documents, pictures, and audio files for example. For this research, the term "unstructured" data will focus specifically on text heavy documents that may also contain multiple data types such as numbers and dates. With this level of disorder inherent in large data sets, it is necessary to implement a method to scour through and transform this data it into understandable information that can be effectively analyzed. This method of analysis takes the form of text mining and its applications can be found across various professional fields such as biomedical research, marketing, and intelligence organizations.

2.4 Applied Text Mining

Text and data mining are having a profound effect on the medical field, specifically in terms of biomedical research [15]. The Medical Literature Analysis and Retrieval System Online (MEDLINE) is a large bibliographic database containing life science and biomedical information including articles from academic journals.

MEDLINE holds information from a variety of medical fields and subjects including biology, biochemistry, medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care, to name a few. In 2004, the MEDLINE database, already containing 12.5 million records, was experiencing a growth rate of 500,000 new citations each year [16]. A considerable growth rate such as this, would only increase the difficulty of keeping up to date with the new discoveries and theories being found across varying fields of biomedical research. The potential for information overload in this arena increased the chance for important connections to be missed for new medical findings, on top of the potential for currently undiscovered connections already missed within the database.

Biomedical researchers addressing these issues have recently implemented biomedical text mining, a combination of techniques from natural language processing and text mining with the goal of allowing researchers to identify needed information more efficiently. This method of analysis also allowed for a general shift of the burden of information overload from the researcher to the computer [16]. To further expand on the capabilities and implementation of biomedical text mining, communities, such as the chemical compound and drug named entity recognition (CHEMDNER) task group, were formed to establish a united effort to evaluate biomedical text mining applications [17]. Application and integration of these novel biomedical research initiatives has resulted in the creation of multiple tools to assist with analysis. An example of how the biomedical community is utilizing text mining is through the use of Named Entity Recognition (NER), whose goal is to identify the frequency for names of specific drugs throughout a body of text or more specifically with a collection of journal articles. Methods such as

NER will identify these drug names in the hopes to extract further relationship information from the entity [16].

Advances such as this are allowing biomedical researchers to model, analyze, and understand complex biomedical, biological, and chemical systems at an accelerated pace through the analysis of text [18]. Applying text mining techniques to biomedical datasets Hu et al. [19] explores a statistical epistasis networks (SEN) approach to bladder cancer data. In short, the application of text mining, in concert with association-mining techniques, have assisted in the production of a methodology which serves to be a promising tool to identify, previously identified, higher-order genetic relationships involved with potential tumor cell expansion.

The evolution of unstructured data in the digital universe has allowed businesses to connect with their customers in ways that were previously inaccessible. Consumer-generated content, data posted by consumers online through different mediums such as social media pages, blogs, and product reviews, is being used by companies to find novel combinations of customer "needs" that represent profitable new opportunities [20]. In a study by Netzer et al. [21], they illustrate the application of text mining to explore the market structure and brand-associative network derived from online customer forums discussing specific product categories. Throughout the course of the study, they leveraged associative and semantic networks to assess the proximity (or similarity) between several terms based on the frequency of their co-occurrence. The data that this study focused primarily on was customer data concerning cars. Here, the motivation was to find co-occurrences between two car brands, models, terms used to describe them and

all combinations in between. The researchers were able to identify connections that consumers were making about specific vehicles. For example, consumers frequently commented on the "plastic parts and interior" of the Toyota Corolla, as well as its "good mileage". Using text mining in concert with network analysis, the researchers were able to develop a market structure from consumer-generated text that was highly correlated with the market structure derived from traditional data collection methods such as survey, transactional, and brand-switching data. The application of text mining to market research is now allowing firms to assess online consumer discussions of their products, placement, branding and allows them to monitor their market position from a higher resolution position.

2.5 Information Extraction

The starting point for computers to begin analyzing unstructured data is to use information extraction. The goal of information extraction is to locate specific pieces of data from a corpus of natural-language texts [2]. This process is used to identify items such as people, places, and time to further provide meaningful information and insight to the researcher, and to uncover relationships within the text. Constructing an information extraction tool can often times be a complex undertaking, however. In order to alleviate this complexity, the task of information extraction naturally decomposes into a series of processing steps, typically including sentence segmentation, tokenization, part-of-speech assignment, and the identification of named entities, i.e. person names, location names, and names of organizations [11]. Figure 1 provides a visual example of an entity extraction pipeline architecture.

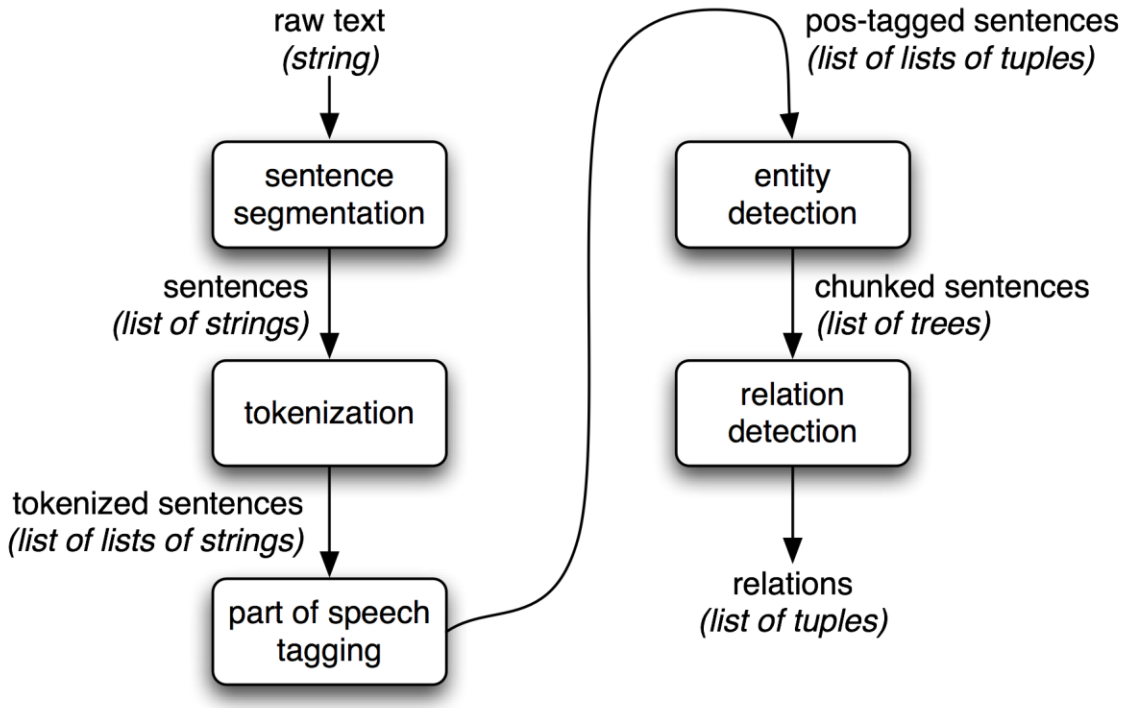


Figure 1. Example of Entity Extraction Pipeline Architecture [22]

Figure 2 provides an example of how text can be identified and tagged using entity extraction principles.

Figure 2. Example of Entity Extraction Technique [23]

2.6 Named Entity Recognition

NER is the task of locating and classifying names in text [24]. According to Jiang et al. [25] NER is probably the most fundamental task in information extraction. She goes

further to explain that extraction of more complex structures such as relations and events depends on accurate named entity recognition as a preprocessing step. NER presents its own set of complexities as with information extraction. An entity may be mentioned over multiple instances throughout the course of text, but may be represented by multiple titles. For instance, "President Donald J. Trump" maybe titled as "President Trump," "Donald Trump," or simply "Trump". While the text mining task could identify four unique entities, the user is aware that all four represent the same individual. As explained in [25], an entity also maintains the potential to be context-dependent. For example, the recognized entity, "JFK", may refer to the person "John F. Kennedy," or the location "JFK International Airport". Here, it is necessary for the context of the entity to be considered to determine the entity type for "JFK" occurring in a document.

Context, is one of multiple challenges faced in application NER technology. Additional fundamental challenges that persist in this application are in the determination of the boundaries of the entity names in the text. This comes to fruition when analyzing entity names such as "Procter and Gamble" where that is the name of the single entity, but has the potential to be recognized as two distinct entities, "Procter," and "Gamble". This is problem is referred to as the entity delimitation problem [24]. Downey et al. describes another fundamental challenge to supervised NER techniques as the unseen classes problem. This problem focuses on the idea of ambiguity in the named entity recognition technology and the impracticality of hand tagging elements of each entity class to train supervised techniques.

2.6.1 Rule-Base Approach

There are multiple methods used in approaches to NER. One approach is to manually develop information extraction rules by encoding patterns. Typically, in rule-based methods a set of rules is either manually defined or automatically learned. Each token in the text is represented by a set of features. The text is then compared against the rules and a rule is fired if a match is found [25]. However, due to the variety of forms and contexts in which the desired information can appear, manually developing patterns is very difficult and will rarely result in a robust system [26]. Manually developed patterns are optimized to work well in specific situations whereas, it is likely that both the language and user's needs will evolve over time. Additionally, this method of manually developed patterns may present difficulties in reflecting text that are less well behaved such as text that contains misspelled words or foreign words/phrases [24]. While rule-based approaches to named entity extraction and information extraction have provided sufficient results in the field of text mining, there has been more recent work accomplished in developing NER technologies via the use of statistical machine learning [27].

2.6.2 Statistical Learning Approach

Machine learning is programming computers to optimize a performance criterion using example data or past experience. It uses the theories of statistics in building mathematical models, because the core task is making inference from a sample [28]. Using statistical approaches, many NER algorithms treat the task as a sequence labeling problem. To map NER to this type of problem, each word in a sentence is treated as an

observation. The class labels must clearly indicate both the boundaries and the types of named entities with the sequence. Like that of most statistical analysis applications this learning approach can be categorized as either being a supervised or unsupervised learning processes. During supervised learning, the program can learn to classify a given set of labeled examples by the user who is teaching the program. Alternatively, unsupervised learning requires the program to build representations from the data without feedback from the user. Most unsupervised learning techniques used for NER are not completely unsupervised [29], however. Various approaches to statistical learning for NER include Hidden Markov Models, Decision Trees, Conditional Random Field (CRF), and Maximum Entropy models.

2.6.3 Dictionary Based Approach

Drug name recognition (DNR) is the designation given to the text mining method which seeks to recognize the mention of various drugs (title and by component) in unstructured medical texts. DNR is an NER task which seeks to classify these names into pre-defined categories [30]. An issue that biomedical researchers face is that drug names and designations in biomedical research vary widely among different authors and publications. This, coupled with the fact that there are no set rules for how drug names must be entered and the potential of typographical errors, presents many challenges for simple queries and the reliance on well-known medical dictionaries for name matching. To combat these issues, researchers have implemented numerous supervised machine learning approaches, retrieving benchmark training data sets from various sources such as medical case reports such as ADE [31] and CHEMDNER [32].

A study by Liu et al. [30] provides an overview of the DNR system process and methods used amongst the biomedical research community. As one could expect, different researchers approach this problem in many different manners, establishing unique procedures for each method. Figure 3 illustrates a typical approach to the DNR system.

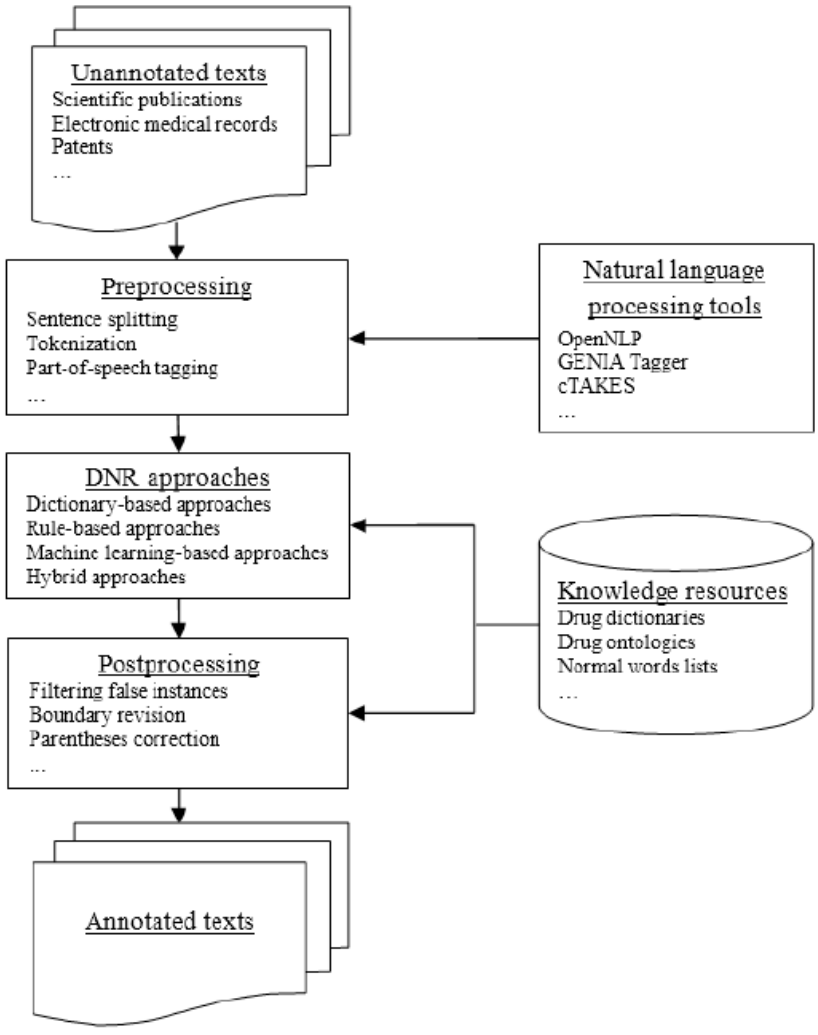


Figure 3. Typical Procedure of a DNR System [30]

Here, the preprocessing step refers to the transformation of the original input texts into representations that enrich them with lexical and syntactic information. Many of these enrichments have been identified in earlier sections of this research.

The Dictionary Approach, listed under the DNR Approach category will be the focus here. Hettne et al. [33] speaks to the usefulness of the dictionary approach and its dependence on how well the terms are suited for natural language processing. Biomedical research application of the dictionary approach have involved the creation of combined dictionaries [34] as well as the use of online databases of chemical compounds [35]. Many of the techniques used in the application of dictionary approaches for NER rely on semi-supervised and supervised means of statistical classification to identify and categorize entities and events. One of the major difficulties with this method is found in the base creation of the dictionary itself. Hettne et al. [33] identified some key issues involving the quality of some of the dictionary structure representations including the mis-association of chemical names with chemical entities and the overall assertation of what is to be considered a correct chemical structure and who asserts that it maintains that specific representation. The determination of how a compound is represented is based on the collective efforts of the company registering the compound, the patent, and the association of the compound amongst various databases. This system of determination must be taken into account when considering the validity of a compound's insertion into a dictionary as it will be an essential component to the results of the classification algorithm. To combat these issues, a curation platform was established in

which chemists could participate in the validation of many relationships within a dictionary.

2.7 Corpus Exploration

The concept of quickly exploring large document collections to discover useful and insightful information has been an exciting notion among researchers. One of the challenges with this area of texting mining is there is no definitive answer to how to perform this level of analysis. Eisenstein et al. [36] proposed TopicViz, an interactive environment that utilizes topic modeling in concert with various visualization techniques to identify latent themes found throughout large collections of documents. Lagus et al. [37] presented another method denoted as WEBSOM, an unsupervised method for the automatic organization of full-text document collections which utilizes the self-organizing map (SOM) algorithm. This method was found to be especially useful for exploration tasks in which the user only had a limited view of the contents contained within the corpus. Using this method, the algorithm would order the documents in some meaningful pattern based on their content. These results would then be presented to the user in the form of a document map in which the user could explore the overall view of what the document space looked like. While these methods both present forms of exploration into a large corpus of documentation, each primarily focuses efforts on the grouping of similar documents based either on latent topics or similarly related subject matters of documents. This does not satisfy exploration and visualization of the purely the content of a corpus however.

As a more in depth approach to the exploration of large corpuses of documentation, Ignat et al. [38] developed an automatic text analysis software which separates documents into clusters and extracts information such as a list of keywords, geographical locations mentioned, names of individuals and organizations and a list of user specified terms located within each cluster. The described system identifies keywords and names found within the text and provides linkages the exact phrases in which these terms are mentioned as well as to external websites to provide additional information. While the tool developed here presents users with an abundance of insight into the data through combination of IR and IE techniques, it does not detect linkages between terms found in the text. The research presented here provided a sound launch point for this study to utilize in its development.

2.8 Word Relationships

This section introduces key concepts in the form of extended definitions, focused on developing relationships between linguistic units. The information presented here is based heavily on the research and documentation presented by Manning and Schutze [39]. They describe statistical inference as the concept of taking data, that has some unknown probability distribution, and making inferences on its distribution. As an example, through the examination of a training set of data, a researcher may seek to statistically infer, or predict, the next phrase given of a sentence of an equivocal class. To perform this type of prediction effectively, it is necessary to group words based on their histories. Using the Markov assumption is one possible way of doing this. This concept essentially looks at the last few words and determines how they affect the next word to

follow. To do this *n - 1* words are placed together creating an $(n\text{-}1)^{th}$ order Markov model, also known as an n-gram model. Provided enough training data, researchers may derive a good probability estimate with

$$p(w_n|w_1, \dots, w_{n-1}) = \frac{p(w_1,\dots,w_n)}{p(w_1,\dots,w_{n-1})}.$$

This is one method of uncovering relationship between different n-grams found in text. A secondary method would be in the analysis of the relationships between common words that exist within the same document within the corpus. These relationships can be discovered through correlation analysis. Here, the focus is to determine how often words appear together, within the same document, relative to how often they appear on their own. To implement this formulation, this research focuses on the phi coefficient which strongly relates to the Pearson correlation coefficient. This correlation coefficient measures the strength and direction of linear association between two variables [40], or words for the purposes of this research. The specific method of correlation used in this research will be discussed more thoroughly in the following sections.

2.9 Summary

This chapter presented some of the highlights of the previous works related to text mining techniques. The chapter opened with a definition of data mining, the challenges that unstructured data poses to data mining, and how text data mining can alleviate those issues. The focus then shifted to illustrate how text mining applications are being used in both the medical field and private sector. Following this, discussion moved to discuss

various methods of discovering insightful information and making connections with the application of text mining. The chapter concluded introducing two concepts for discovering word relationships, n-grams, and the use of the correlation coefficient.

III. Methodology

3.1 Chapter Overview

The purpose of this chapter is to provide an introduction into the various text mining techniques applied to the methodology developed during this thesis research. Discussion begins with the explanation of how users of this package can navigate their exploratory analysis. Then the discussion focuses on the concepts underlying n-gram development and analysis. Next is a discussion of term frequency analysis, the measure of how frequently terms occur in documentation, followed by a discussion of term correlation analysis, a statistical technique used to expose relationships between terms in the same document. Following this is a brief overview on the application of topic modeling to a corpus of documents. This chapter ends with a discussion focused on the graphical tools used to visualize the results of the aforementioned text mining techniques.

3.2 Corpus Exploration Method

The purpose of this section is to introduce the methodological flow of analysis this research seeks to develop. Figure 4, is a visual representation of the application of these various methodologies in sequence. This flow chart serves to outline the possible directions an analyst may take when conducting their analysis.

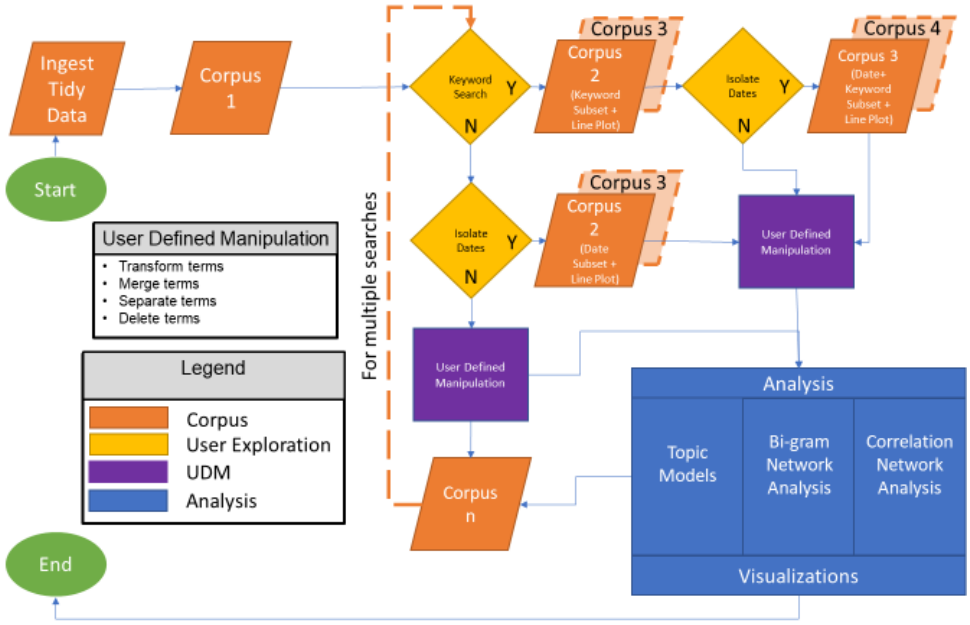


Figure 4. Flow of Analysis

The program is initialized with the ingestion of data. To correctly implement this software package, data is required to be in a data frame format with, at minimum, variables identified for the Date, Text Body, and Article Identification. This package utilizes "tidy data" techniques [41] for all data structures. Once tidied, data will enter first into the "User Exploration" phase where, as the name suggests, the analyst will be provided the functionality to explore the dataset to obtain any preliminary insight that is to be gained. Following this, the user will transition into the "User Defined Manipulation" phase where, applying their subject matter expertise, they have the ability to transform, merge, separate, and delete terms found throughout the exploration phase. Once the user has successfully manipulated the dataset, the package can move into the "User Analysis" phase in which they can use text mining techniques such as bi-gram

network analysis, correlation network analysis, and topic models to visualize the results of their analysis and make inferences into the dataset.

3.3 Text Mining Techniques

This section will discuss the various text mining techniques utilized through this research. It begins with a discussion of the bag-of-words methodology and its relation to n-gram analysis. Discussion continues with definitions of temporal document and term frequency, and correlation analysis. This section concludes with an explanation of the topic modeling method used and visualizations of the results.

3.3.1 Bag-of-Words

Text documents are considered unstructured as compared to the highly structured fields of numerical data. In this form, it is difficult to analyze the content of textual documents, and even more so to mine them for valuable insight. To lessen this difficulty one of the most widely used approaches to text mining is the concept of representing text within a document as a "bag-of-words". Before explaining how this concept is used in text mining it is first necessary to provide an explanation of vector space representation. This concept takes a word and defines it as a value of numerical importance [11]. This model represents each document d as vector in m-dimensional space with each described by a numerical feature vector $w(d) = \left(x(d, t_1), \ldots, x(d, t_m)\right)$. This vectorization allows documents to be compared using various vector operations.

The bag-of-words concept seeks to be a method of extracting features from documents by examining text in documentation as unique numerical values and

measuring the frequency of those words throughout the document[42]. Here, each word count is considered a feature.

3.3.2 Preprocessing

Prior to each analysis performed throughout this research, it is necessary to perform data preprocessing on all text within the corpus. Initially, upon ingestion, documents maintain their original structure which includes all words founds in the text, punctuations, capitalized letters, and full and correct spellings of each term. The method of preprocessing this data for analysis purposes include applying computational procedures such as the removal of stop words, transforming all text to its lower-case form, and the removal of all punctuation and white space to all for effective tokenization of each term. This research elected not to use a stemming algorithm throughout the preprocessing. Traditionally, stemming seeks to remove the ends of words in an effort to reduce words to a base form, but this application is often performed in a crude manner procedure [12]. It was elected to omit stemming from data cleansing processes to not skew the results of the analysis.

3.3.3 N-grams

N-gram analysis [43] focuses on identifying a contiguous sequence of *n* terms in a given sequence of text. To identify contiguous sequences, it is necessary to tokenize the textual content of the documents within a corpus. Tokenization [44] is the process of segmenting the linear sequences of text into smaller linguistic units. These units may consist of segments such as punctuations, word(s), sentences and paragraphs to name a few. The *n* is designated by the number of terms contained within a given token. When *n*

is equal to 1, this is designated as a unigram. When *n* is equal to any value greater than 1 it denotes how often *wordX* is followed by *wordY* and beyond. Table 1 illustrates the n-gram relationship and how *n* is used. N-gram analysis seeks to provide a level of enhanced insight into a corpus' content.

Table 1. N-gram Example

N-gram	(n)	Token
unigram	1	"trump"
bi-gram	2	"donald trump"
tri-gram	3	"president donald trump"
four-gram	4	"u.s. president donald trump"

3.3.4 Term Frequency Analysis

Term frequency (tf) is a measure of how frequently a term occurs in a document. Every document within a corpus should be considered a unique instance in which it is possible that a term can appear many more times in some than others. For use in this project, this calculation is simply a count.

For a more formal description of the term frequency algorithm, let D be the set of documents and $T = \{t_1, \dots, t_m\}$ be the different terms occurring in D, then the absolute frequency of $t \in T$ in document $d \in D$ is given by tf(d, t). When applied to a subset of terms, the term frequency can be written as $tf(d, T') \coloneqq \sum_{t \in T} tf(d, t)$.

Applying term frequency, the analyst focuses on the most frequent terms that appear throughout the corpus. The most frequent terms will provide some indication of

the main topics of discussion throughout the text. Term frequency is used in concert with n-grams at both the document and term level to provide values for some of the plots used throughout this research.

3.3.5 Term Correlation Analysis

Term correlation analysis seeks to expose the relationships between terms that are found in the same document, but may not co-occur such as with n-grams [45]. To accomplish this, pairwise correlation is used to indicate how often terms appear together relative to how often they appear independently. Generating these correlation values, the pairwise correlation relies on the calculation of the phi coefficient [33],

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{[\![(n_{1\cdot})(n)]\!]_{0\cdot})(n_{\cdot 0})(n_{\cdot 1})}}$$

where n_{11} is the case where both words appear, n_{00} is the case where neither word appears, and n_{01} and n_{10} are cases where either word appears. Table 2 illustrates the how the term adjacency matrix is generated to populate these values.

Table 2: Term Adjacency Matrix Example

	y = 1	**y = 0**	**total**
x = 1	n_{11}	n_{10}	$n_{1\cdot}$
x = 0	n_{01}	n_{00}	$n_{0\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

If most of the values between two terms falls on the main diagonal, then a positive correlation is indicated.

3.3.6 Topic Modeling

Topic modeling is a very widely used unsupervised classification method for documentation. For this research, the Latent Dirichlet Allocation (LDA) [47] method of topic modeling was used. The main concept of LDA is that within a corpus, latent (natural) topics are present among the documents with each word in each document contributing to a specific topic. LDA treats each document as a mixture of topics, with each topic as a mixture of words [45]. The purpose of this analytical approach is to discover hidden topics that pervade the corpus. Following this discovery, each document will be tagged with its respective topic, organized, explored, and then the content of these hidden topics will be summarized. The two assumptions made during this analysis are that (1) each document exhibits each topic in different proportions and (2) each word in each document is drawn from one of the topics [48]. As a simple explanation of the complex math used in LDA, imagine that for each of the possible number of topics (k), the algorithm will multiply the frequency of the term found in the topic by the number of other terms found in the document already belonging to that topic. This explanation is expressed by,

$$P(Z|W,\ D) = \frac{(\#\ of\ word\ W\ in\ topic\ Z) + \beta_w}{(total\ tokens\ in\ Z) + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha),$$

where Z represents possible topics within each document, W represents each word in a document, D represents each document, and α and β are user defined hyperparameters.

When using LDA, the appropriate determination of the number of topics is one of the more crucial elements involved in the accuracy of the algorithm. When the parameter

k is to small the potential for loss information pervades the analysis. Conversely, when k is to large LDA cannot capture all correlations while also inferring the documents posterior distribution over topics leading to an inaccurate representation of the data [49]. To determine the appropriate number of topics (*k*), this research utilizes the *ldatuning* package [50], which realizes four metrics in its selection of the appropriate number of topics for LDA modeling. The following is a summarization of each of the methodologies utilized in the determination of the appropriate number of topics used in LDA, a sample of the visualization of expected results, and a discussion on how an analyst utilizing this tool can interpret the results.

3.3.6.1 Cosine Distance Minimization Method

Cao, Xia, Li, Zhang, and Tang [49] present an approach to the determination of the appropriate number of topics by adaptively selecting the best LDA model based on topic density. In this approach they determine that the best *k* is correlated with the distances between topics. The cosine distance is used to measure the correlation between topics with

$$corre(T_i, T_j) = \frac{\sum_{v=0}^{V} T_{iv}, T_{jv}}{\sqrt{\sum_{v=0}^{V} (T_{iv})^2} \sqrt{\sum_{v=0}^{V} (T_{jv})^2}}.$$

The smaller the $corre(T_i, T_j)$ the more independent the topics, therefore establishing this as a minimization metric. The output of the cosine measure is then used in the calculation of the average cosine distance between every pair of topics. Using this method, Cao et al. were able to establish a correlation between the best *k* and the distances between topics. Given a topic and the average cosine distance, they establish the topic density as the

number of topics within this radius. The k^{th} topic with the smallest topic density is then deemed the optimal k.

3.3.6.2 KL-Divergence Minimization Method

As a secondary approach to a minimization metric, Arun, Suresh, Madhavan, and Murty [51] propose a measure to correctly identify the number of topics in a corpus through the use of the Symmetric Kullback-Leibler (KL) divergence. This research focuses on the notion that LDA's probabilistic generative model can be viewed as a non-negative matrix factorization method. Assessing the concept of data in this way allows for the separation of a Document-Word Frequency Matrix *M* into a Topic-Word Matrix *M1* and a Document-Term Matrix *M2* where the values between the two matrices are identical. Before presenting the divergence model, Arun et al. proceeds to explain how the concepts of Singular Value Decomposition (SVD) and topic splitting are incorporated into the theory of their model. They introduce the theory of the Symmetric KL divergence explaining that if given a random Topic-Word Matrix *R*, the vector representing the distribution of row L_1 norms, R_{l1}, and the vector representing the distribution of L2 norms, Rl2, will be very similar component-wise when the number of words *W* in a corpus is large enough. As *W* becomes large enough the Symmetric KL-divergence will go towards zero. With this proposition they presented,

$$ProposedMeasure(M1, M2) = KL(C_{M1}||C_{M2}) + KL(C_{M2}||C_{M1}),$$

where C_{M1} is the distribution of singular values of Topic-Word matrix M1, and C_{M2} is the distribution obtained by normalizing the vector L*M2, (where L is 1*D vector of lengths of each document in the corpus and M2 is the Document-Topic matrix).

With this measure, these researchers support that the appropriate number of topics will have the lowest calculated measure.

3.3.6.3 Information Divergence Maximization Method

Deveaud, Sanjuan, and Bellot [52] build upon the methodologies of Cao et al. and Arun et al. In the review of the works of these researchers, Deveaud et al. acknowledges the differences in their approaches, but note the similarities in the calculation of the distance metrics between topics over several instances of the model, all varying in the number of topics. They propose a method, a simple heuristic, which seeks to estimate the number of latent topics in a corpus by maximizing the information divergence *D* between all pairs of potential LDA topics (k_i, k_j). $\hat{K}$ represents the number of topics determined by this heuristic and represented by,

$$\hat{K} = \underset{K}{argmax} \frac{1}{K(K-1)} \sum_{(k,k')\in \mathbb{T}_K} D(k||k'),$$

where K is the number of topics given as a parameter to LDA, $\mathbb{T}_K$ is the set of K topics modeled by LDA, and D is the Jensen-Shannon divergence.

The method presented here uses the Jensen-Shannon divergence, stating that it is a symmetrized version of the KL divergence. The outcome of this method estimates the number of topics $\hat{K}$ and its associated topic models.

3.3.6.4 Markov Chain Monte Carlo Maximization Method

Griffiths and Steyvers present a unique method also focused on the concept of a maximization model. The researchers seek to obtain estimates for the topic's multinomial distributions over the words *W*, ϕ, and the set of document's multinomial distributions over the topics, θ, by considering the posterior distribution over the assignments of words to topics, $P(\boldsymbol{z}|\boldsymbol{w})$, where $\boldsymbol{z}$ is the vector of latent variables indicating the topics. Considering this distribution, the researchers found they were able to find estimates for ϕ and θ. To evaluate $P(\boldsymbol{z}|\boldsymbol{w})$, among other mathematical components necessary for this model, but beyond the scope of this thesis research, Griffith et al. also implemented the use of a Markov chain Monte Carlo procedure for their algorithm to converge to a target distribution. The Markov chain portion of this procedure called for the use of Gibbs sampling [53] as an indicator of when the next state would be reached. The Markov model portion of the algorithm leads to the development of a conditional distribution. The information developed here would then be used to determine the initial state of the Markov chain. At this point the chains are run over multiple iterations. After enough iterations the chain would begin to approach at target distribution which maximizes $P(\boldsymbol{z}|\boldsymbol{w})$.

3.3.7 Network Graph Visualizations

Data visualization [54] is the science of visually representing data, either categorical or quantitative, that has been abstracted in some form for units of information. Network theory is the study of graph visualizations representing complex systems of relationships between discrete objects. These graphs are typically represented by nodes,

which represent entities, and edges, which represent the interactions between these entities. For the purpose of networks towards the application of this thesis, focus is concentrated on undirected graphs and networks, Figure 5 [55].

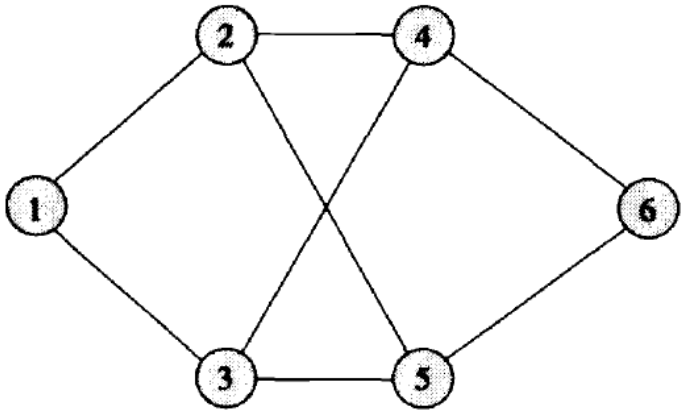


Figure 5. Undirected Network Graph Example [55]

Here nodes are represented by each individual term found within the documents in the corpus. Edges are represented by some value metric such as count or correlation.

IV. Results and Analysis

4.1 Chapter Overview

This chapter presents the application of the text mining and graphical techniques described in Chapter III to a large corpus of news articles in the form of case studies. The chapter begins with a detailed discussion of the data generation process. Following this, the focus shifts to a thorough investigation of the overall dataset. The remainder of the chapter is dedicated to the examination of the overall dataset using two case studies. The chapter concludes with the validation of the inferential results using unbiased, nationally recognized, and expert sources.

4.2 Data Generation

Data generation was executed in three distinct stages. Stage 1 involved the selection of eight viable news sources to be used throughout this study. Stage 2 involved a mass URL retrieval of the selected news sources from the Global Database of Events, Language and Tone (GDELT) Project API. Stage 3 involved the creation and application of web-crawling algorithms to extract specific data from each URL retrieved from the GDELT Project API to be compiled into a robust data frame.

4.2.1 Stage 1: News Source Selection

As stated previously, the purpose of this research is to develop a methodology to allow users to sift efficiently and effectively through a corpus of textual data to glean insightful information. With this purpose in mind, it was imperative that the data sources be of an accurate, reliable, and reputable nature. For the algorithmic convenience of web scraping, news sources were also selected based on the consistency of the webpage structure between articles. Table 3 identifies the eight, well-known and well-trusted, news agencies selected for this study.

Table 3. Summary of News Sources

Reuters	BBC	CNN	Fox News	CBS News	USA Today	Washington Post	New York Times

4.2.2 Stage 2: GDELT API 2.0

Data generation for this research utilized the GDELT GEO 2.0 API which debuted on April 26, 2017. Through many iterations, this API has become a robust engine that allows users to explore world events through a host of insightful visualizations. While this API maintains many advanced features, this research simply accessed the API for the generation of a list of URLs from the eight specific news sources identified above. The API was used to generate a collection of URL links, dates, and titles of online news articles of user specified news sources. The API was not without its limitations in use however. The API limited the user's data request to a maximum of 85 days from the date of use. For example, if a user initiated a pull request on July 17, 2017 the first date the API would retrieve would be on 18 April 2017. Additionally, each

request to the API was limited to a maximum of 250 records to be extracted at a time. Table 4 illustrates examples of the method this research used to access the API.

Table 4. News Source API Access Examples

CNN	http://API.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:cnn.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
BBC	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:bbc.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
CBS News	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:cbsnews.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
Fox News	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:foxnews.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
Reuters	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:reuters.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
USA Today	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:usatoday.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
New York Times	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:nytimes.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv
Washington Post	http://api.gdeltproject.org/api/v2/doc/doc?mode=artlist&query=sourcelang:english+domain:washingtonpost.com+sourcecountry:unitedstates&startdatetime=20170418000000&maxrecords=250&format=csv

Raw data files were captured through the development of an algorithm the author created which looped through each day of each month in the API links above. From the point of instantiation, the algorithm captured 250 records per news source, per day. Following this data capture, all files were merged, sorted, and deduplicated. Raw data was collected as Comma-Separated Values (csv) delimited files. Each csv file will produce the following four variables for each specified date.

• URL - the address of the World Wide Web page
• Mobile URL - the address of the World Wide Web page specifically for use on a mobile device
• Date - time and date in which the article was published online. (given in YYYY-MM-DD HH:MM:SS format)
• Title - title of the article generated by the URL

Table 5 provides an example csv output file.

Table 5. CSV Output Example

URL	Mobile_URL	Date	Title	NewsSource
http://www.cbsnews.c	http://www.cbsnews.	4/18/2017 0:00	Saturn moon Tit	cbsnews
https://www.usatoday	http://amp.usatoday.c	4/18/2017 0:15	1 dead , 2 hurt a	usatoday
http://www.cbsnews.c	http://www.cbsnews.	4/18/2017 0:15	Leading Democr	cbsnews
https://www.nytimes.c	http://mobile.nytimes.	4/18/2017 1:00	The (Other) Ot	nytimes
http://www.latimes.co	http://www.latimes.c	4/18/2017 1:15	Candidate endor	latimes

4.2.3 Stage 3: Web Scraping

To extract information from each URL, this project utilized R's *rvest* function [56] to scrape data from each news article's webpage. The *rvest* package required that we isolate the website specific html wrapper containing the information. Html wrappers identified for this research contained elements for the author's name and the article text. To accomplish this task, it was necessary to either enter the developer console of the website and extract the information or use a tool such as SelectorGadget which automates the process. Each of the news source's webpages had a unique, developer specific, structure. Due to the variation in webpage architecture, it was necessary to develop specific case structures within the algorithm to extract each desired html element from

each website. In some instances, multiple cases per website were required to scrape the necessary data.

The information which was collected from this portion of the project was comprised of each news article's author and text in separate columns. Fields where the author's information could not be scraped, the identifier "unknown" was used. Entries containing no information for the text column or duplicate information were removed completely.

4.3 Data Storage

After scraping each of the previously retrieved URLs for author and text information, it was necessary to amalgamate this information with the URL, mobile URL, publication date, title, and news source ID information previously captured. Following this combination, the data was then converted into a data frame. Table 6 illustrates an example of the data frame's structure.

Table 6. Example Data Frame Structure

Title	Author	Date	URL	NewsSource	Text	ArticleNo
Mother gives birth to one of largest	unknown	20170501 00:00:00	http://www.cbsnev	cbsnews	SACRAMENTO -- A newborn	1
Ex - Governor Bob McDonnell on S	unknown	20170501 00:00:00	http://www.cbsnev	cbsnews	Former Virginia Gov. Bob McD	2
At least 10 dead , 2 missing after to	unknown	20170501 00:00:00	http://www.foxnev	foxnews	At least 13 people had died and	3
OCC culinary students are now serv	Alex Chan	20170501 00:00:00	http://www.latime	latimes	Ahoy, hungry mateys! After yea	4
Nine new sculptures unveiled in Ne	Hannah Fry	20170501 00:00:00	http://www.latime	latimes	Newport Beach welcomed nine	5
Finding the Will to Party on a Mute	KATIE ROGERS	20170501 00:00:00	https://www.nytim	nytimes	WASHINGTON — It was not s	6

4.4 Overall Dataset

The analyzable dataset used for this research consists of 82,688 news articles. Table 7 outlines the date ranges over which the data was collected.

Table 7. Overall Data Date Range

Start Date/Time	End Date/Time
April 17, 2017, 20:00:00 EDT	August 31, 2017, 19:30:00 EDT

Table 8 outlines the number of articles provided by each news source per month and in total.

Table 8. Overall Data News Source Totals

News Source	**April**	**May**	**June**	**July**	**August**	**Articles per source**
Reuters	2920	5605	5938	7373	6708	**28,544**
USA Today	946	2439	2429	2252	2079	**10,145**
The Washington Post	960	2020	2039	2453	2303	**9,775**
The New York Times	1126	2729	3049	2209	0	**9,113**
CNN	828	1950	1960	1678	1467	**7,883**
CBS News	853	1974	1829	1656	1546	**7,858**
BBC	718	1835	1936	1665	1443	**7,597**
Fox News	1659	36	32	26	20	**1,773**
Total	**10010**	**18588**	**19212**	**19312**	**15566**	**82,688**

Table 9 provides the number of terms, or words, used for each news source summed across each month and in total.

Table 9. Overall Dataset Term Counts

News Source	**April**	**May**	**June**	**July**	**August**	**Terms per source**
Reuters	1,179,773	2,269,136	2,571,684	3,098,205	2,879,524	**11,998,322**
USA Today	615,922	1,514,874	1,502,924	1,343,390	1,253,782	**6,230,892**
The Washington Post	808,359	1,695,352	1,723,522	2,167,324	2,037,405	**8,431,962**
The New York Times	1,004,654	2,440,889	2,947,946	2,177,838	0	**8,571,327**
CNN	617,979	1,355,480	1,389,175	1,272,586	1,096,254	**5,731,474**
CBS News	504,029	1,134,221	1,072,666	1,019,930	877,912	**4,608,758**
BBC	386,309	1,014,080	1,072,114	871,837	746,779	**4,091,119**
Fox News	663,114	29,566	42,620	78,343	106,317	**919,960**
Terms per month	**5,780,139**	**11,453,598**	**12,322,651**	**12,029,453**	**8,997,973**	**50,583,814**

Use of this package requires that data be stored in a data frame with the body of text being analyzed to be stored under the column name “text” and the dates stored under the column name “date”. If analysis is being conducted on a corpus of text that does not maintain unique time/date features, but rather a sequential order such as that of chapters in a book, this information will need to be given the heading of “date”.

4.5 Case Studies

This section walks through the analysis of two distinct research avenues taken on the overall dataset. The first section presents a Directed Search case study and describes a possible temporal based scenario in which the application of this thesis research could be very beneficial. The next section focuses on an Undirected Search case study and denotes an alternative scenario which highlights additional levels of the exploratory capabilities within this thesis project. Analysis of each case study is performed using a different approach to demonstrate the versatility of the package. Each case study concludes with the presentation of an inference into the content of each sub-corpus, as well as a validation of said inference through a comparison with external and reputable sources.

4.5.1 Directed Search: Ballistic Missile Proliferation

Suppose this notional scenario: a business account manager returns to the office on September 01, 2017 after an extended absence. At the morning meeting, the account manager is informed that the company predicts that recent North Korean ballistic missile tests are expected to adversely affect profits for the year. The manager is expected to understand what sectors these recent missile tests may affect and a potential area where the firm can re-focus its financial efforts within the hour. The Vice President of Operations provides the account manager with a storage device containing 82,688 proprietary reports collected between April 18th, 2017 and August 31st, 2017 (Corpus 1). The account manager can try to find as much relevant information from this data source as he can in the limited amount of time. For the purposes of this scenario, it will also be

assumed that the specific information contained within Corpus 1 is sensitive and is not available on the internet.

In this case, a logical way to begin finding relevant information is to create a sub-corpus focused solely on articles containing the word "Korea." Using R to search for that term, the account manager would be able to separate and reduce the focus to a select number of articles. He then may want to focus on only those articles, within this already reduced set, that also reference the term "ballistic". This additional level of specification further reduces the corpus to 1,336 articles. This new sub-corpus, hereinafter referred to as Corpus A, contains 1,336 articles and 1,117,711 words, and would become the basis of the account manager's search for information.

Having created this detailed corpus, a logical progression for the account manager would be to perform an investigation into the frequency of documents produced during the given date range (18 April 2017 – 31 August 2017). Figure 6 illustrates the corpus plot for Corpus A. To denote areas where noticeable increases in articles have taken place yellow boxes have been added to this plot.

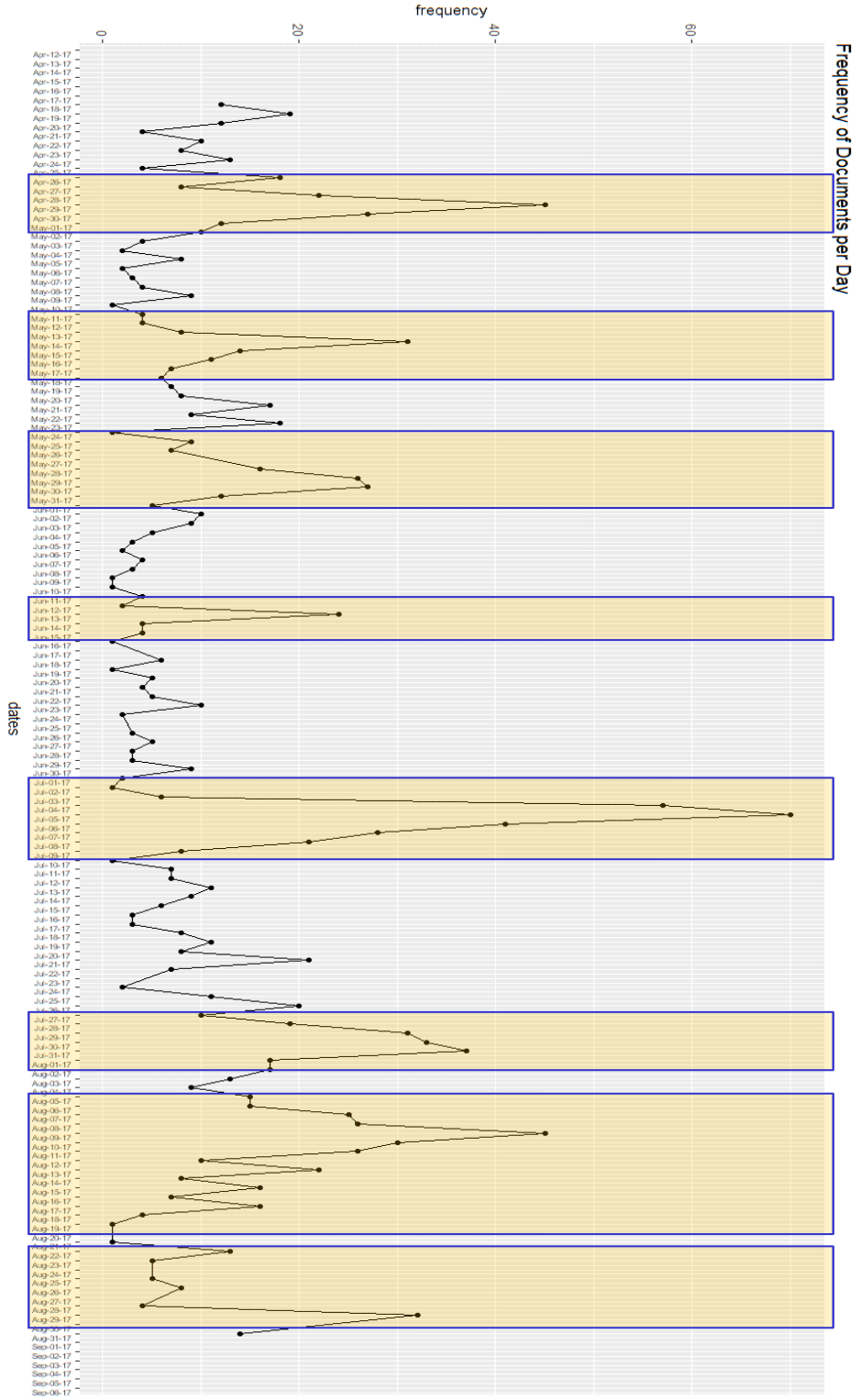


Figure 6. Corpus A Document Frequency

The above figure identifies eight distinct ranges where the article frequency has made noticeable shifts. One might assume that it's likely a North Korean ballistic missile was either launched or talked about on these days. To further investigate the events within each of these eight peaks, the account manager may elect to isolate each occurrence based on each respective date range. Table 10 denotes the summary information for each new sub-corpus, identifying each corpus's name, start and end dates, and the total number of articles and terms per corpus.

Table 10. Summary of Corpus A's Sub-Corpuses

Name	Start Date	End Date	Total Number of Articles	Total Number of Terms
Corpus A.1	26 April 2017	01 May 2017	120	101,994
Corpus A.2	11 May 2017	17 May 2017	69	48,463
Corpus A.3	24 May 2017	31 May 2017	90	70,679
Corpus A.4	11 June 2017	15 June 2017	27	14,116
Corpus A.5	01 July 2017	09 July 2017	224	176,346
Corpus A.6	27 July 2017	01 August 2017	136	111,541
Corpus A.7	05 August 2017	19 August 2017	250	267,191
Corpus A.8	22 August 2017	29 August 2017	64	41,607

Having effectively created these eight sub-corpuses, the account manager is now able to explore each and uncover information contained within. The following sections focus on the analysis of each sub-corpus, providing three examples of various types of information the account manager maybe able to uncover. Each example will demonstrate the visualizations this research supports and potentially inferences that could be made on the resulting information.

4.5.1.1 Corpus A.1: 26 April – 01 May 2017

The account manager may begin the examination of the first corpus, Corpus A.1. with an n-gram analysis to gain an initial understanding of the dataset. Examining the twenty most frequent bigrams found throughout the corpus he would produce Figure 7.

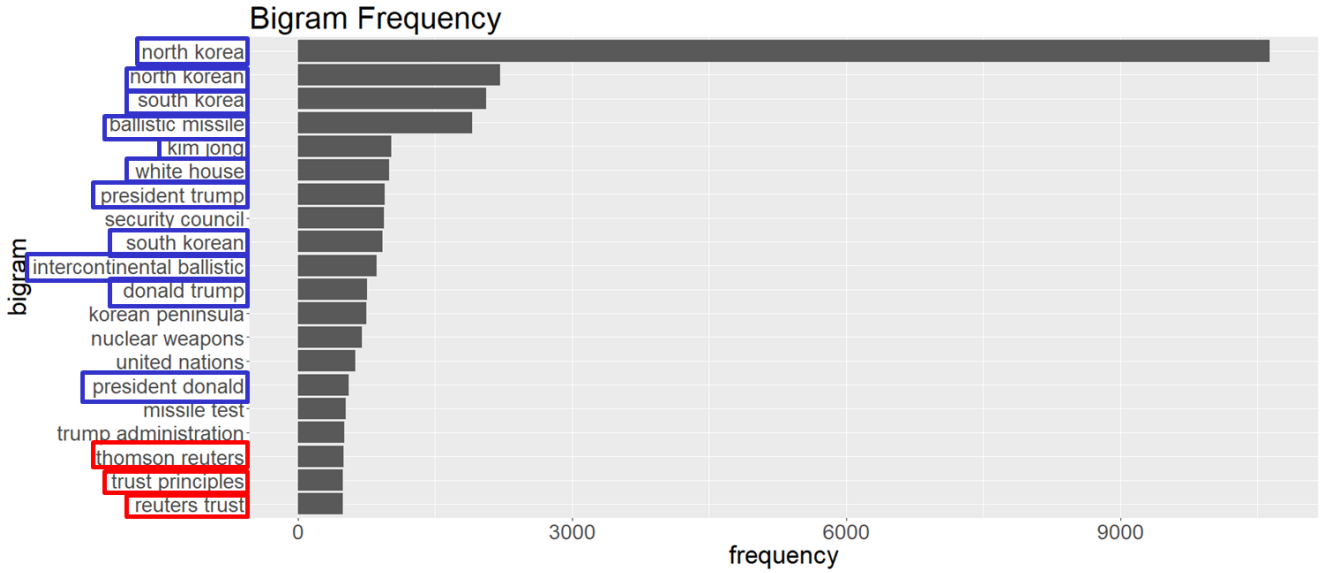


Figure 7. Corpus A.1 Bigram Frequency

This plot provides the account manager with a broad view of the context of the articles. This plot could be used for multiple purposes, two of which could be to identify the frequency distribution of these bigrams and make the determination of which words should be combined and/or deleted. Here, the first nine terms surrounded by blue rectangles designate bigrams that require merging. The following four terms surrounded by red rectangles indicate terms that have no meaning or relevance to the context of the articles within the corpus. While the application of n-gram analysis, for this illustration,

was conducted at the bigram level, the account manager maintained the ability to extend the analysis up to the five-gram level. To ensure that any changes made to Corpus A.1 are applied to the remaining sub-corpuses, he could choose to apply the 'merge terms' function in R to Corpus A and update all sub-corpuses. Table 11 identifies some of the terms that the journalist may have manipulated along the course of their research.

Table 11: Corpus A.1 Summary of Merged/Deleted Terms

Replacement Term	Term
PDJT	u.s president donald trump president donald trump president trump president donald trump
PBO	president barack obama barack obama
CPXJ	chinese president xi jinping president xi jinping jinping
White_House	white house
NKJong	north korean leader kim jong kim jong
ICBM	intercontinental ballistic missile icbm intercontinental ballistic missile
ballistic_missile	ballistic missile
S_Korea	south korean south korea
N_Korea	north korean north korea
100_days	100 days
USS_CV	uss carl vinson
DELETED	thomson reuters trust principles york times reuters fox news

Upon updating Corpus A.1, the account manager could now re-create the results of the n-gram analysis. Figure 8 is the visual representation of the bigram analysis following the updates that were applied.

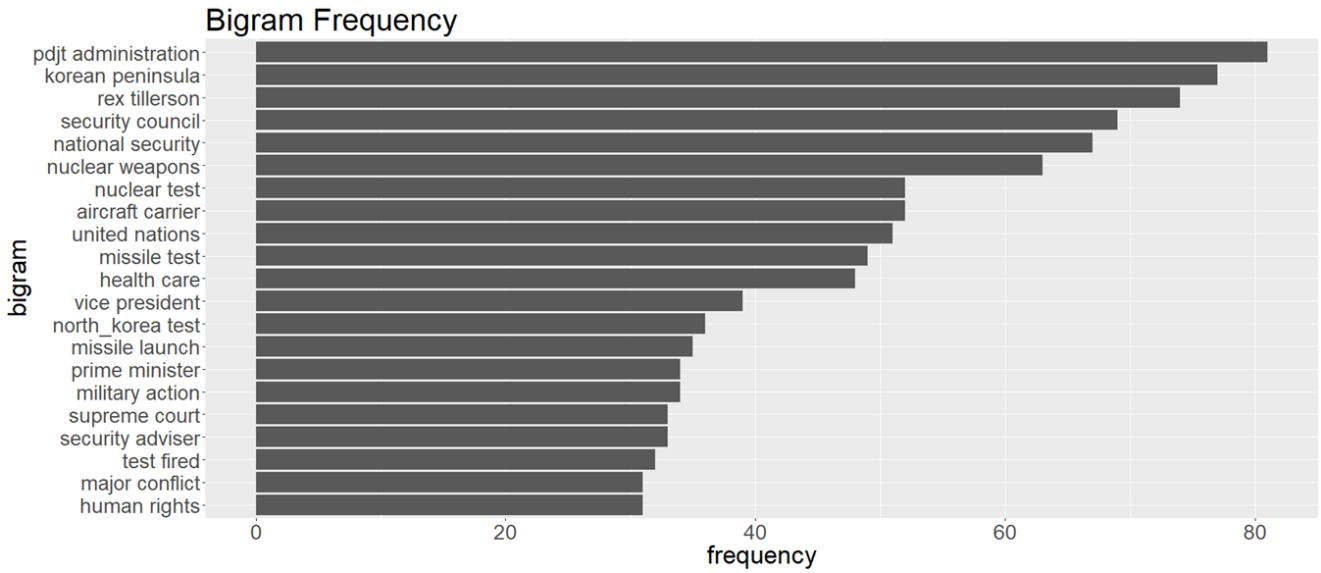


Figure 8. Corpus A.1 Bigram Frequency

Updates made to Corpus A, and its sub-corpuses, can be seen throughout this figure with terms such as "pdjt" and "north_korea."

This cleaned dataset can provide the account manager with a more accurate account of the relevant information contained within the dataset. To uncover more of the information held within this corpus, he may decide to examine the bigram network, Figure 9.

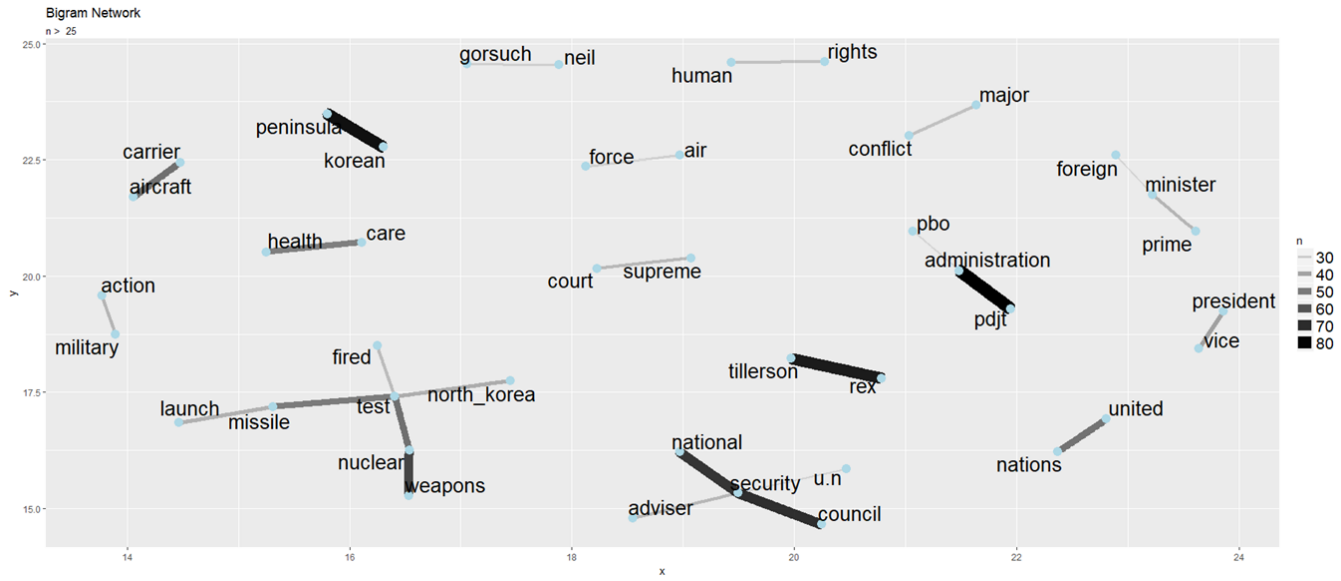


Figure 9. Corpus A.1 Bigram Network

This plot indicates that the articles within this sub-corpus contain information concerning the testing of a North Korean nuclear missile. Output would also appear to suggest the possibility the munition was near the coast of the Korean Peninsula and that it's testing may have had the potential to spark a major conflict. Additionally, it appears that there was international interest in the test with the inclusion of the term "United Nations".

Another method of information discovery the account manager could use is the correlation network. Figure 10, provides further insight into the events that occurred during this time.

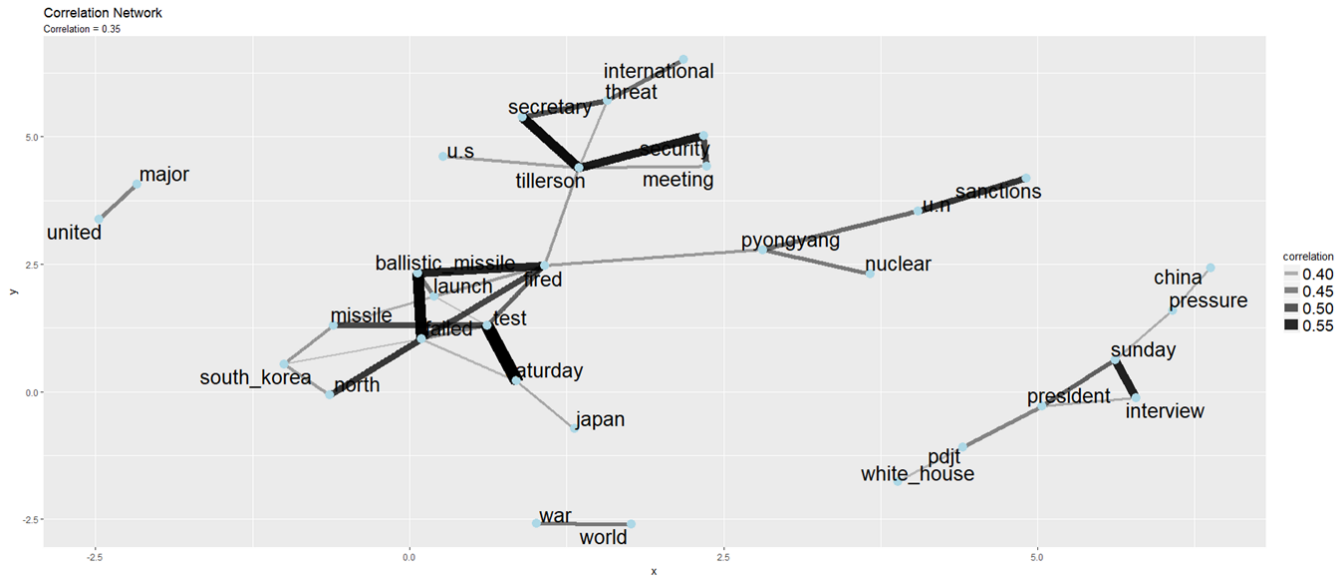


Figure 10. Corpus A.1 Correlation Network

These results indicate that North Korea performed a nuclear ballistic missile test on Saturday April 29, 2017, but the test failed. Further, it would appear, this launch was considered an international threat by United States Secretary of State Rex Tillerson and had potential implications to the term "world war".

4.5.1.2 Corpus A.2: 11 May 2017 – 17 May 2017

The account manager may advance their study by focusing on the analysis of Corpus A.2. Since the text in all corpuses have been updated with Corpus A.1, he may favor proceeding with the examination of the bigram-network plot, Figure 11.

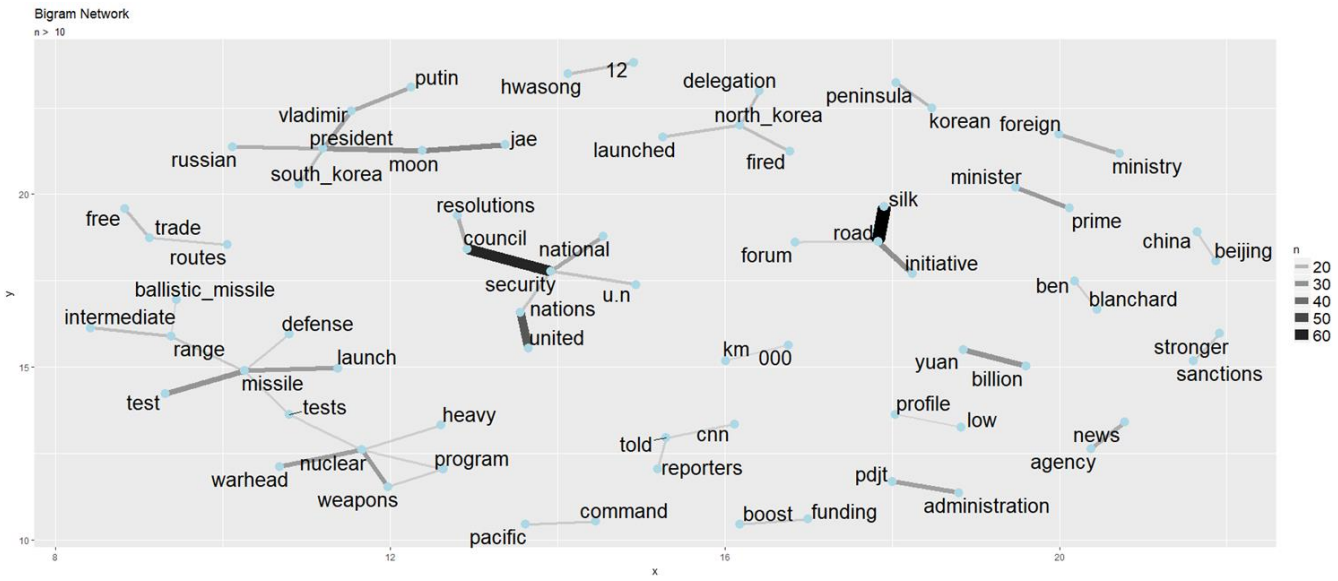


Figure 11. Corpus A.2 Bigram Network

Results from this plot suggest the conduct of another intermediate range nuclear ballistic missile test that now has the interest of the United Nation's Security Council. Additionally, there was no indication that this missile test failed as was the case with Corpus A.1. It was also noted that Russian President Vladimir Putin and South Korean President Moon Jae were mentioned frequently. The account manager may note that this analysis also uncovered topics that may not have any direct relevance to North Korea's ballistic missile development. Words such as "silk", "road", and "free trade routes" may be more focused on another subject that may have indirect ties to the ballistic missile development topic.

He may find these terms interesting, but may ultimately decide that they should be removed from the current analysis for noise reduction purposes and investigated later for future research. To remove articles focused on terms such as "silk", a method the account

manager may choose is to create another sub-corpus from Corpus A.2 based on the keyword "silk" Articles shared between Corpus A.2 and the "silk" Corpus would be identified and removed from Corpus A.2. Having further cleaned the dataset, the journalist can now re-examine the bi-gram network, Figure 12. Following analysis of the plot it can be noted that the output is altered significantly, yet the term "Hwasong-12" continues to be present in the analysis.

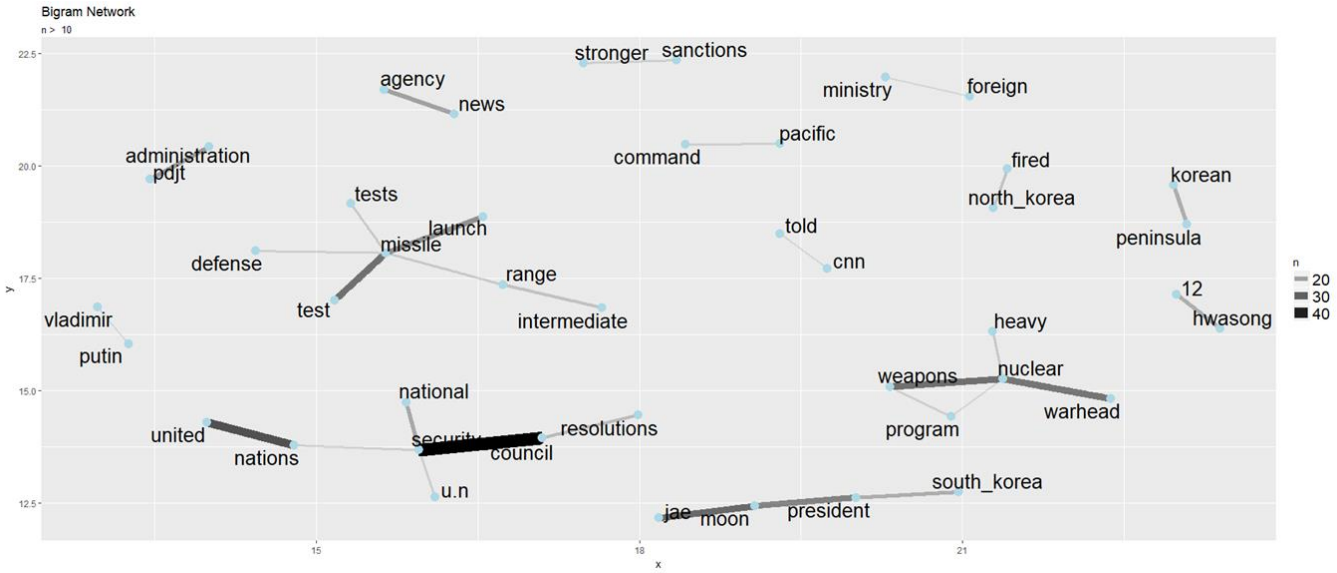


Figure 12. Corpus A.2 Bigram Network

As a secondary exploratory avenue, the account manager can examine the correlation network analysis, Figure 13.

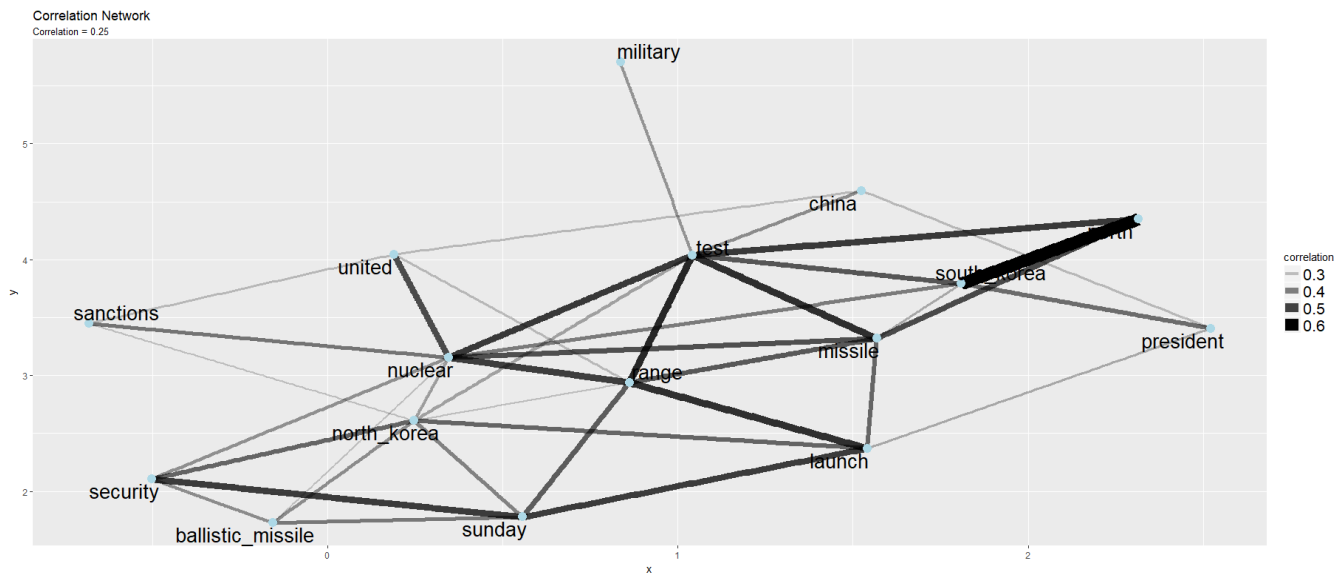


Figure 13. Corpus A.2 Correlation Network

This output indicates that sanctions were being discussed in response to this event. Further, the calendar day "Sunday" was identified as being significant to the analysis, indicating that the event may have occurred on 14 May 2017.

Returning to an early section of this analysis, the account manager may have found the unique term "Hwasong-12" to be of interest. To identify more information on this word it may be elected to perform a term association analysis on the word "Hwasong", Figure 14.

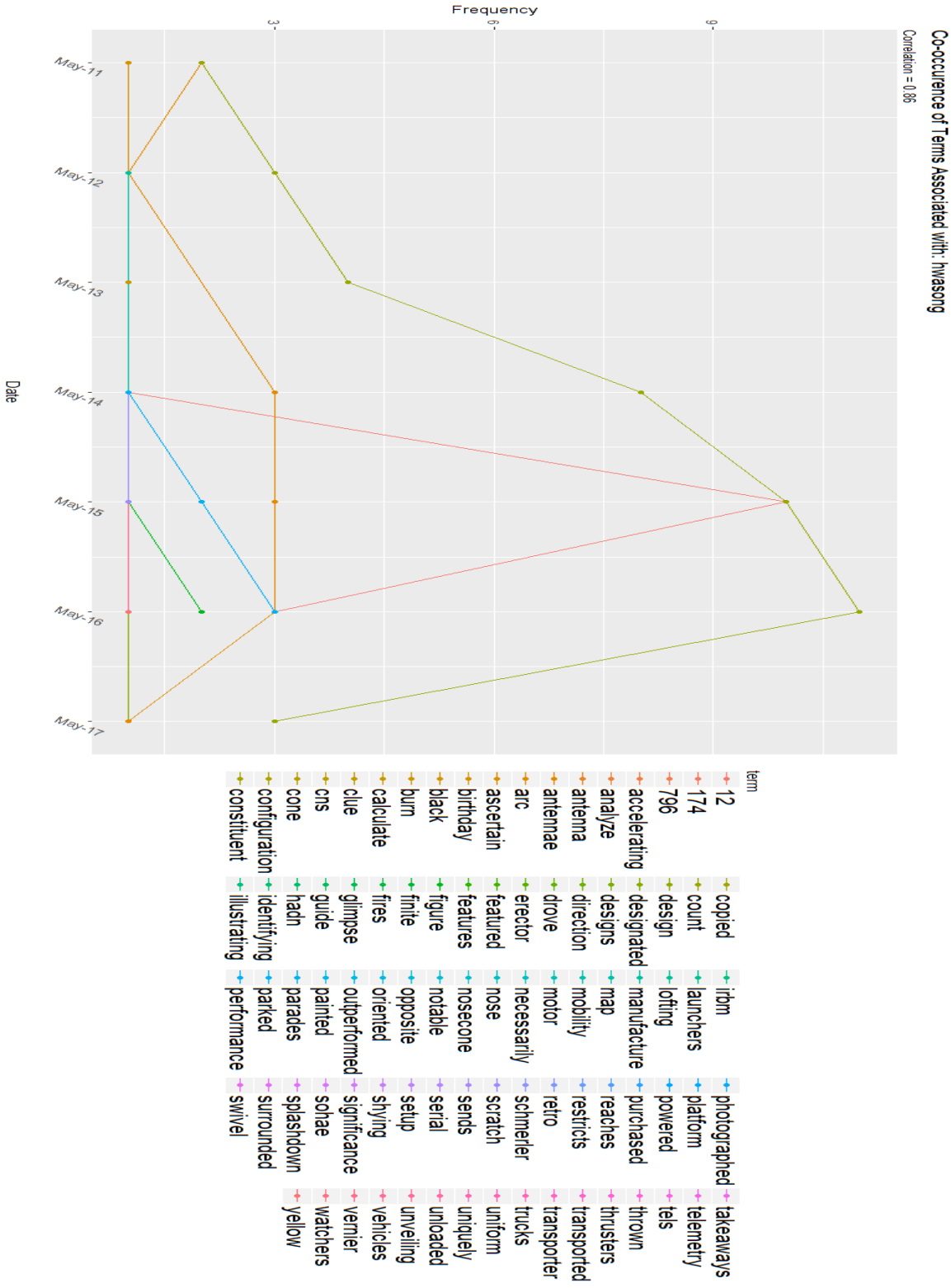


Figure 14. Corpus A.2 Term Association: Hwasong

Analyzing this output, some terms found to be highly correlated with "Hwasong" include "vehicles", "design", "direction", "nosecone", "motor", and "engine", terms often used when describing weapon systems such as ballistic missiles. Applying the results of this analysis the account manager may infer "Hwasong-12" is the name of the ballistic missile that was tested.

Uncovering these pieces of information and evaluating them in concert the account manager may infer that on May 14, 2017 North Korea successfully tested the Hwasong-12 ballistic missile.

4.5.1.3 Corpus A.4: 11 June 2017 – 15 June 2017

Examining a later portion of his analysis, the account manager may extend the analysis into the analysis of Corpus A.4. Extending the investigation into a bigram-network plot the account manager would produce the following, Figure 15.

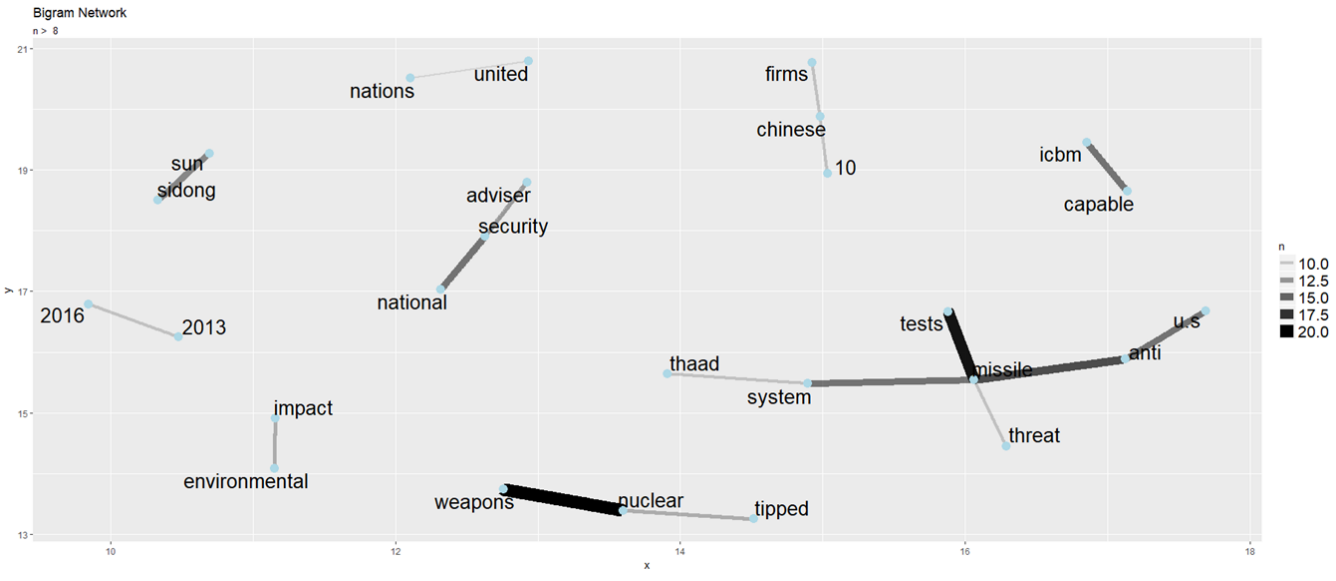


Figure 15. Corpus A.4 Bigram Network

Analyzing this network, he may find that the output does not allude to any new missile testing between these date ranges. Furthermore, it would appear that most news coverage during this period, discussed the implementation of the Terminal High Altitude Area Defense (THAAD) anti-ballistic missile system. The results were similar analyzing n-gram analysis outputs as well as correlation analysis.

4.5.1.4 Inferenced Information Summary

Completing his exploratory analysis into each of the eight sub-corpuses identified, the account manager may want to create a summary of the events in which he believes to have taken place between the dates of 18 April and 31 August 2017. This summary may consist of the following inferences:

Corpus A.1

North Korea performed a nuclear ballistic missile test on Saturday April 29, 2017, with the result of the test being a failure.

Corpus A.2

North Korea successfully conducted testing of the Hwasong-12 nuclear ballistic missile on Sunday May 14, 2017.

Corpus A.3

North Korea successfully conducted testing of an intermediate range ballistic missile on Monday, May 29, 2017.

Corpus A.4

No missile tests were conducted between 11-15 June 2017. Most news appeared to focus on the implementation of the THAAD anti-ballistic missile system.

Corpus A.5

North Korea successfully conducted testing of its intermediate range InterContinental Ballistic Missile (ICBM), the Hwasong-14 on Tuesday July 04, 2017.

Corpus A.6

North Korea successfully conducted testing of a variant of the Hwasong-14 ballistic missile system on Friday July 28, 2017.

Corpus A.7

No missile test occurred during the period of 05-19 August 2017. Topics covered here appear to cover topics concerned with economic sanctions being implemented by the UN Security Council and U.S. military preparation in response to pervious demonstrations of North Korea's nuclear missile capability.

Corpus A.8

North Korea successfully conducted testing of the Hwasong-12 ballistic missile system on Tuesday August 29, 2017.

4.5.1.5 Data Validation

The Center for Strategic and International Studies' (CSIS) Missile Defense Project is dedicated to providing an authoritative analysis on missile defense as well as

other related issues [57]. The Advisory Board for this project is comprised of a host former government officers, military officials, scholars, and business leaders who meet periodically to update and discuss the direction of the project. The Missile Defense Project focuses on the analysis of a wide variety of policy, program, and strategic issues related to missile defense and operates the Missile Threat website, an open source initiative related to the proliferation of cruise and ballistic missiles around the world. This database provides detailed information missile types, tests, and the countries in possession of such munitions. An infographic found on the Missile Threat website, Figure 16 , provides a summary of the proliferation of North Korean ballistic missile testing between 1984 and 2017.

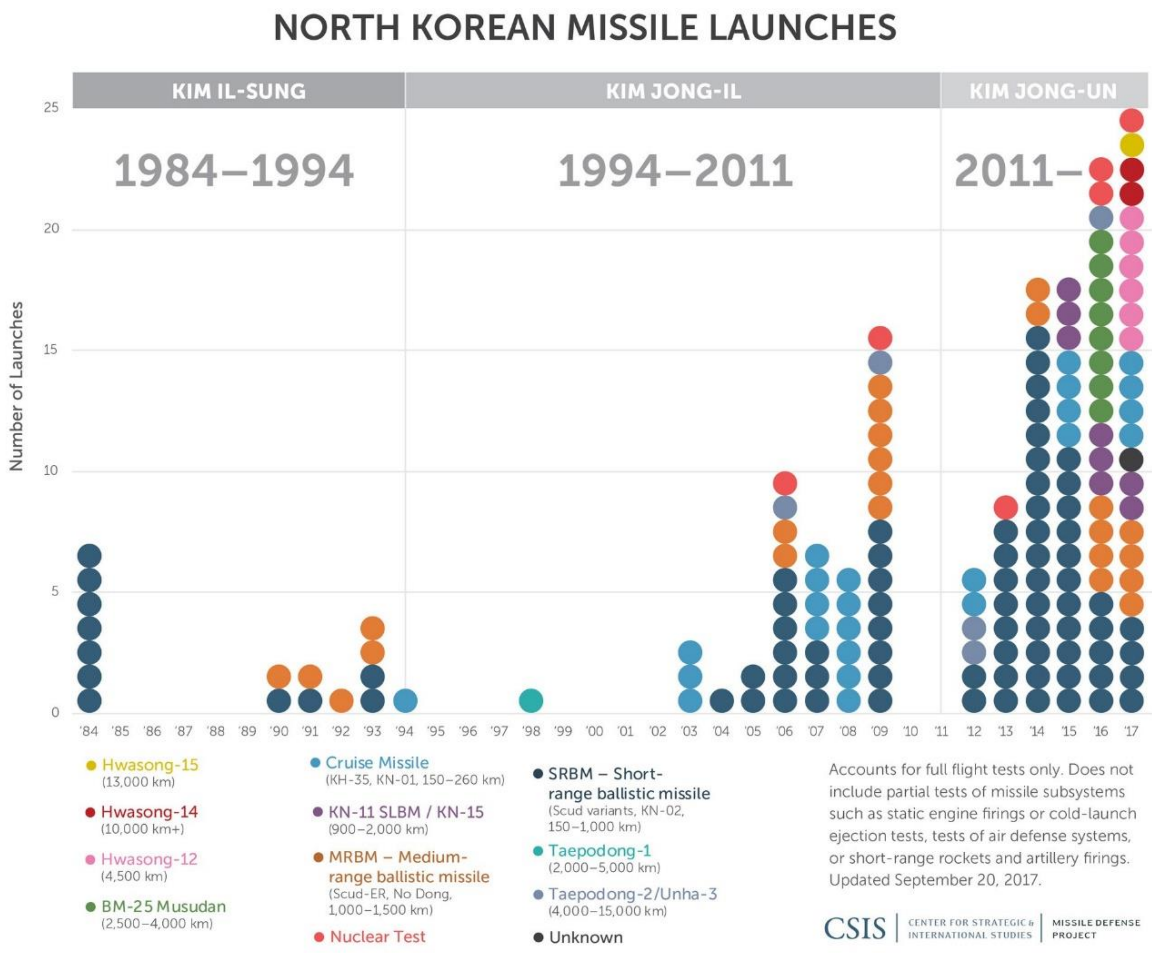


Figure 16. North Korean Missile Launches 1984-2017 [58]

Additionally, the database provides a table detailing each North Korean ballistic missile test which occurred during 2017. Table 12, produced by CSIS, identifies the date of the missile test, the type of missile tested, and the number of munitions released during the test. This table represents test that occurred between the date range of 18 April and 31 August 2017.

Table 12. North Korean Ballistic Missile Test Summary (18 April - 31 August 2017) [58]

Date	Missile Type	Number Launched
August-28-2017	Hwasong-12 (IRBM)	1
August-26-2017	KN-21 Scud variant (SRBM)	3
July-28-2017	Hwasong-14 (ICBM)	1
July-4-2017	Hwasong-14 (ICBM)	1
June-7-2017	Kumsong-3 (ASCM)	4
May-28-2017	KN-18 MaRV Scud-variant (SRBM)	1
May-21-2017	KN-15 (MRBM)	1
May-14-2017	Hwasong-12 (IRBM)	1
Apr-29-2017	Hwasong-12 (IRBM)	1

Applying the information discovered through the application of exploratory analysis through R, the following timeline, Figure 17, was constructed.

Figure 17. Estimated Timeline of North Korean Ballistic Missile Development

Figure 18 compares the results of the inferences made using the R package to the information provided by the CSIS Missile Defense Project. Check marked green boxes represent the occurrence of a North Korean ballistic-missile test and the identification of said test by the CSIS Missile Defense Project and/or with a corpus exploration package in R.

	April 28/29	May 14	May 21	May 28/29	Jun 7	Jun 11-15	Jul 4	Jul 28	Aug 5-17	Aug 26	Aug 28
Missile Test	Hwasong-12 (Fail)	Hwasong-12	KN-15	KN-18	Kumsong-3	No Test	Hwasong-14	Hwasong-14	No Test	KN-21	Hwasong-12
CSIS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CorpEx	✓	✓	X	✓	X	✓	✓	✓	✓	X	✓

Figure 18. Estimated vs Accurate Event Occurrences

Using R, the account manager would have effectively been able to identify six of nine ballistic missile tests, and two instances where there was a large circulation of news dealing with North Korea, but no missile tests. This would result in the account manager

correctly identifying North Korean ballistic missile tests conducted between 18 April and 31 August 2017 with an accuracy of ~73%.

4.5.2 Undirected Search: Silk Road Initiative

This thesis now considers this secondary conceptual scenario: the account manager from Case 1, along with being tasked to become knowledgeable on North Korean ballistic testing in recent accounts, was also tasked with finding an area his firm could re-focus their financial efforts. The account manager decides to revisit some of the interesting terms that emerged during analysis of Corpus A.2. These terms included "silk", "road", and "free trade". Still in possession of the storage device (Corpus 1), the account manager may want to begin this exploration by separating articles containing the term "silk". Focusing Corpus 1 on the word "silk", he could examine a line chart of articles containing this term, Figure 19.

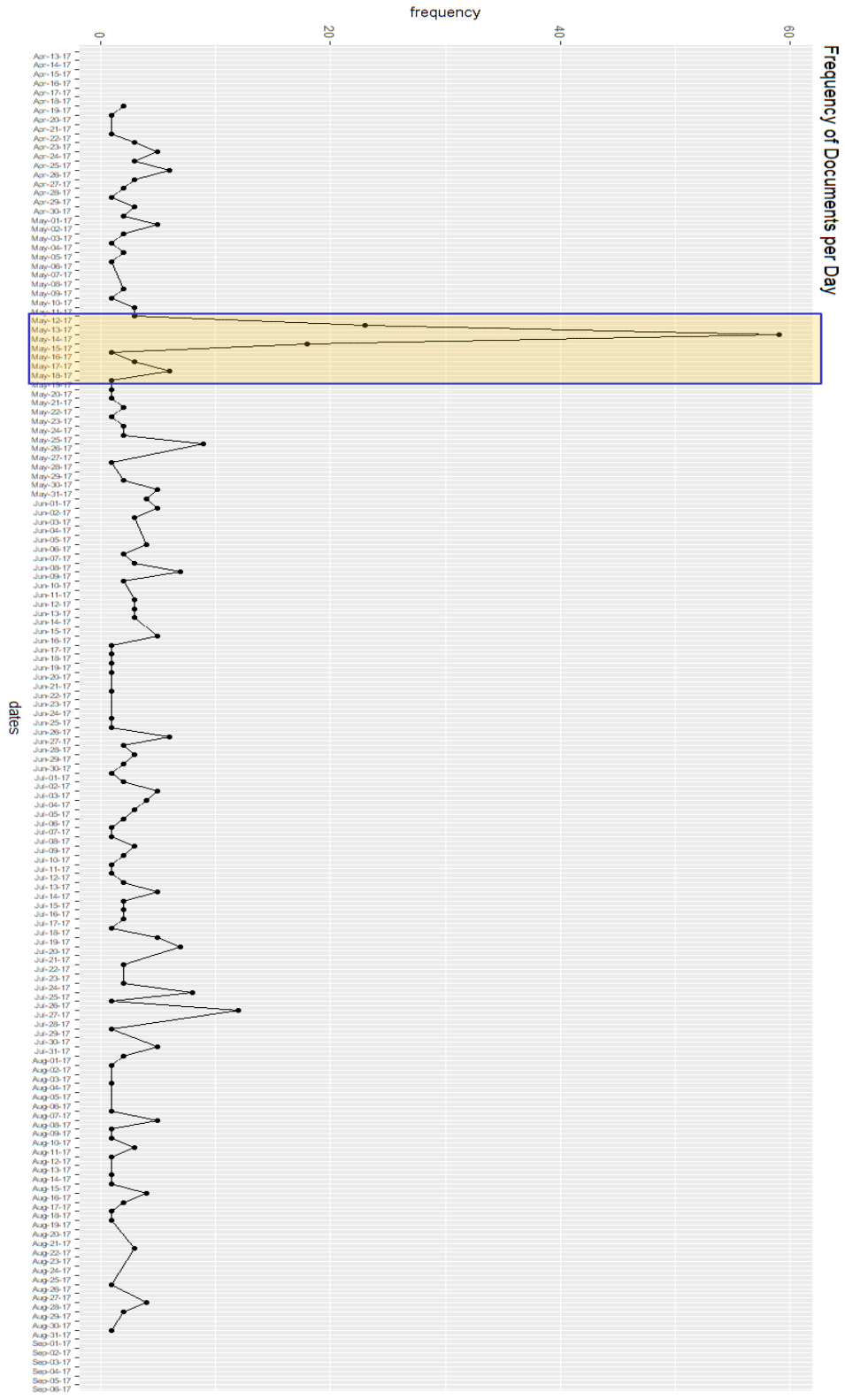


Figure 19. "Silk" Time Series Document Frequency Plot

The highlighted date range (12 May to 19 May 2017) denotes an increase in the frequency of articles produced that the account manager may find interesting. At this point he may elect to only select articles within this date range for his exploration.

This reduced corpus, articles containing the term "silk" within the date range of 12-19 May 2017, hereinafter be referred to as Corpus B, contains 373 articles and 323,359 terms and will be the focus of Case 2.

4.5.2.1 Corpus B

A logical progression for this account manager's analysis would be to perform an examination of n-grams. Figure 20 and Figure 21 illustrate the bigram and trigram frequency plots respectively.

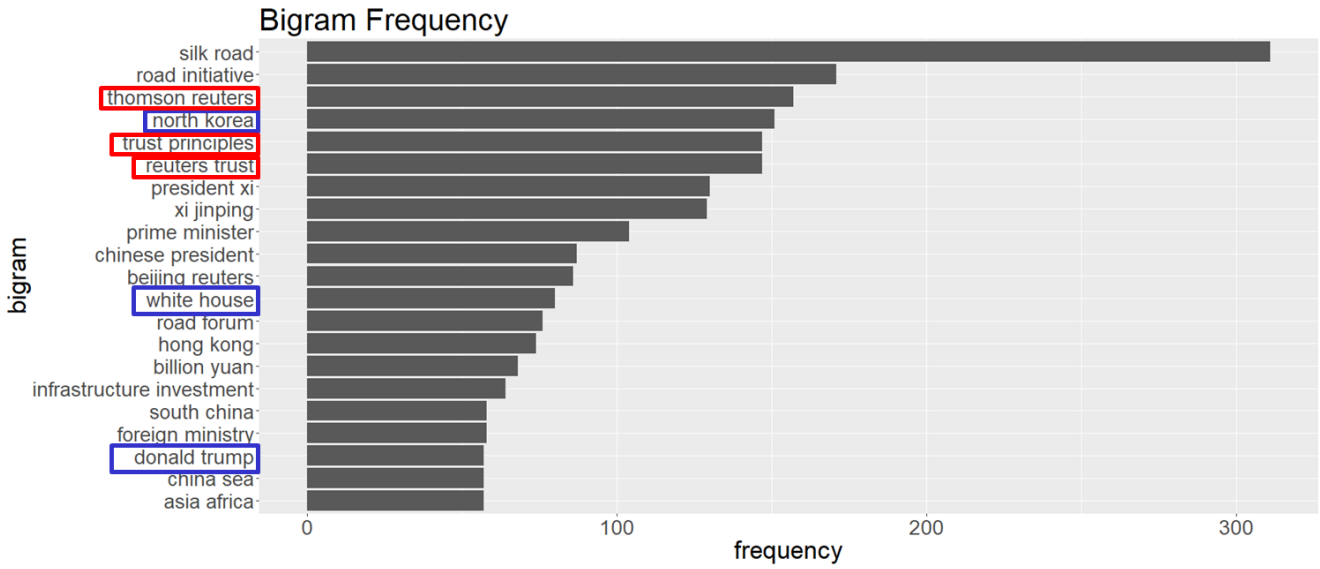


Figure 20. Corpus B Bigram Frequency

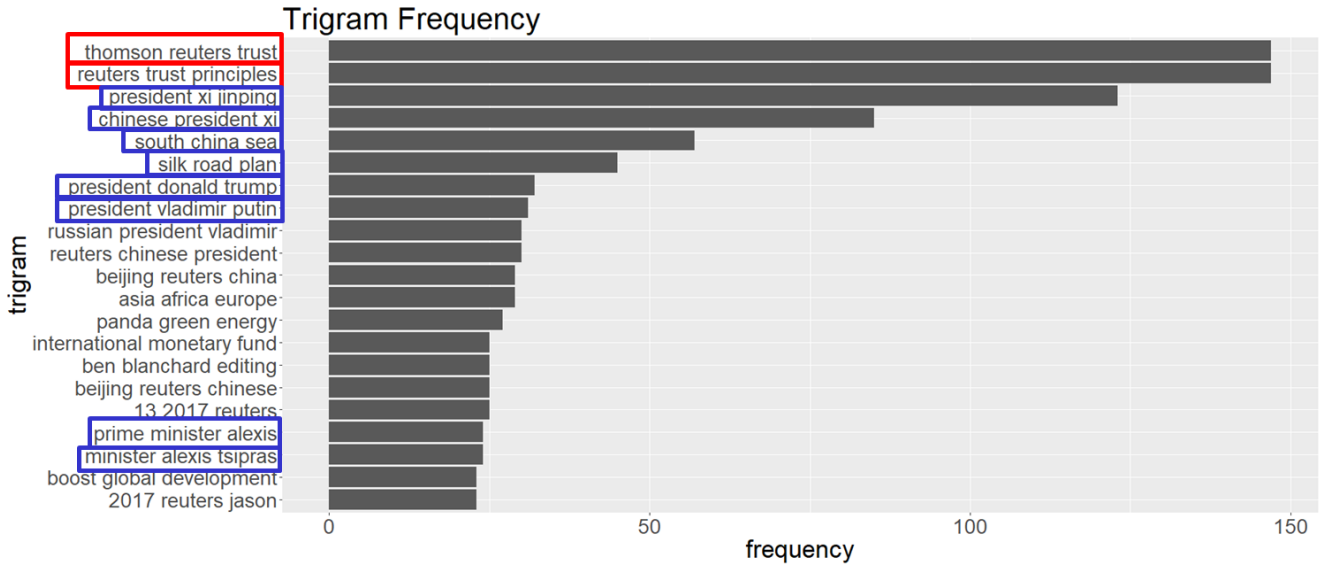


Figure 21. Corpus B Trigram Frequency

Similar to Case Study 1, this n-gram analysis serves to indicate terms that offer no relevant information to the analysis, terms surrounded by a red rectangle, and terms that should be intuitively merged, terms surrounded by a blue rectangle. At this point, the account manager may elect to implement a series of merge/delete functions to clean the data set.

Upon cleaning this corpus, the journalist may reexamine Corpus B through the investigation of a bigram network analysis, Figure 22.

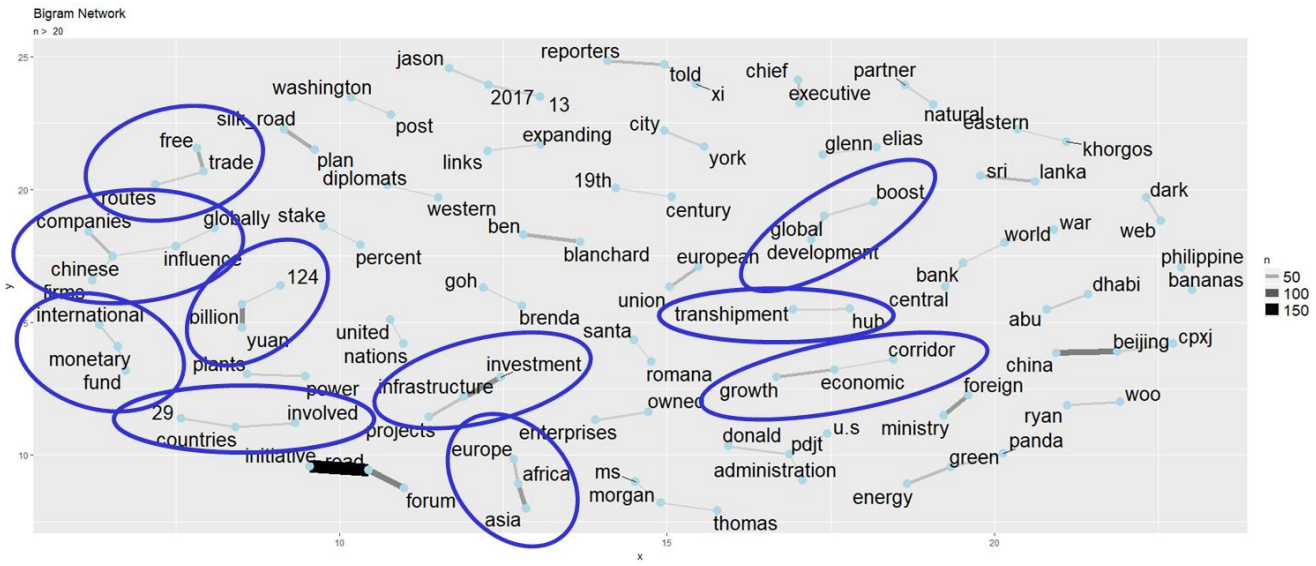


Figure 22. Re-Examination of Corpus B Bigram Network

This plot uncovers some very insightful information, circled in blue. The account manager could make the inference that Corpus B, centered on the term "silk", is dealing with some actions being taken by China that may expand their influence globally. There are indications that there is a substantial investment (124-billion-yuan) in various infrastructure projects that will ultimately lead to a boost in global economic growth. Chinese global influence may be spread across Asia, Africa, and Europe through the use of transshipment hubs. Additionally, this event may affect the global development of free trade, and it appears that 29 countries are involved in this transaction.

To further analyze connections made throughout Corpus B, he may next explore the correlation network, Figure 23.

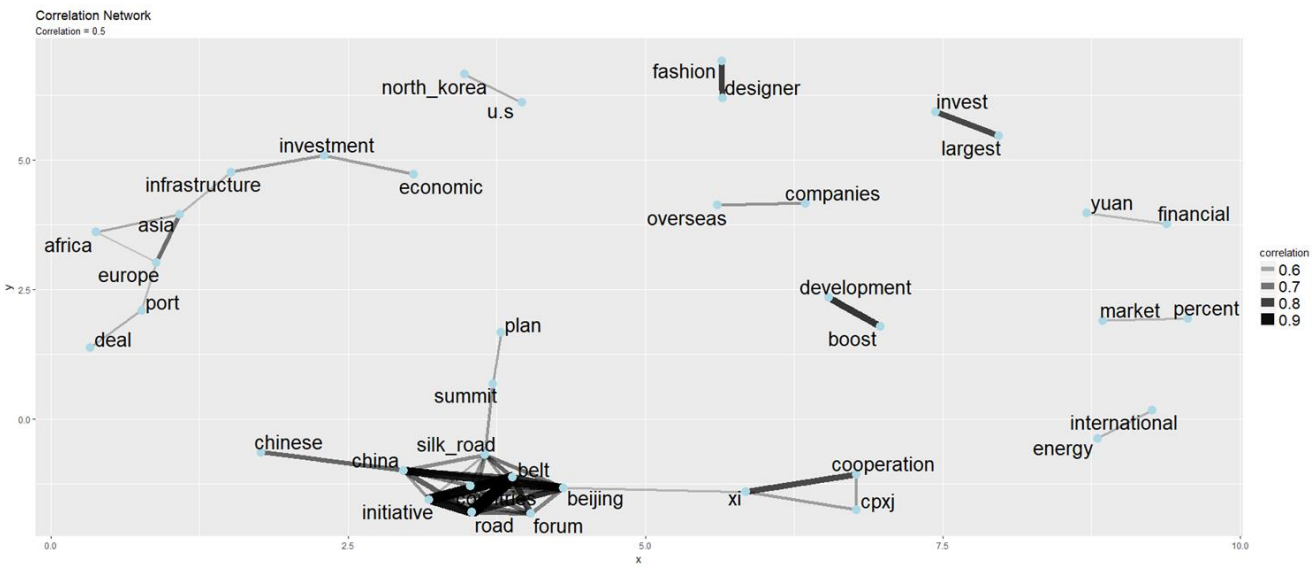


Figure 23. Corpus B Correlation Network

Figure 23 further solidifies the information presented in the previous plot, speaking to the notion that there is an investment in infrastructure dealing with Asia and European ports. This plot also provides additional information on a "silk road summit" that was held, likely involving multiple countries.

The account manager may deem it necessary to have a deeper understanding of the information surrounding the "silk road" and seek to uncover hidden topics found within the corpus. To perform the application of topic modeling, it is mandatory that the account manager first determine the appropriate number of topics to create. Using the topic modeling functionality of R and the "silk" corpus as an input into the *topic number* feature, he is presented with an output denoting possible topic number options, Figure 24.

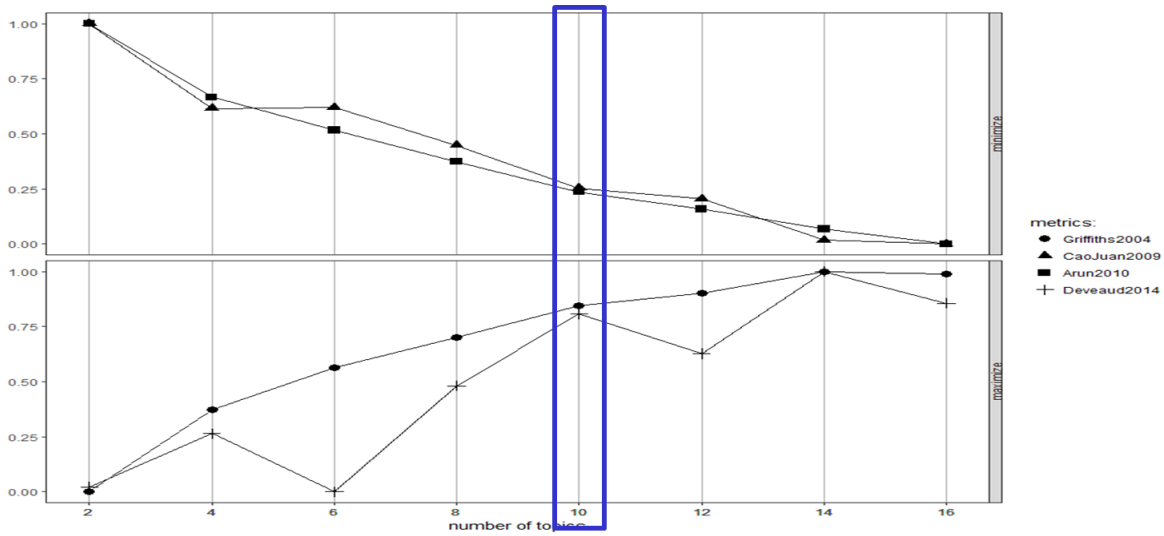


Figure 24. Topic Number Analysis Output

CaoJuan2009 and Arun2010 are minimization metrics that indicate the appropriate number of topics as those reaching the normalized value of zero. Griffiths2004 and Deveaud2014 are maximization metrics seeking to identify the optimal number of topics based on reaching a normalized value of one. The account manager notices that all four metrics approach their maximum (minimum) values near 16 topics. Making the business case that the value of 16 maybe too large for such a small dataset, the account manager may elect to select only 10 topics for their continued analysis.

Applying 10 topics to the topic modeling algorithm, he would achieve the following output, Figure 25.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	said	silkroad	pdjt	china	asia	belt	percent	will	china	greece
2	nkorea	new	president	beijing	bank	road	said	said	said	new
3	forum	billion	one	chinese	projects	trade	growth	investment	south	cooperation
4	foreign	china	said	told	new	countries	investment	initiative	sea	infrastructure
5	plan	summit	day	plan	project	initiative	monday	road	two	international
6	missile	yuan	year	government	will	beijing	economy	belt	philippines	primeminister
7	summit	development	can	editing	initiative	said	financial	railway	duterte	deal
8	countries	countries	first	standards	economic	britain	year	company	sovereignty	ministry
9	united	said	people	reporting	world	two	global	billion	vietnam	saying
10	delegation	also	like	xinjiang	pakistan	open	cpxj	port	last	state

Figure 25. Topic Models

This output is a visual representation of the top 10 words associated with 10 user specified topics. Upon achieving this output, it is at the determination of the user to infer what each topic is related to. Table 13 are some potential inferences that the account manager may have made upon examination of Figure 25.

Table 13. Topic Inferences

Topic	Inference
1	North Korean missile forum
2	Summit on new silk road development
3	President Donald J. Trump's response to an issue
4	Chinese government commenting on a plan
5	World economic projections on initiative
6	Trade opening between Beijing and Britain
7	Global economies financial growth
8	Cost of investment of new railways
9	Policy issues with china and Vietnam
10	New cooperation with Greece due to infrastructure

The account manager may find the inference on Topic 2, "Summit on new silk road development" to be of interest. Sub-setting the articles linked to Topic 2, he now can investigate this topic further. Figure 26 is the visualization of the bigram network analysis of Topic 2.

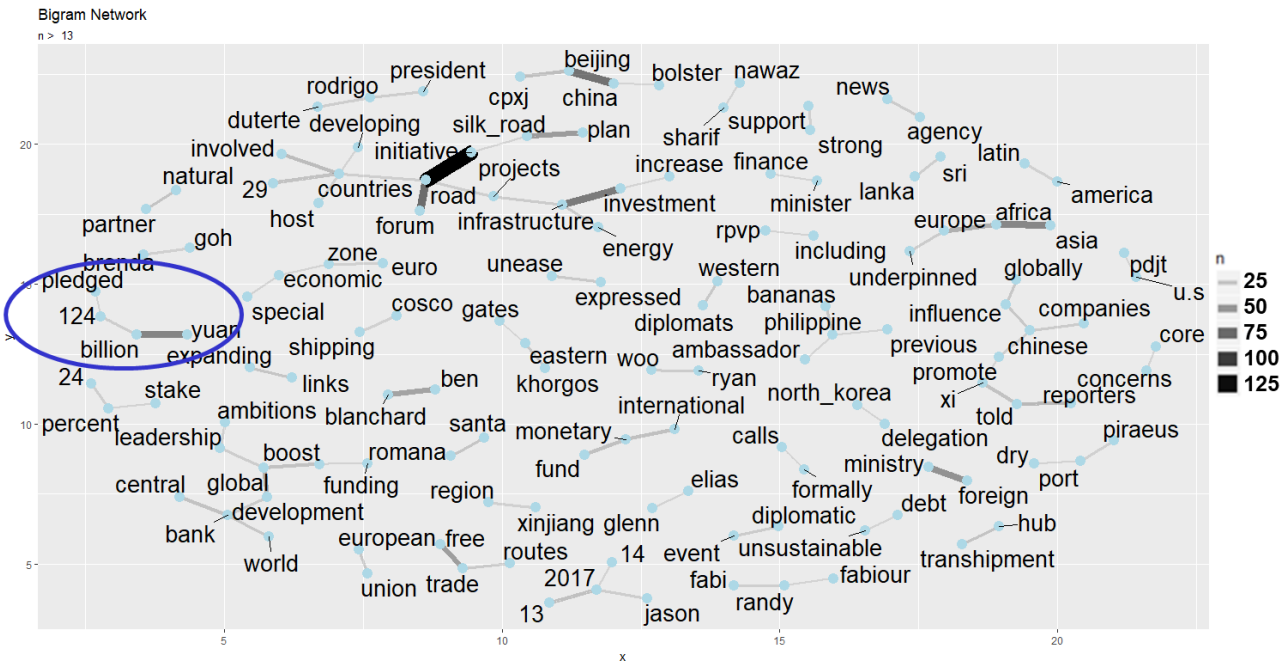


Figure 26. Topic 2: Bigram Network

It can be noted that most of the information retrieved from this bigram network analysis confirms the information uncovered previously. It can be further noted that there was more detailed information in this plot than in previous plots. Analysis of this output, the area circled above, indicates that Chinese President Xi Jinping pledged 124 billion yuan to the support of the silk road initiative. While there is not enough detail present here, it can also be inferred that this initiative will have an influence that affects things on the level of billions.

4.5.2.2 Data Inference

From the analysis of this corpus, the account manager may be able to infer that there is an economic initiative in China known as the Silk Road Plan which may have some of the following properties:

- Potential free trade initiative which appears to involve world trade and boosting global development
- Project likely requires the creation of transportation hubs which connect trade routes in China to Africa, Europe, and other Asian countries
- 29 Countries maybe involved in the form of a summit
- Expected 124-billion-yuan investment

4.5.2.3 Data Validation

China's One Belt, One Road (OBOR) initiative, illustrated in Figure 27, is a strategy proposed by Chinese President Xi Jinping in 2013 [59]

Figure 27. China's One Belt One Road Map [60]

The trade initiative seeks to develop the "Silk Road Economic Belt", the land component of OBOR which would increase connectivity between China, Central and South Asia, the Middle East and Europe. Additionally, the OBOR strategy calls for the development of the 21st Century Maritime Silk Road, the oceanic component of the trade route which would link Southeast Asia to the Middle East, Africa, and Europe [61]. China has embarked on this momentous endeavor with the goal of improving trade relationships, primarily through investments in infrastructure. This lofty and lengthy goal will ultimately call for a significant amount of cooperation and funding between many countries. In the opening speech of the "Belt and Road" Summit Luncheon, delivered by Chairman Jin Qi on May 18, 2016, he described the goal of the initiative as the creation of a community of mutual interests, economic integration, and cultural tolerance. During his speech he also mentions that about 30 countries have already come together in support of the effort, through the signing of memorandums of understanding [62]. China established the Silk Road Fund in 2014, providing US $40 billion to finance various initiatives linked to this project [63]. Additionally, established in 2015 to compete with United States financial organizations, The Asian Infrastructure Investment Bank (AIIB), has reportedly been provided US $65 billion in initial capital to support these investments [64]. As of February 2016, at least US $250 billion in project investments in various stages has said to have been tracked with the prediction that this initiative will receive up to US $1 trillion of outbound state financing in the next 10 years [65].

Comparing the findings identified through the account manager's probe of Corpus 3 to what publicly available information exists the results may be summarized as the following, Table 14.

Table 14. Silk Road Inference Accuracy

Data Inference	Data Validation	Confirmed Information
Potential free trade initiative which appears to involve world trade and boosting global development	·Connections to Southeast Asia to the Middle East, Africa, and Europe	✓
Project likely requires the creation of transportation hubs which connect trade routes in China to Africa, Europe, and other Asian countries	·Silk Road Economic Belt ·21st Century Maritime Silk Road	✓
29 Countries maybe involved in the form of a summit	·About 30 countries signed memorandum of agreement	✓
Expected 124-billion-yuan investment	·Silk Road Fund (2014) – US$40 billion ·AIIB (2015) – US$65 billion ·Estimated (2016) – US$250 billion ·Expected (10 year outlook) – US$1 trillion	✓
	Accuracy	100%

V. Conclusion

5.1 Results

The methodology developed in conjunction with this thesis research allows users to effectively sieve through a large corpus of textual documentation quickly and accurately to identify information contained within. Results for the temporal analysis of North Korean ballistic missile proliferation between 18 April and 31 August of 2017 provided a 73% accuracy, identifying no false munition tests. Additionally, results obtained through open exploration into the corpus identified detailed information concerning China's One Belt One Road initiative. Each of the inferences made through the exploration were able to be confirmed through investigation of open source, publicly available information.

5.2 Research Conclusion

The explosive growth that the digital universe has realized with unstructured data is expected to continue clear into the future. This anticipated increase in the volume of unstructured data has the potential to overburden the analyst needing to explore datasets to gain useful information. Additionally, this threat expands to the organizational level, requiring the organization to support time and resource intensive measures to ensure their analysts have the ability to provide actionable and accurate intelligence. Applying various text mining techniques, this research developed a framework for the exploration of large

corpuses of textual documents. Ultimately, the work developed throughout this research can be used to significantly reduce the time required of analysts to maneuver through large corpuses of text data uncovering insightful and dependable information.

5.3 Future Research

The methodology developed for the purposes of this research was intended to establish the foundational framework for a potentially very robust text exploration tool to be used in the open source R environment. While the current methodology enhances the user's ability to identify and separate documents with a corpus based on user specified characteristics, transform the content of the corpus to increase analytical efficiency, and allows the user to perform a host of text mining techniques to analyze this form of data, there are further avenues in which future research can add value. Potential areas include the development of an information extraction tool, development of a method to allow the user to track and undo changes made during the UDM phase, adaptation of the topic modeling framework to analyze how correlated multiple source's depiction of the same event are, development of an anomaly detection system for the frequency of documents produced as well as the use of user identified terms, and an improvement to the functionality of the time series metrics to account for rate of change with time series data.

The addition of an information extraction tool would require the implementation of a part-of-speech (POS) tracker which would identify how each term in a sentence is being used and then subsequently tag each term in the document. This tagged information would give the analyst the ability to apply a NER algorithm to further define recognizable terms within the document. With these terms defined, the analyst can now make

connections using linkages such as person, organization, location, currency, etc., providing a deeper investigation into the analysis.

The UDM phase of this methodology provides the user the ability to alter the content of the corpus to apply organizational knowledge or subject matter expertise directly into the data. Currently, this process requires the user to make inputs into the software identifying the changes to be made. There is currently no method of tracking what changes are made, requiring the user to keep a mental note of how they have progressed through the data. Providing the user, the ability to actively track, save, and undo changes would increase the efficiency of this phase.

Integrating a method to provide the user with the ability to validate the similarity between sources may have very beneficial effects. This concept is tailored towards the analysis of multiple reports focused on a singular subject. In an ideal examination of the sources, all reports should maintain high correlation in the information presented. Should one of the sources indicate information different than the others, this report would then be identified as an anomaly. This anomaly represents a deviation in the accuracy of the information with two potential outcomes being that this source is providing inaccurate information, or this source is providing accurate information and the other sources are in fact incorrect. Analysis at this level could provide valuable insight into the validity of an array of text documents ranging from military field reports to competitor market analysis.

In its current stage, this methodology allows the analyst to detect fluctuations in the frequency of documents produced through the examination of document frequency plots. Should a sudden increase in the number of documents produced occur, the user

would be informed of this shift through the examination of the sudden upward trajectory in the time series visualization. While this method of abstracting these fluctuations sufficiently identifies useful information, as demonstrated in the Directed Search case scenario, the development of an anomaly detection system may prove useful in identifying statistically significant shifts in the frequency of documents produced. The development of this addition could also be extended to the analysis of how terms are used throughout documents. Here, the user would be provided with an output noting any statistically significant shifts in the usage of specified terms. The user could then track the use of terms as they are used across a corpus.

The final additional research avenue that could provide benefit to this methodology is in the advancement of the functionality for time series data metrics. Currently these metrics and visualizations are based only on the valuation of their frequencies. Allowing the user, the ability to examine this information with the added ability to focus on rates of change, could provide additional insight into the corpus. Development of this feature could greatly benefit this research.

Appendix A: R Packages Used

R Version 3.4.3("Kite-Eating Tree"), was used for this thesis research. Additionally, RStudio version 1.0.143 used. The following is a list of the packages used in this research to perform the methodology used to explore the corpus.

Table 15. R Package Summary

Package	Author	Version	Purpose
dpylr	Wickham et al. [66]	0.7.4	Data Manipulation (Pipe Operation)
lubridate	Spinu et al. [67]	1.7.1	Date Manipulation
stringr	Wickham [68]	1.2.0	Text Manipulation
tm	Feinerer et al. [69]	0.7-3	Text Mining Package
tidytext	Silge et al. [70]	0.1.6	Text Mining Package
ggraph	Pedersen [71]	1.0.0	Grammar Graphics
igraph	Csardi [72]	1.1.2	Network Analysis and Visualization
ggplot2	Wickham et al. [73]	2.2.1	Data Visualizations
topicmodels	Grün et al. [74]	0.2-7	Topic Model Development
widyr	Robinson [75]	0.1.0	Widen, Process, then Re-Tidy Data
reshape2	Wickham [76]	1.4.3	Reshape Data
ldatuning	Murzintcev [77]	0.2.0	Tuning LDA Model Parameters
scales	Wickham et al. [78]	0.5.0	Scales Functions for Visualizations
tidyr	Wickham et al. [79]	0.7.2	Tidy Data Functionality

The Application of Text Mining and Data Visualization Techniques to Textual Corpus Exploration

Captain Jeffrey Smith
Advisor: LTC Christopher Smith, Ph.D.
Committee Member: Bradley Boehmke, Ph.D.
Department of Operational Sciences (ENS)
Air Force Institute of Technology

Sponsor:
United States Army Intelligence and Security Command (INSCOM)

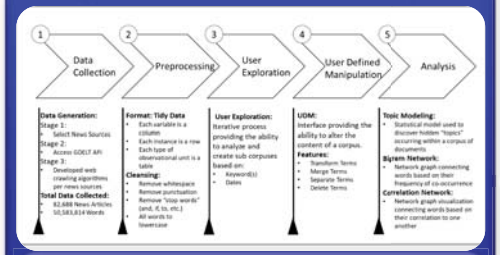
Problem Statement

Unstructured data in the digital universe is growing rapidly and shows no sign of slowing anytime soon. This growth in digital data makes the prospect of information overload a much more prevalent threat to analysts. Utilizing various text mining techniques such as n-gram analysis, document and term frequency analysis, correlation analysis, and topic modeling methodologies, this research developed a tool to allow analysts to maneuver effectively and efficiently through large corpuses of potentially unknown textual data.

Research Questions

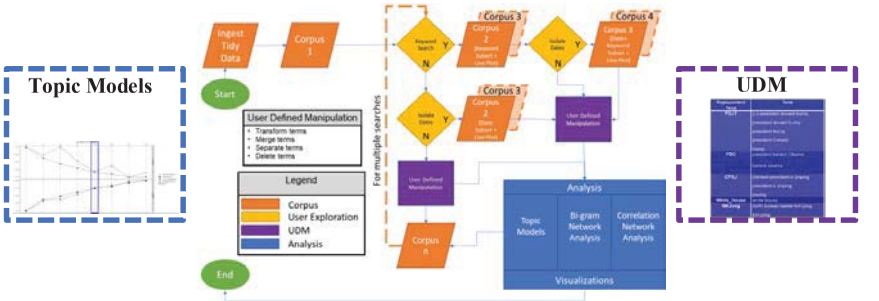
- What methodologies exist to permit analysts to identify and separate documents within a corpus based on user specified characteristics?
- Are there methods for an analyst to alter the content of a corpus for further analysis?
- What data visualization techniques are best leveraged to aid in the discovery of insightful information contained within a corpus of text documents?

Corpus Exploration Pipeline



Methodology

Ex. Visualizations

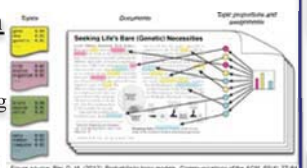


Text Mining Techniques

Latent Dirichlet Allocation

- A topic modeling algorithm that automatically discovers and tags hidden topics existing within documents in a corpus.
- Iteratively calculates the probability of each word (W) belonging to each topic (Z) given words found in each document (D)

$$P(Z|W,D) = \frac{(\#\ of\ word\ W\ in\ topic\ Z) + \beta_w}{(total\ tokens\ in\ Z) + \beta} * (\#\ words\ in\ D\ that\ belong\ to\ Z + \alpha)$$



Term Correlation Analysis

- Exposes relationships between terms that are found in the same document, but may not co-occur such as with n-grams
- Phi coefficient (ϕ) (± 1): Measure of association for two binary variables

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{1\cdot})(n_{0\cdot})(n_{\cdot 0})(n_{\cdot 1})}}$$

	y = 1	y = 0	total
x = 1	n_{11}	n_{10}	$n_{1\cdot}$
x = 0	n_{01}	n_{00}	$n_{0\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 0}$	n

n_{11} - case where both words appear
n_{00} - case where neither word appears
n_{01}, n_{10} - case where either word appears

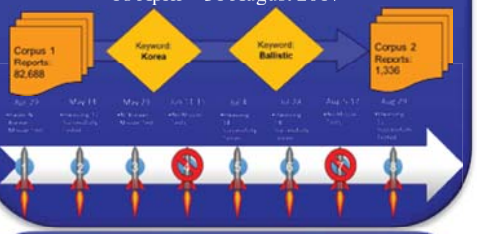
User Defined Manipulation (UDM)

- Alter the content of a corpus based on subject matter expertise or organizational knowledge.
- Replace terms
  - Replace given terms with user defined terms
- Merge/Separate terms
  - Unite multiple terms that represent a single entity
- Delete terms
  - Remove terms from the analysis

Conclusions

- Allowed the user to effectively sieve through a large corpus of text documentation quickly and accurately to identify information contained in the corpus.
- Directed Search:

Identification of North Korean Ballistic Missile Tests between:
18 April – 31 August 2017



Future Work

- Development of an Information Extraction tool
- Allow users to track changes made during the UDM phase
- Adapt topic modeling to track relatedness of single events from multiple sources
- Document frequency anomaly detection functionality
- Application of rate of change functionality to time series metrics

AIR FORCE INSTITUTE OF TECHNOLOGY

Works Cited

[1] Gantz, J., & Reinsel, D. (2013). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east*. IDC iView: IDC Analyze the Future (2012), 1–16. Retrieved from https://www.emc.com/collateral/analyst-reports/idc-digital-universe-united-states.pdf

[2] Nahm, U. Y., & Mooney, R. J. (2002). Text mining with information extraction. *Proceedings of the AAAI Technical Report SS-02-06*, Stanford *CA*. 60–67. Retrieved from www.aaai.org/Papers/Symposia/Spring/2002/SS-02-06/SS02-06-013.pdf

[3] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Maryland Heights, MO: Elsevier.

[4] Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos*. IDC Iview, 1142(2011), 1–12.

[5] Berry, M.W., & Castellanos, M. (2008). Survey of text mining II. *Computing Reviews, 45(9),* 548

[6] Solka, J. L. (2008). Text Data Mining: Theory and Methods. *Statistics Surveys, 2(0),* 94–112. doi: 10.1214/07-SS016

[7] R Core Team. (2017). "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing [Computer software]. Retrieved from http://www.R-project.org

[8] *The GDELT Story*. (2016). Retrieved March 07, 2017, from The GDELT Project, http://gdeltproject.org/about.html

[9] *Big Data and the Challenge of Unstructured Data*. (2017, January 01). Retrieved from Ciklum Blog, https://www.ciklum.com/blog/big-data-and-the-challenge-of-unstructured-data/

[10] Vijayarani, S., Ilamathi, J., & Nithya, M. (2015). Preprocessing Techniques for Text Mining - An Overview. *International Journal of Computer Science & Communication Networks, 5(1),* 7–16. Retrieved from http://www.ijcscn.com/Documents/Volumes/vol5issue1/ijcscn2015050102.pdf

[11] Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 20*, 19–62. doi: 10.1111/j.1365-2621.1978.tb09773.x

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval (Vol. 1)*. Cambridge, United Kingdom: Cambridge university press Cambridge.

[13] Aggarwal, C. C., & Zhai, C. X. (2013). *Mining text data. Mining Text Data (Vol.*

9781461432). Springer Science & Business Media. doi: 10.1007/978-1-4614-3223-4

[14] Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development, 2(4),* 314–319. doi: 10.1147/rd.24.0314

[15] Hardin, S. (2017). Text and Data Mining Meets the Pharmaceutical Industry: Markus Bundschus Speaks. *Bulletin of the Association for Information Science and Technology, 43(3),* 42–44. doi: 10.1002/bul2.2017.1720430314

[16] Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics, 6(1)*. doi: 10.1093/bib/6.1.57

[17] Krallinger, M., Leitner, F., & Rabal, O. (2013). Overview of the chemical compound and drug name recognition (CHEMDNER) task. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop, 2*, 2–33. Retrieved from https://doi.org/ISBN 978-84-933255-8-9

[18] Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Shen, B. (2013). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics*. doi: 10.1016/j.jbi.2012.10.007

[19] Hu, T., Andrew, A. S., Karagas, M. R., & Moore, J. H. (2013). Statistical epistasis networks reduce the computational complexity of searching three-locus genetic models. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, 397–408. doi: 10.1142/9789814447973_0039

[20] Urban, G. L., & Hauser, J. R. (2004). "Listening In" to Find and Explore New Combinations of Customer Needs. *Journal of Marketing, 68(2),* 72–87. doi: 10.1509/jmkg.68.2.72.27793

[21] Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science, 31(3),* 521–543. doi: 10.1287/mksc.1120.0713

[22] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol CA : O'Reilly Media, Inc.

[23] Wilmer A., (2017, November 02). Re: Named Entity Recognition at RAVN - Part 1 [Web log message]. Retrieved from iManage, https://imanage.com/blog/named-entity-recognition-ravn-part-1/

[24] Downey, D., Broadhead, M., & Etzioni, O. (2007). Locating complex named entities in web text. *In IJCAI International Joint Conference on Artificial Intelligence*, 2733–2739. doi: 10.1.1.104.6523

[25] Jiang, J. (2012). Information extraction from text. In: Aggarwal C., Zhai C. (Eds.) *Mining Text Data*.(pp. 11–41). Boston MA: Springer

[26] Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using

information extraction. *ACM SIGKDD Explorations Newsletter, 7(1)*, 3–10. doi: 10.1145/1089815.1089817

[27] Nadeau, D., Turney, P. D., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 4013 LNAI)*. 266–277. doi: 10.1007/11766247_23

[28] Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge MA: The MIT Press.

[29] Prasad, G., Fousiya, K. K., Kumar, M. A., & Soman, K. P. (2015). Named Entity Recognition for Malayalam language: A CRF based approach. *In 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings.* (pp. 16–19). doi: 10.1109/ICSTM.2015.7225384

[30] Liu, S., Tang, B., Chen, Q., & Wang, X. (2015). Drug name recognition: Approaches and resources. *Information (Switzerland). 6(4)*. 790-810. doi: 10.3390/info6040790

[31] Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics, 45(5)*, 885–892. doi: 10.1016/j.jbi.2012.04.008

[32] Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., & Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics, 7*. 1-11. doi: 10.1186/1758-2946-7-S1-S1

[33] Hettne, K. M., Williams, A. J., Van Mulligen, E. M., Kleinjans, J., Tkachenko, V., & Kors, J. A. (2010). Automatic vs. manual curation of a multi-source chemical dictionary: The impact on text mining. *Journal of Cheminformatics, 2(1)*. doi: 10.1186/1758-2946-2-3

[34] Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J. M., Schijvenaars, B. J. A., Van Mulligen, E. M, Kors, J. A. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics, 25(22)*, 2983–2991. doi: 10.1093/bioinformatics/btp535

[35] Pence, H. E., & Williams, A. (2010). ChemSpider: an online chemical information resource. *Journal of Chemical Education, 87(11)*, 1123-1124. doi.org/10.1021/ed100697w

[36] Eisenstein, J., Chau, D. H., Kittur, A., & Xing, E. (2012). TopicViz: Interactive topic exploration in document collections. *Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts*, 2177–2182. doi: 10.1145/2212776.2223772

[37] Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1996). Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration. *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. 238-243. Retrieved from https://pdfs.semanticscholar.org/722d/585cdfdee14fa536766075eda298c32801bf.pdf

[38] Ignat, C., Pouliquen, B., Steinberger, R., & Erjavec, T. (2006). A tool set for the quick and efficient exploration of large document collections. *Proceedings of the Symposium on Safeguards and Nuclear Material Management. 27th Annual Meeting of the European SAfeguards Research and Development Association (ESARDA-2005)*. Retrieved from https://arxiv.org/abs/cs/0609067.

[39] Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge MA: The MIT Press.

[40] Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ: British Medical Journal (Online), 345*. 1-2. Retrieved from https://fhs.mcmaster.ca/anesthesiaresearch/documents/Sedgwick2012Pearsonscorrelationcoefficient.pdf

[41] Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59(10)*. doi: 10.18637/jss.v059.i10

[42] Brownlee J. (2018, January 01). Re: A Gentle Introduction to the Bag-of-Words Model [Web log message]. Retrieved from Machine Learning Mastery, https://machinelearningmastery.com/gentle-introduction-bag-words-model/

[43] Cavnar, W. B., Trenkle, J. M., & Mi, A. A. (1994). N-Gram-Based Text Categorization. *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. 161–175. doi: 10.1.1.53.9367

[44] Trim C. (2018, January 13). Re: The Art of Tokenization [Web log message]. Retrieved from IBM Developerworks, https://www.ibm.com/developerworks/community/blogs/nlp/entry/tokenization

[45] Silge, J., & Robinson, D. (2017). *Text Mining with R - A Tidy Approach*. Sebastopol, CA : O'Reilly Media, Inc.

[46] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3(Jan)*, 993–1022.

[47] Kumar, A., & Paul, A. (2016). *Mastering Text Mining with R*. Birmingham, UK: Packt Publishing Ltd.

[48] Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72(7–9)*, 1775–1781. doi: 10.1016/j.neucom.2008.06.011

[49] Murzintcev N. (2017, November 06). Re: Select number of topics for LDA model

[Web log message]. Retrieved from RPUBS, http://rpubs.com/siri/ldatuning

[50] Arun, R., Suresh, V., Madhavan, C. E. V., & Murty, M. N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 6118 LNAI)*. 391–402. doi: 10.1007/978-3-642-13657-3_43

[51] Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17(1)*, 61–84. doi: 10.3166/dn.17.1.61-84

[52] Resnik, P., & Hardisty, E. (2010). *Gibbs sampling for the uninitiated*. Maryland Univ College Park Inst for Advanced Computer Studies.

[53] Friendly, M., & Denis, D. J. (2001). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*.[Web Document]. Retrieved from http:// www.datavis.ca/milestones/

[54] Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1988). *Network flows*. New York City, NY: Pearson Education.

[55] Wickham, H. (2017, June 21). Re: rvest: easy web scraping with R [Web log message]. Retrieved from RStudio Blog, http://blog.rstudio.com/2014/11/24/rvest-easy-web-scraping-with-r/

[56] Missile Defense Project, Center for Strategic & International Studies. (2018, January 02). Retrieved from https://www.csis.org/programs/international-security-program/missile-defense-project

[57] North Korean Missile Launches & Nuclear Tests: 1984-Present, CSIS Missile Defense Project. (2018, January 02) Retrieved from https://missilethreat.csis.org/north-korea-missile-launches-1984-present/

[58] Tsui, S., Wong, E., CHI, L., & Tiejun, W. (2017, January 1). One Belt, One Road. *MonthlyReview, 68(08)*. Retrieved from https://monthlyreview.org/2017/01/01/one-belt-one-road/

[59] Gilchrist K. (2017, August 24). China's 'Belt and Road' initiative could be the next risk to the global financial system. *CNBC*, Retrieved from https://www.cnbc.com/2017/08/24/chinas-belt-and-road-initiative-could-be-the-next-risk-to-the-global-financial-system.html

[60] Chin, G. T. (2016). Asian Infrastructure Investment Bank: governance innovation and prospects. *Global Governance: A Review of Multilateralism and International Organizations, 22(1)*, 11–25.

[61] Qi, Jin. (2016, May 18). Re: Chairman Jin Qi at the 'Belt and Road' Summit Luncheon Speech[Web log post]. Retrieved from Silk Road Fund, www.silkroadfund.com.cn/cnweb/19930/19938/32726/index.html

[62] Jinchen, T. (2016). *One Belt and One Road: Connecting China and the World*. Mckinsey. Retrieved from https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/one-belt-and-one-road-connecting-china-and-the-world

[63] Ferdinand, P. (2016). Westward ho—the China dream and "one belt, one road": Chinese foreign policy under Xi Jinping. *International Affairs, 92(4)*, 941–957. doi: 10.1111/1468-2346.12660

[64] Van Der Leer, Y., & Yau, J. (2016). *China's new silk route:The long and winding road*. Retrieved from Pwc Growth Markets Centre on January 02, 2018, from https://www.pwc.com/gx/en/growth-markets-center/assets/pdf/china-new-silk-route.pdf

[65] Wickham, H., Francois, R., Henry, L., Müller, K., & RStudio. (2017). *Title A Grammar of Data Manipulation - Package "dplyr."*. Retrieved from https://github.com/tidyverse/dplyr%0Ahttps://github.com/tidyverse/dplyr/issues

[66] Spinu, V., Grolemund, G., Wickham, H., Lyttle, I., Constigan, I., Law, J., Lee, C. H. (2017). *R: Package "lubridate."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/lubridate/lubridate.pdf

[67] Wickham, H. (2017). *R: Package "stringr."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/stringr/stringr.pdf

[68] Feinerer, I., Hornik, K., & Artifex Software Inc. (2017). *R: Package "tm."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/tm/tm.pdf

[69] Silge, J., De Queiroz, G., Keyes, O., & Robinson, D. (2018). *R: Package "tidytext."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/tidytext/tidytext.pdf

[70] Pedersen, T. L. (2017). *R: Package "ggraph."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/ggraph/ggraph.pdf

[71] Csardi, G. (2017). *R: Package "igraph."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/igraph/igraph.pdf

[72] Wickham, H., Chang, W., & RStudio. (2016). *R: Package "ggplot2."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf

[73] Grün, B., & Hornik, K. (2017). *R: Package "topicmodels."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/topicmodels/topicmodels.pdf

[74] Robinson, D. (2017). *R: Package "widyr."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/widyr/widyr.pdf

[75] Wickham, H. (2017). *R: Package "reshape2."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/reshape2/reshape2.pdf

[76] Murzintcev, N. (2016). *R: Package "ldatuning."*. Retrieved January 1, 2018, from

https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

[77] Wickham, H., & RStudio. (2017). *R: Package "scales."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/scales/scales.pdf

[78] Wickham, H., Henry, L., & RStudio. (2017). *R: Package "tidyr."*. Retrieved January 1, 2018, from https://cran.r-project.org/web/packages/tidyr/tidyr.pdf

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
22-03-2018	Master's Thesis	Oct 2016 – Mar 2018

4. TITLE AND SUBTITLE

The Application of Text Mining and Data Visualization Techniques to Textual Corpus Exploration

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jeffrey R. Smith Jr. Capt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENS-MS-18-M-163

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally Left Blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A Approved For Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Unstructured data in the digital universe is growing rapidly and shows no evidence of slowing anytime soon. With the acceleration of growth in digital data being generated and stored on the World Wide Web, the prospect of information overload is much more prevalent now than it has been in the past. As a preemptive analytic measure, organizations across many industries have begun implementing text mining techniques to analyze such large sources of unstructured data. Utilizing various text mining techniques such as n-gram analysis, document and term frequency analysis, correlation analysis, and topic modeling methodologies, this research seeks to develop a tool to allow analysts to maneuver effectively and efficiently through large corpuses of potentially unknown textual data. Additionally, this research explores two notional data exploration scenarios through a large corpus of text data, each exhibiting unique navigation methods analysts may elect to take. Research concludes with the validation of inferential results obtained through each corpus's exploration scenario.

15. SUBJECT TERMS

Text Mining, Corpus Exploration, Term Correlation, Topic Modeling

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Christopher M. Smith, AFIT/ENS
U	U	U	UU	101	**19b. TELEPHONE NUMBER** *(include area code)* (937) 255-3636

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18