2016-06

# Applications of text analytics in the intelligence community

Hall, Daniel M.

Monterey, California: Naval Postgraduate School


# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**APPLICATIONS OF TEXT ANALYTICS IN THE INTELLIGENCE COMMUNITY**

by

Daniel M. Hall

June 2016

Thesis Advisor: Johannes O. Royset
Second Reader: Jon Alt

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

| REPORT DOCUMENTATION PAGE | | | *Form Approved OMB No. 0704–0188* |
|---|---|---|---|



| **1. AGENCY USE ONLY** *(Leave blank)* | **2. REPORT DATE** June 2016 | **3. REPORT TYPE AND DATES COVERED** Master's thesis |
|---|---|---|

| **4. TITLE AND SUBTITLE** APPLICATIONS OF TEXT ANALYTICS IN THE INTELLIGENCE COMMUNITY | **5. FUNDING NUMBERS** |
|---|---|
| **6. AUTHOR(S)** Daniel M. Hall | |
| **7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)** Naval Postgraduate School Monterey, CA 93943-5000 | **8. PERFORMING ORGANIZATION REPORT NUMBER** |
| **9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)** N/A | **10. SPONSORING / MONITORING AGENCY REPORT NUMBER** |

**11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol number __NA__.

| **12a. DISTRIBUTION / AVAILABILITY STATEMENT** Approved for public release; distribution is unlimited | **12b. DISTRIBUTION CODE** |
|---|---|

**13. ABSTRACT (maximum 200 words)**

We evaluate Anseri, a commercial text analytics software, and its ability to assist a military intelligence analyst in the planning phase of major operations. The intelligence cycle involves extensive, timely, and detailed analysis of the operating environment. This requires a lot of reading by intelligence analysts to fully analyze the content. Tools that automate the initial summarization of the topic themes in a large body of text reduce the amount of time spent reading the material and focus the analyst's research efforts by providing a method to prioritize documents based on their relevance to the research topic. Anseri's utility is tested on a corpus of Islamic State press releases to demonstrate the analyst's ability to quickly gain a basic understanding of the thematic nature of the corpus and prioritize deeper research.

| **14. SUBJECT TERMS** topic analysis, intelligence, text analytics, Islamic State | **15. NUMBER OF PAGES** 69 |
|---|---|
| | **16. PRICE CODE** |

| **17. SECURITY CLASSIFICATION OF REPORT** | **18. SECURITY CLASSIFICATION OF THIS PAGE** | **19. SECURITY CLASSIFICATION OF ABSTRACT** | **20. LIMITATION OF ABSTRACT** |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | UU |

THIS PAGE INTENTIONALLY LEFT BLANK

**Approved for public release; distribution is unlimited**

# APPLICATIONS OF TEXT ANALYTICS IN THE INTELLIGENCE COMMUNITY

Daniel M. Hall
Captain, United States Marine Corps
B.S., United States Naval Academy, 2009

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the

**NAVAL POSTGRADUATE SCHOOL**
**June 2016**

Approved by: Dr. Johannes O. Royset
Thesis Advisor

LTC Jon Alt
Second Reader

Dr. Patricia Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

We evaluate Anseri, a commercial text analytics software, and its ability to assist a military intelligence analyst in the planning phase of major operations. The intelligence cycle involves extensive, timely, and detailed analysis of the operating environment. This requires a lot of reading by intelligence analysts to fully analyze the content. Tools that automate the initial summarization of the topic themes in a large body of text reduce the amount of time spent reading the material and focus the analyst's research efforts by providing a method to prioritize documents based on their relevance to the research topic. Anseri's utility is tested on a corpus of Islamic State press releases to demonstrate the analyst's ability to quickly gain a basic understanding of the thematic nature of the corpus and prioritize deeper research.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS


**I. INTRODUCTION....1**
- **A. THE BENEFIT OF TEXT ANALYSIS TOOLS IN INTELLIGENCE....1**
- **B. THE BENEFIT OF SPEED IN INTELLIGENCE....4**
- **C. THE SCOPE OF THIS RESEARCH....8**
- **D. THESIS STURCTURE....8**

**II. TEXT ANALYTICS METHODS....9**
- **A. COMMON TERMS AND DEFINITIONS....9**
- **B. PROBABILISTIC MODELS....9**
  - **1. Latent Dirichlet Allocation....10**
  - **2. Latent Semantic Analysis....11**
- **C. SPARSE LEARNING METHODS....12**
  - **1. LASSO Regression....12**
  - **2. Principal Component Analysis....13**
  - **3. Singular Value Decomposition....14**
- **D. TEXT ANALYTICS BY ANSERI....15**
- **E. SUMMARY....16**

**III. ANSERI: FAST TEXT ANALYSIS SOFTWARE....17**
- **A. ISLAMIC STATE CORPUS....17**
- **B. FORMATTING TEXT DATA FOR ANALYSIS WITH ANSERI....19**
- **C. TOPIC ANALYSIS PRODUCED BY ANSERI....20**
- **D. CASE STUDY 1: PARAMETER TUNING....20**
  - **1. Singular Values....25**
- **E. SUMMARY....30**

**IV. INTELLIGENCE COMMUNITY IMPLICATIONS OF ANSERI....31**
- **A. WORKING WITH ANSERI AS AN ANALYST....31**
- **B. CASE STUDIES: ANALYST BENEFITS OF ANSERI....32**
  - **1. Case Study 2: Topics Changing over Time....32**
  - **2. Case Study 3: Author Sentiments....35**
    - ***a. Keyword Study....38***
    - ***b. File Path Display....42***
- **C. SUMMARY....42**

**V. CONCLUSION ..... 45**
**A. RECOMMENDATIONS FOR FUTURE WORK ..... 45**

**LIST OF REFERENCES ..... 47**

**INITIAL DISTRIBUTION LIST ..... 49**

# LIST OF FIGURES


Figure 1. The Joint Intelligence Process. Source: U.S. Joint Chiefs of Staff, (2013).....5

Figure 2. Relationship of Data, Information, and Intelligence. Source: U.S. Joint Chiefs of Staff (2013).....6

Figure 3. Joint Intelligence Preparation of the Operational Environment. Source: U.S. Joint Chiefs of Staff (2013).....7

Figure 4. 100 Topic LDA Model Fit to 17,000 Articles. Source: Blei (2012).....11

Figure 5. Geometric Representation of LSA. Source: Landauer et al. (2013).....12

Figure 6. Comparison of Computing Time Between Sparse PCA and LDA. Source: Godbehere (2015).....14

Figure 7. Document Identification in the Islamic State Corpus. Source: Whiteside (2014).....19

Figure 8. Graphical Representation of the Term/Document Matrix and the Effects of Different nf and nd Values. Source: Godbehere (2016).....23

Figure 9. Singular Values Produced During the SVD Factorization of the File Level versus Sentence Level Term/Document Matrices. .....26

Figure 10. Singular Values Associated with First 50 Topics in the Entire Corpus. .....27

Figure 11. Singular Values Associated with the First Eight Topics from the 2004 Sentence Level Term/Document Matrix.....28

Figure 12. Singular Values Associated with the First Eight Topics from the 2005, 2006, and 2008 Sentence Level Term/Document Matrices.....29

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES


Table 1. Topic Summarization Produced by Anseri on the 2004 File Level Data with Default Settings ..... 21

Table 2. Topic Summarization Produced by Anseri on the 2004 Sentence Level Data with Default Settings ..... 22

Table 3. Anseri Performance on 2004 Sentence Level Data When Using the Recommended Settings of nf = 20 nd = 200. ..... 24

Table 4. First Eight Topics with Singular Values from the 2004 Sentence Level Term/Document Matrix. ..... 28

Table 5. Sentence Level Data Performance at Recommended Settings ..... 33

Table 6. Topic Summarization of Each Author Corpus ..... 36

Table 7. Common Naïve Keyword Search Terms that Yield Results from Each Author in the Corpus. ..... 39

Table 8. Naïve Keyword Search Results from the Zarqawi Sentence Level Data. ..... 40

Table 9. Focused Keyword Search Results from the Zarqawi Sentence Level Data. ..... 41

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| GUI | graphical user interface |
| IPOE | intelligence preparation of the operational environment |
| IS | Islamic State |
| J-2 | intelligence directorate of the joint staff |
| JSON | JavaScript object notation |
| LASSO | least absolute shrinkage and selection operator |
| LDA | latent Dirichlet allocation |
| LSA | latent semantic analysis |
| LSI | latent semantic indexing |
| nf | number of features per topic |
| nd | number of documents per topic |
| PCA | principal component analysis |
| PIR | priority intelligence requirement |
| pLSA | probabilistic latent semantic analysis |
| SPCA | sparse principal component analysis |
| SVD | singular value decomposition |

THIS PAGE INTENTIONALLY LEFT BLANK

# EXECUTIVE SUMMARY

The recent technology push to digitally store media has made access to large amounts of documents easy and has raised the expectation of the amount of research one can conduct on countless topics. The intelligence community has identified the handling of large amounts of digitally stored data as a concern for future intelligence analysts. The amount of digital text available to individual analysts today makes it nearly impossible to conduct a thorough analysis of the material manually. Tools that automate the initial summarization of the topic themes in a large body of text reduce the amount of time spent reading the material and focus the analyst's research efforts by providing a method to prioritize documents based on their relevance to the research topic. Text analytics algorithms are able to accomplish such tasks on large text corpora and assist analysts in gaining a basic understanding of the thematic nature of the included texts.

To evaluate the utility of the commercial text analytics software Anseri, a corpus of Islamic State press releases is analyzed for its content. Anseri is a fast text analysis software that can be used by intelligence analysts to gain efficiency in the processing and exploitation portion of the intelligence process. We will attempt to classify these documents in a manner that would be practical to an intelligence analyst creating a time-sensitive product. If these text analytics tools can streamline the processing and exploitation portion of the intelligence process, they will significantly reduce the amount of time spent reading the information required to produce timely actionable intelligence products to decision makers. Three case studies are used to demonstrate the capabilities of Anseri.

Case Study 1 tunes the input parameters to achieve efficient results from the topic summarization. This case study verified that Anseri performed as expected with the Islamic State press releases. When Anseri analyzes the corpus with 20 features and 200 documents per topic, the results are readable and meaningful. This case study also explores the significance of the magnitude of the singular values used in the singular value decomposition that produces the topics. Case Study 2 gains insight into the analyst's ability to use Anseri to analyze the topic trends over time. Anseri is capable of

representing the topics within the corpus and when the texts are separated by date the analyst can identify topic trends over time. Case Study 3 demonstrates Anseri's ability to conduct semantic analysis when the corpus is segregated by author. This case study also analyzes the utility of the keyword search and file path display functions of Anseri. These functions increase the efficiency of the analyst by identifying the documents that closely relate to the commander's priority intelligence requirements.

Anseri increases the efficiency of intelligence analysts. When conducting intelligence support to operational planning timely and accurate production is essential to success. Adding Anseri to the military intelligence analyst's toolkit would reduce the amount of time and effort spent processing data. As a result, more time can be spent properly analyzing information and producing high quality intelligence products for decision makers.

# ACKNOWLEDGMENTS

I would like to express my appreciation to my advisor, Dr. Johannes Royset, for the support and guidance in this research endeavor. His mentoring was essential to successfully completing this project. I would also like to thank the Naval Postgraduate School faculty of the Operations Research Department for their dedication to the students, ensuring growth through a challenging curriculum. Thank you as well to the United States Marine Corps for providing this opportunity to advance my education.

Most of all, I want to thank my wife, Jessica, for all the love and support she provided me during my time at the Naval Postgraduate School. None of this is possible without her support.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

The recent technology push to digitally store media has made access to large amounts of documents easy and has raised the expectation of the amount of research one can conduct on many topics. The amount of digital text data available to individual analysts today makes it nearly impossible to conduct thorough analysis of the material manually. Planning military operations is reliant on timely and accurate intelligence of the operating environment. The threat environment for the United States military is growing more uncertain due to non-state actors using unprecedented media methods for recruiting and strategic communication. This threat needs to be researched and analyzed with speed and accuracy; however, the amount of available information is a distraction to intelligence analysts. The intelligence community would benefit from an automated tool that assists in the tasks of collection, processing, and exploitation. Software tools that use text analytics can conduct topic analysis and reduce the amount of reading required of human analysts. Anseri is a fast text analysis software, designed for business, which can be used by intelligence analysts to gain efficiency in the intelligence process. Text analytics automates analysis of unstructured text, identifying patterns and themes, and producing a synthesized intelligence product (What is text analytics, 2016). Software will not replace humans in the intelligence process but it will make analysts more effective and efficient.

## A. THE BENEFIT OF TEXT ANALYSIS TOOLS IN INTELLIGENCE

The intelligence community has identified the handling of large amounts of digitally stored data as a concern for future analysts. As technology continues to allow for greater access to individuals around the world, the amount of data that will need to be analyzed will expand. Intelligence analysts are relied upon to be the experts in their area of responsibility and as the amount of available information grows so does the potential workload of that analyst to keep abreast of daily events. As stated in the Joint Intelligence Publication, “during execution, intelligence must stay at least one step ahead of operations and not only support the current phase of the operation, but also

simultaneously lay the informational groundwork required for subsequent phases" (U.S. Joint Chiefs of Staff, 2013, p. xv). Having a tool that can estimate the underlying thematic nature of large amounts of text allows the analyst to focus on the most urgent articles first when building an understanding of the operational environment.

Analyzing text can be challenging for a machine due to the nuances of language and interpersonal relationships. Understanding the thematic nature of a body of texts creates several challenges. For instance, words with positive connotation can be considered negative under certain circumstances (Vinodhini & Chandrasekaran, 2012). Trained intelligence analysts paired with machines that are designed to understand text could reduce the amount of time spent learning the nuances of the languages used in the text. Intelligence analysts can be overwhelmed by the glut of information in the form of reports articles, imagery, and message traffic. Intelligence personnel are often trying to find a needle in a haystack and careful application of these techniques could help them in that effort.

In support of operational planning, commanders will continue to rely on their intelligence staff to develop a comprehensive understanding of the situation in these areas. The current mechanism for focusing intelligence efforts are the Priority Intelligence Requirements (PIR). PIRs are mission specific and will be continually updated throughout each phase of the operation. Well written PIRs will "ask only one question, focus on a specific fact, event, or activity, and provide intelligence required to support a single decision" (U.S. Department of the Army, 1994). The Army's Field Manual 34–2 (1994) provides the following examples of good PIRs: "Will the enemy use chemical agents on our reserve force before it leaves the assembly area?" or "Which bridges over the river are intact?" These PIRs are specific to a portion of the operation and there is a time limit for the relevance of the information. The Intelligence Officer receives the requirements from the commander, focuses the scope of each requirement, and assigns the focused PIRs to his staff to answer. The commander needs a timely and accurate response to support his decision.

During the planning phase, mainly background research is used to answer the PIRs. Information from the web, news media, and other intelligence reporting agencies

can be gathered and sorted into various directories with relative ease. The time-consuming task is then reading each document to evaluate its relevancy and producing an intelligible product. Text analysis tools could be used to read an entire directory of reports and products within seconds. The resultant topic summarization can direct the analyst to documents that contain relevant information to the PIRs.

Sparse machine learning techniques have been identified as being a useful tool in the field of text analytics. Previous work in this field has included multi-document text summarization, comparative analysis, and the visualization of large text corpora (El Ghaoui et al., 2013). Businesses have also identified the need to analyze the enormous amount of consumer, industry, product and company information that can be gathered from the Internet (Chen, Chiang, & Storey, 2012). The use of social and news media can have great effects on marketing strategies and text analytics techniques can identify trends and increase productivity. Machine learning techniques have been used to classify text since the 1990s (Sebastiani, 2002); however, the incredible growth in Internet connectivity worldwide has increased the need for efficient tools to classify and categorize texts. These techniques, added to the tool kit of the Military Intelligence Analyst, could streamline the research process and eventually aid in the predictive analysis.

Search engines have evolved in the last 20 years into the powerful commercial systems they are today. Many of the foundational text processing and indexing techniques in the field of text analytics are employed in document searching and management systems (Chen et al., 2012). Anseri, a commercial software, has been developed by SumUp LLC to apply sparse machine learning techniques to large text corpora in an attempt to identify patterns in the text and visualize the implications of the text. SumUp LLC was founded to create customized business solutions for the world's leading retailers and manufacturers using large-scale text analytics and statistical modeling. SumUp's Anseri platform enhances integrated customer experience management, market intelligence, document management and eCommerce optimization. Anseri is an integrated toolset for ingesting, managing, and analyzing documents from virtually any source (SumUp Corporate Fact Sheet, 2016). Anseri offers a unique

capability to the intelligence community by changing the way in which large corpora of text is processed.

## B. THE BENEFIT OF SPEED IN INTELLIGENCE

Intelligence support to decision making requires a timely and accurate assessment of the situation. The role of the intelligence section in a military unit is to support the commander in the decision making process with information and assessments that support mission accomplishment (U.S. Joint Chiefs of Staff, 2013). The Joint Forces Commander is responsible for knowing his environment, objectives and mission. The Intelligence directorate of the joint staff (J-2) is guided by specific tasks such as, "inform the commander, describe the environment, identify, define and nominate objectives, support planning and execution of operations, counter adversary deception and surprise, support friendly deception efforts and assess the effectiveness of operations" (U.S. Joint Chiefs of Staff, 2013, ix). Figure 1 shows a graphical representation of the Joint Intelligence Process as defined by Joint Publication 2–0. The goal of intelligence is to support all phases of operations with timely and actionable intelligence that supports the decision maker's understanding of the situation. This thesis work targets the intelligence support to planning with an evaluation of a specific tool that supports the processing and exploitation phase of the intelligence process.


Figure 1. The Joint Intelligence Process. Source: U.S. Joint Chiefs of Staff, (2013)

Text analytics can be used to reduce the time spent on the initial collection and processing and exploitation phases of the intelligence process. Figure 1 displays each step of the intelligence process in boxes of equal size. If the boxes were to be sized by level of importance then "Dissemination and Integration" would be biggest because that is the point in the process when the intelligence is finally put to use. "Analysis and Production" would be the next biggest because understanding the information is important. The "Collection" and "Processing and Exploitation" phases are lower on the list of importance but take a disproportionate amount of time. It is important that the correct information be collected and processed but these steps do not directly benefit the warfighter and need to be completed in a short amount of time so the decision maker can make timely decisions. Figure 2 attempts to show the process of converting information

collected from the operational environment into intelligence. The grey-shaded area represents the amount of information that is available and as it passes through the lenses of "Collection," "Processing and Exploitation," and "Analysis and Production" the useful information is extracted into intelligence. Text analytics tools can be used to decrease the efforts required in the "Processing and Exploitation" phase through automation of the initial understanding of the thematic nature of the text. This allows more time to analyze the information and produce an accurate intelligence product.

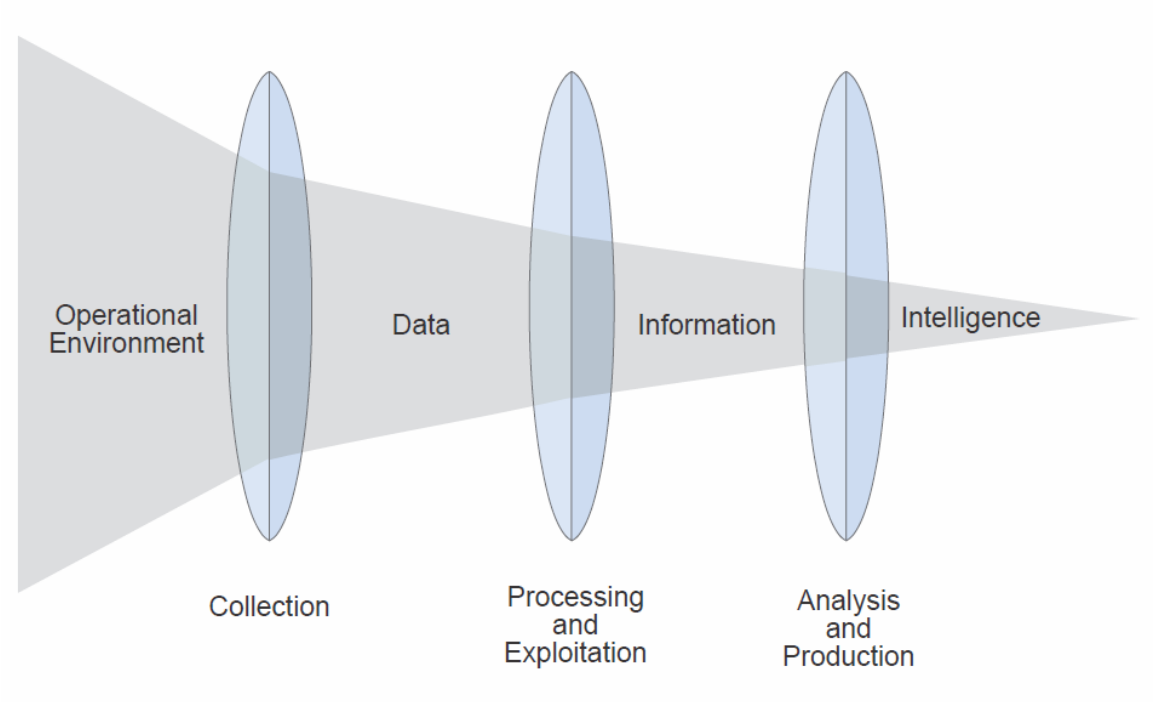



Figure 2. Relationship of Data, Information, and Intelligence.
Source: U.S. Joint Chiefs of Staff (2013)

The main deliverable for the intelligence staff during the planning phase of operations is the Intelligence Preparation of the Operating Environment (IPOE). Figure 3 shows the four-step process used by intelligence professionals to describe the operating environment to the decision makers. IPOE provides a systemic methodology to analyze information about the operational environment and identifies adversary courses of action by probability and consequence (U.S. Joint Chiefs of Staff, 2013). IPOE is an extensive

report that takes a lot of time and energy to produce. With faster processing, developing IPOE can be given more attention to ensure it is accurate and clearly communicates the analysis conducted.

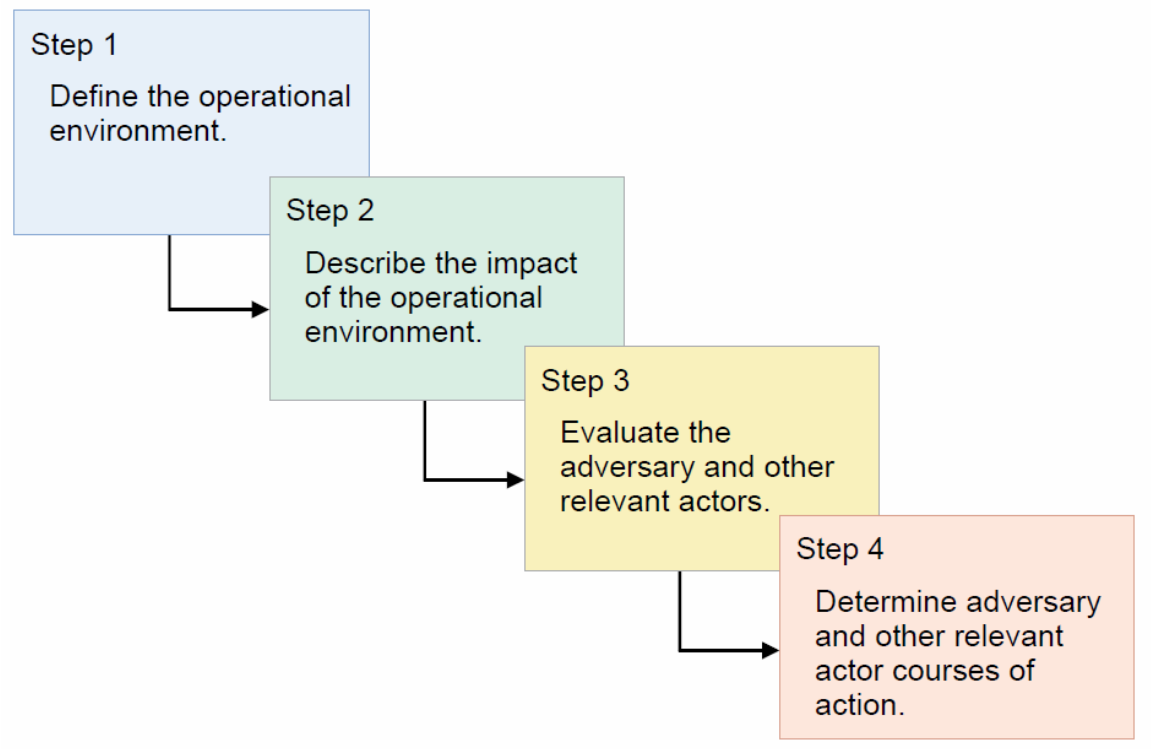



Figure 3. Joint Intelligence Preparation of the Operational Environment. Source: U.S. Joint Chiefs of Staff (2013)

Speed is important in warfare. Decision makers need to make timely decisions to maintain the initiative and outmaneuver the enemy. Those who support the decision makers need to be able to produce timely and accurate reports that ensure their decisions are based on a thorough understanding of the operating environment. Good decisions made faster than the enemy give commanders the best chance to achieve victory.

## C. THE SCOPE OF THIS RESEARCH

To study the capability of Anseri, a corpus of Islamic State[1] (IS) press releases will be analyzed for its content. The dates of the documents range from 2004 to 2013. We will attempt to classify these documents in a manner that would be practically useful to an intelligence analyst creating a time sensitive product. If these text analytics tools can streamline the processing and exploitation portion of the intelligence process, they will significantly reduce the amount of time spent reading the information required to produce timely actionable intelligence products to the decision makers. Using understanding of the underlying thematic nature of the corpus one can identify the documents that require the immediate attention of intelligence analysts to assess the significance of the resource.

Several case studies are explained to show the relevance and utility of text analytics software to the intelligence professional. The case studies evaluate the ability of Anseri to quickly and accurately identify significant documents in the corpus, streamline the processing of text data, and identify themes and opinions of various authors in the corpus. With this type of tool in the analyst's toolkit, a skilled analyst can be more effective and efficient while supporting the command.

## D. THESIS STURCTURE

Chapter II features a literature review of previous work in the field of text analytics and topic analysis. Chapter III describes the first case study conducted with Anseri and reports the results. Chapter IV discusses two case studies from the practical perspective of an analyst and reports the implications to the fields of intelligence analysis and text analytics. Finally, Chapter V concludes and offers recommendations for future work to be conducted with Anseri.

[1] This thesis refers to the organization commonly known as ISIS or ISIL as "The Islamic State." The use of the term "Islamic State" does not recognize the legitimacy of the organization but is using the name that the organization uses to describe itself. The organization will be referred to as the Islamic State (IS) for the remainder of the thesis.

# II. TEXT ANALYTICS METHODS

This chapter serves as a literature review and a description of existing text analysis methods. The process of analyzing text incorporates techniques from statistical learning, data analysis, and regression. Text analysis is important because differences in language, dialect, and interpretation combined with the great volume of the corpus prove to be a time-consuming task for an analyst to perform (Ingersoll, Morton, & Farris, 2013). Autonomous software tools that can identify the thematic nature of texts provide the analysts a basic understanding of the thematic nature of the text in far less time than it would take to conduct a detailed reading of the text. Current methods for text analysis can be separated into two major categories: probabilistic models and sparse learning methods.

## A. COMMON TERMS AND DEFINITIONS

The text analysis concepts and models operate on a common set of terms and definitions. A feature is a word or string of characters in text, separated by a space. Punctuation is removed from the text when formatting the corpus. A document is a segment of text that generates a row in the term/document matrix. Common text analysis tools will allow for a document to be as large as an entire text file saved in the corpus or as small as an individual sentence. A term/document matrix is a numerical representation of the co-occurrence of features and documents in the corpus. A term/document matrix is represented by a matrix $A$ with $m$ rows of documents and $n$ columns of features or n-grams. An n-gram is a combination of $n$ features used to analyze terms that occur together as one feature. Stop words are features that are filtered out of the corpus because they do not have meaning in topic analysis. There is no one prescribed set of stop words common to all algorithms, but most include words that would not be capitalized in a document title. A topic is specified by a numerical vector that corresponds to a group of prevalent features within the term/document matrix.

## B. PROBABILISTIC MODELS

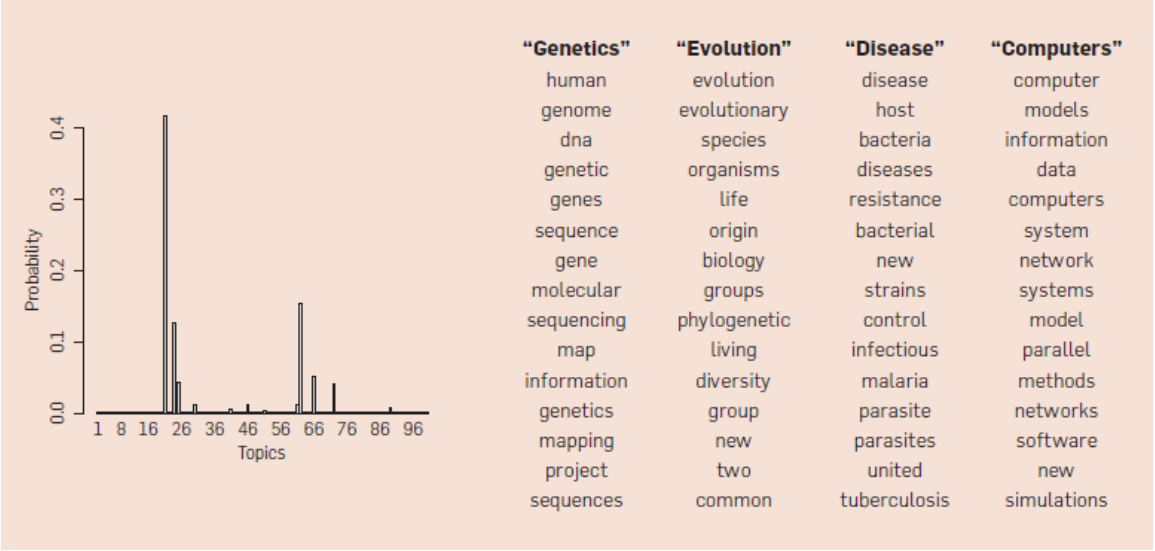
In an effort to improve upon current keyword search methods used by common search engines, machine learning researchers offer a group of probabilistic topic

modeling algorithms to identify the themes of large and unstructured corpora based on more than the occurrence of keywords. Key advantages of topic models include they can be applied to large collections of documents, organize the collection according to the discovered themes, and be adapted to many kinds of data (Blei, 2012). Particularly advantageous to the intelligence analyst is the fact that topic models do not require prior understanding or classification of the documents in the corpus to accurately determine the thematic nature of the corpus.

### 1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that explains the themes of sets of text by clustering groups of words into topics with an assumed underlying set of topics in the corpus (Ingersoll et al., 2013). LDA "provides a tool for discovering and exploiting the underlying thematic structure in large archives of text" (Blei, 2012). LDA uses a bag-of-words assumption meaning the order of the features does not matter. Also, the order of documents does not matter and "the number of topics is assumed known and fixed" (Blei, 2012). LDA was developed as follow on work from a previously developed probabilistic model, probabilistic latent semantic analysis (pLSA). LDA uses a distribution of words and topics to best estimate the themes of the documents in the corpus.

To implement LDA one must define a topic over a fixed vocabulary. Then, as described by Blei (2012), "for each document, the topics are generated in a two-step process." The first step, Blei continues, is to "choose a probability distribution over topics." Then for each word in the document, "choose a topic from the distribution over topics and choose a word from the corresponding distribution over vocabulary" (Blei, 2012). The results are an understanding of each document based on the proportion of the document that matches each topic. Figure 4 provides an example of what can be expected as output from an LDA topic model. These models are effective but require a prior understanding of the nature of the corpus and much computing time.




100 topic LDA model fit to 17,000 articles from the journal *Science*. At left are the inferred topic proportions from one article in the journal. At right are the most frequent words from the most frequent topics found in the article.

Figure 4. 100 Topic LDA Model Fit to 17,000 Articles. Source: Blei (2012)

### 2. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a well-known technique for applications in text analysis and information retrieval. The goal of LSA is to find a "data mapping which provides information well beyond the lexical level and reveals semantical relations between the entities of interest" (Hofmann, 1999). The semantic level of the text refers to the understanding of what is meant beyond what is written on the page. The problems that LSA needs to address are polysemy and synonymy because multiple meanings for words and phrases can cause text analysis and indexing algorithms to perform poorly. LSA has been contested for its claims that the meaning of a text can be assumed by the occurrence of words in its composition (Landauer, McNamara, Dennis, & Kintsch, 2013).

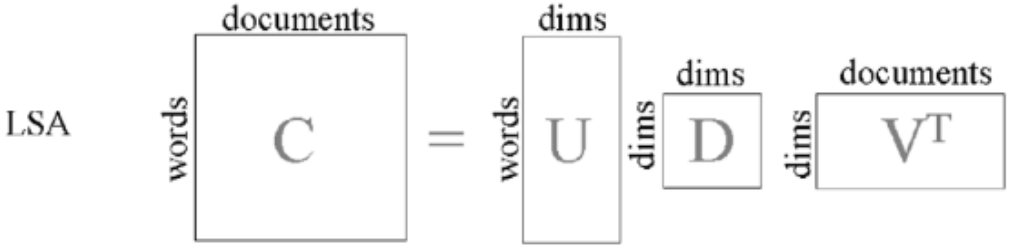



Figure 5. Geometric Representation of LSA. Source: Landauer et al. (2013)

Hofmann (1999) introduces a probabilistic LSA (pLSA) based on a statistical latent class model. The introduction of the probabilities into the pLSA algorithms uses a local maximum likelihood function to reduce the number of variables used to analyze the topic structure. The reduced complexity of the model allows for faster runtimes and achieves accuracy that is as good if not better than that achieved with the traditional LSA. The desire to reduce the number of prediction variables is the driving force behind sparse learning methods discussed in the next section.

## C. SPARSE LEARNING METHODS

Sparse machine learning methods applied to text can address several text analytics tasks to include topic summarization, division of multiple corpora, and visualization and clustering (El Ghaoui et al., 2013). Topic summarization divides the corpora into two classes, one that corresponds to the topic, and one that does not. This division can be used to highlight the terms with the most predictive power for the model. Topic summarization is closely related to LDA (El Ghaoui et al., 2013). Discrimination between several text corpora is used to find terms that best describe the differences between corpora. Visualization and clustering models provide insight into the topic themes in large data sets (El Ghaoui et al., 2013).

### 1. LASSO Regression

LASSO regression is a variation of least squares regression that, “performs both variable selection and regularization to enhance the prediction accuracy and interpretability of a statistical model” (El Ghaoui et al., 2013, p.3). In text analysis,

LASSO can be used to determine which combination of features holds the most predictive power within the term/document matrix formed by the corpus. Commonly, the $L_1$-norm is the penalty to encourage sparsity in the number of predictors in the model. Sparsity is important in a text analysis model because a great deal of superfluous words within a corpus generates noise in the model and can make the results unintelligible (El Ghaoui et al., 2013). LASSO can be used to generate a corpus specific set of stop words by eliminating the features that do not hold statistically significant predictive power.

### 2. Principal Component Analysis

Principal Component Analysis (PCA) is an example of unsupervised learning. In the text analytics context, PCA can be used to reduce the feature space. PCA is a popular tool for dimension reduction in regression and can be applied to the term/document matrix to place greater importance on the variables that have the best predictive power (James, Witten, Hastie, & Tibshirani, 2013). "The first principal component of a set of features is the normalized linear combination of the features that has the largest variance" (James et al., 2013, p. 375). PCA then projects the data onto a subspace that can be used to perform linear regression analysis on the data. The low dimensionality of the data projection from PCA is desirable for text analysis and a version of PCA to be discussed later known as Sparse PCA is used to enhance the readability of results from text analysis algorithms (El Ghaoui et al., 2013).

Sparse PCA can be used for topic discovery in large collections of documents by computing a low rank approximation of the term/document matrix (Godbehere, 2015). This result is desirable for topic analysis because the term/document matrix can become large quickly and the sparsity improves computing time (Calafiore & El Ghaoui, 2014). Sparsity in the feature space and the document space provided by sparse PCA allows the user to decrease the feature space by identifying features and documents that explain the variance in the data (El Ghaoui et al., 2013).

Text analytics involves large datasets that include many more features than documents. Sparse PCA can be implemented in an iterative fashion to approximate the topics in a corpus. Godbehere (2015) explains that this method solves a rank-1

approximation of the term/document matrix and extracts the principal component to form a topic. The matrix is then modified by removing the feature pattern that generates the rank-1 approximation before starting the next iteration through a process called deflation. The process can be iterated as many times as desired to extract the most prevalent topics in the corpus. Sparse PCA performs a similar task to that of LDA; however, the advantages of algorithms that use sparse PCA include savings in computing time, no prior knowledge of the underlying thematic nature of the corpus is required, and a more focused view of the topics. Figure 6 shows a side by side comparison of computation times of algorithms using LDA and Sparse PCA.

| | SPCA | LDA |
|---|---|---|
| Time to 25 % complete: | est. 8 minutes 21 seconds | 7 hours 16 minutes 57 seconds |
| Total Time: | 33 minutes 25 seconds | est. 29 hours 7 minutes 48 seconds |
| Time Per Topic: | 2 seconds | 105 seconds |
| Time to First Response: | 2 seconds | est. 29 hours 7 minutes 48 seconds |

Performance of two methods to extract 1000 topics from a corpus of 415,041 documents. LDA execution was stopped after 25% complete and the remaining times were extrapolated. The Sparse PCA model will return topics as they are discovered and continue to calculate remaining topics independent of the first. The LDA model must run to completion. Thus, the time to first response and total time are equal for LDA.

Figure 6. Comparison of Computing Time Between Sparse PCA and LDA. Source: Godbehere (2015)

### 3. Singular Value Decomposition

Singular Value Decomposition (SVD) of a rectangular matrix is a three term factorization that produces a low rank approximation of the original matrix while preserving much of the original information in the matrix. Text analysis algorithms that use SVD use term/document matrices to perform semantic indexing of corpora. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990) introduced a new method for automatic indexing and retrieval through latent semantic analysis, which relies heavily on SVD. Latent Semantic Indexing (LSI) models are focused on document clustering and retrieval based on the semantic nature of the documents. LSA models use a

high dimensional representation of the term/document matrix and SVD provide a method for reducing that dimensionality while preserving the most informative features and documents included in the matrix. The goal of an LSI is to address shortfalls in term-matching that degrade keyword search performance. Broadly, these deficiencies can be called synonymy and polysemy (Deerwester et al., 1990). Synonymy is when there are multiple words with the same meaning and polysemy is when there are multiple definitions for one word.

With SVD the term/document matrix (*A*) is factored into three matrices *U*, *Σ,* and *V*. In the factorization *U* and *V* have orthogonal, unit length columns and Σ is a diagonal matrix of singular values. The term/document matrix is first preprocessed and reweighted to ensure that the length of the documents does not have undue influence in the model (Deerwester et al., 1990). Equation 1 shows an SVD factorization of any matrix A.

$$A = U \Sigma V^T \tag{1}$$

SVD is the basis for several nonlinear optimization problems and is effective at compressing text data into a more manageable representation (Calafiore & El Ghaoui, 2014). The work of Deerwester et al. (1990) to incorporate SVD in his LSI model created a basis for future work in topic summarization and fast text analytics software. The ability to represent text data in lower dimensions reduces the computation time and memory requirements of text analytics algorithms.

## D. TEXT ANALYTICS BY ANSERI

Anseri uses a combination SVD and Sparse PCA to quickly read and analyze text. The combination of techniques provides the user with a basic topic summarization of a large corpus in seconds. The topic summarization can then be used as a baseline for the user to further explore the content of the corpus. Anseri addresses the shortcomings of LDA and keyword search by analyzing the entire content of the text, identifying topic patterns, and assigning topics without requiring the preprocessing of the data.

## E. SUMMARY

This chapter provided a broad overview of current text analysis methods that provide the ground work for the analytical tool discussed in this thesis. Probabilistic topic models and sparse learning techniques are essential to handling large volumes of text data in a timely manner. The following chapters will use Anseri to conduct topic analysis on a corpus of Islamic State press releases. The results from Anseri will be discussed on both a quantitative technical level and a qualitative non-technical level.

# III. ANSERI: FAST TEXT ANALYSIS SOFTWARE

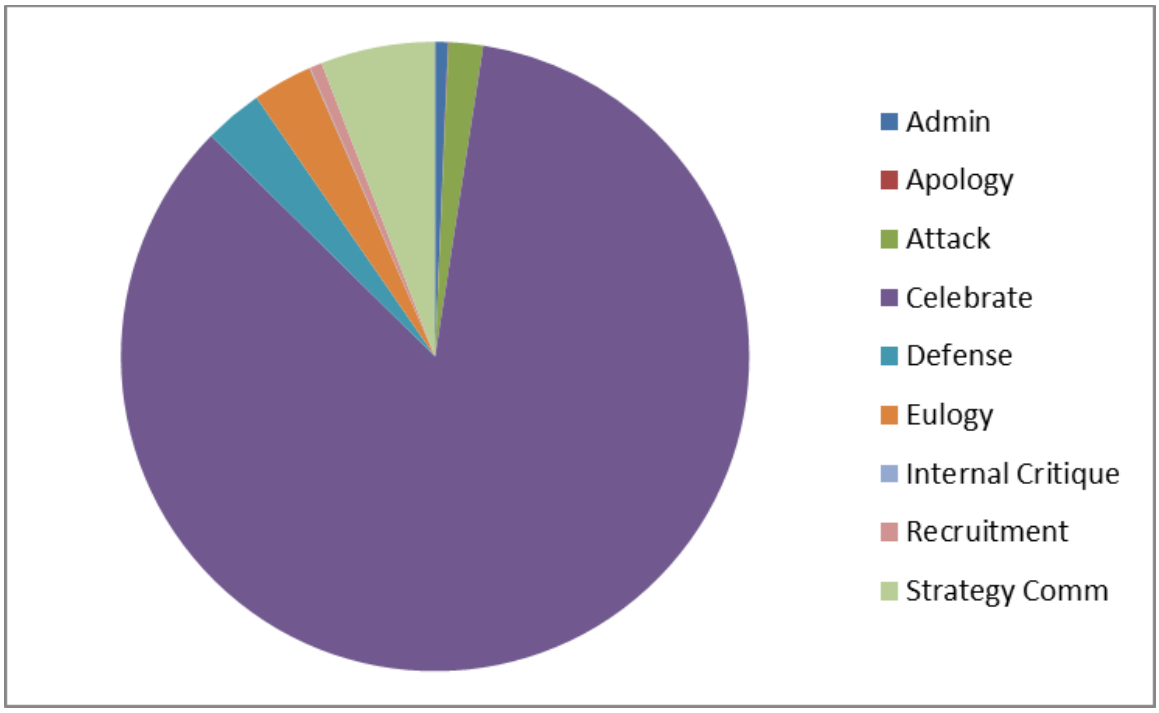
This chapter provides a discussion of the technical attributes of Anseri and the ability to summarize the content of the IS corpus. We begin by describing the data set to be analyzed, proper formatting of the data for ingestion into Anseri and the method in which Anseri summarizes the topics. Finally, a case study based on tuning the various parameters of Anseri, will demonstrate the parameter tuning required to achieve accurate results. For a unit that is preparing to deploy to an unfamiliar area, the intelligence analyst is assigned the task of preparing the operational environment and any pertinent information surrounding the operation. The first step is to compile all reports on the subject area. Often the reports and products will be collected and sorted into various directories creating a reading list for the analyst. An analyst equipped with Anseri can then use this tool to gain initial insights to the content of the corpus without having to read every page. The goal is to be able to prioritize your reading list to streamline the workload.

## A. ISLAMIC STATE CORPUS

The corpus studied in this thesis is a collection of various IS communications from differing levels of leadership. The corpus encompasses the formative years of the IS organization. To better understand the cognitive process of IS leadership, Whiteside (2014) conducted a content analysis of a wide assortment of IS documents. He collected a corpus of captured IS documents from U.S. government sources and press releases from various collections (Haverford College, Global Terror Alert, and jihadist outlets that posted al Furqan and al Fajr Media). There are a total of 2,995 files saved in the corpus covering a date range of 2004 to 2013. Attempting to read and analyze between 3,500 and 4,000 pages is time-consuming and ineffective when asked to produce a timely intelligence report.

The benefit of working with this IS corpus is that it has been previously analyzed. Whiteside (2014) read and analyzed each document for his dissertation and labeled each document as administrative, apology, attack, celebrate, defense, eulogy, internal critique,

recruitment, strategic communication and warning. Figure 7 shows the proportion of each type of document in the corpus. The document labels are listed in the corpus documentation separate from the press releases and will not be input into Anseri. This allows Anseri to conduct unsupervised learning independent of the labels. The corpus documentation will only be used as a form of validation for the results produced and will not affect the performance of Anseri. The corpus is representative of several significant events in the history of IS. The organization undergoes a transition from obscurity to the focus of U.S. military operations to the media spotlight in the period of time covered by this corpus. At the end of the time period encompassed by the corpus, IS becomes a prominent story in the U.S. media (Whiteside, 2014). Having a historically focused corpus will highlight the ability of Anseri to provide assistance in background research through topic summarization. We will attempt to gain an understanding of the documents contained by the corpus in a manner that would be practical to an intelligence analyst creating a time sensitive product.




The Islamic State corpus has been previously processed identifying these nine categories in which each document can be identified.

Figure 7. Document Identification in the Islamic State Corpus. Source: Whiteside (2014)

## B. FORMATTING TEXT DATA FOR ANALYSIS WITH ANSERI

The corpus is input into Anseri as a term/document matrix in JavaScript Object Notation (JSON). JSON is a file format from the C-family of programming languages that uses "human readable text to transmit data objects consisting of attribute/value pairs" (Introducing JSON, 2016). JSON can produce these attribute/value pairs from text written in virtually any language. The corpus can also be saved under many different file formats, which are converted into attribute/value pairs by first reverting all characters to lower case, scrubbing all punctuation and special characters, and removing redundant white spaces. Finally, the system prescribed stop words are removed and what is left is the term/document matrix (El Ghaoui et al., 2013). For our purposes, the files are saved in the directory as .docx and .pdf files because those are fairly common file types downloaded from various web pages.

When the term/document matrix is input into Anseri two databases are produced; one with file level data and the other with sentence level data. In the file level database, a document is represented by the entire file saved in the directory, therefore the document count is based on the number of files saved in the directory. In the sentence level database, files in the directory are divided by each individual sentence written within each file. The document count is then the number of sentences in the directory. This greatly increases the number of documents in the term/document matrix increasing the number of rows in the matrix.

## C. TOPIC ANALYSIS PRODUCED BY ANSERI

Anseri uses SVD to produce a low rank approximation within the term/document space and inserts sparsity to achieve readable results by reducing noise. The decomposition process identifies repeated patterns within segments of text and Anseri extracts these patterns as topics (Godbehere, 2016). Anseri can easily build a corpus and incorporates topic analysis, machine learning and keyword search to gain insights into the thematic nature of the corpus (SumUp Corporate Fact Sheet, 2016). The algorithm uses a process known as deflation to adjust the term/document matrix after extracting each topic to accurately reflect the features that still require analysis.

Anseri's command line arguments include key parameters that can be set by the user to tailor the results regarding the number of features per topic, number of documents, per topic and the number of topics to produce. The default settings for these parameters are eight features per topic, 16 documents per topic, and eight topics. The number of topics parameter dictates the number of iterations Anseri will perform to discover topics and subsequently display for the analyst. We will maintain the number of topics at the default setting. The deflation takes place after each iteration by removing the columns that contain the features used to determine the topic and then the singular values are then recalculated and the process is repeated (Godbehere, 2016).

## D. CASE STUDY 1: PARAMETER TUNING

Case 1 serves as a method of verification of the commands and their general performance on a large data set. First, Anseri was run using the default parameters on

both the file level data and the sentence level data for each one year directory in the corpus. Table 1 and Table 2 provide the output of Anseri when the default settings are run on the 2004 directory. Each column in the table represents one topic in the order that Anseri discovered them. Of note, the top eight topics provided by the file level data and the sentence level data are different. The topics are different because of the bag-of-words assumption that eliminates the effects of word order in each document. When the documents contain many words the co-occurrence of features identifying a topic can be spread within the document. For example, consider a news article that is many pages long, though on the file level it is considered to be one document the number of topics discussed over those pages could vary greatly. If the features identifying a topic only appear at the beginning and end of the document, they may not be representative of the topic theme and therefore the results can be deceiving when analyzing the file level data. The sentence level data restricts the amount of topic variation in each document because there is typically only one topic to each sentence. Anseri will output to the user a numbered topic with the features that were used to create each topic and user needs to interoperate the meaning of the features and assign a topic name.

Table 1. Topic Summarization Produced by Anseri on the 2004 File Level Data with Default Settings

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| qaida, committee, mesopotamia, jihad, allah, december, europe, praise | tawheed, movement, wal, globalterroralert, brothers, september, evan, kohlmann | august, info, www, http, abu, iraq, zarqawi, musab | october, baghdad, bosnian, american, akhbar, allahu, capital, blessing | military, wing, mujahideen, involvement, commission, conflicts, sale, final | bin, afghan, report, laden, network, messenger, july, communiqué | november, ramadan, soldiers, responsibility, apostates, claims, americans, operation | lion, martyrs, battalion, honor, martyrdom, gracious, merciful, benevolent |

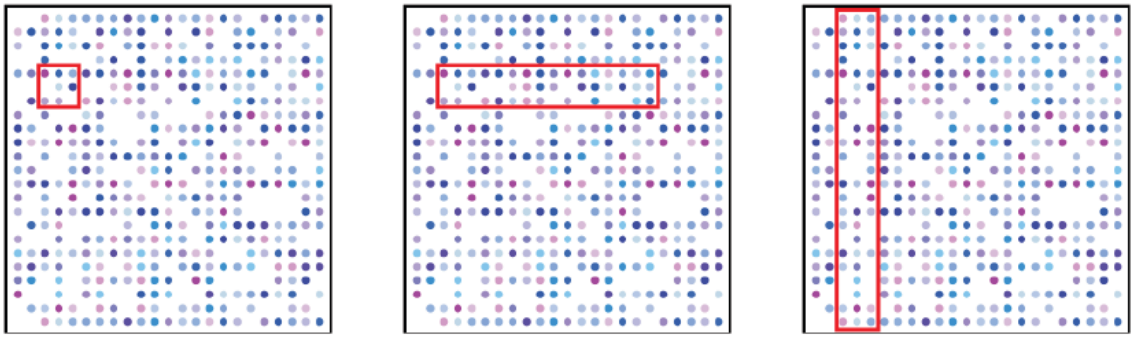
Table 2. Topic Summarization Produced by Anseri on the 2004 Sentence Level Data with Default Settings

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| final, sale, commission, report, july, europe | praise, allah, blessing, iii, operation, successful, fate, oppressors | jihad, qaida, bosnian, involvement, conflicts, afghan, network, laden | globalterroralert, info, www, http, evan, kohlmann, iraq, movement | north, america | lions, protect, heroic, ramadi, mosul, minus, kill, managed | tawheed, wal, abu, kidnappers, purchased, italian, hostages, yesterday | enemy, losses, media, reported, international, town, suffer, observed |

Now that the default performance has been identified, the analyst needs to tune the parameters to the particular dataset being analyzed. To analyze the effects of the number of features per topic (nf) and the number of documents per topic (nd) commands the following exploration was used to build our intuition. Anseri was run 34 times using the default nd settings and varying nf from 5 to 50. For the values of nf =5 up to 30 the number of features was increased by one each time and four additional times at 35, 40, 45, and 50. The variation of the features that identified each topic becomes negligible after nf = 20. Next, Anseri was run 18 times holding nf at 20 and varying the nd parameter. For both the sentence data and file data nd was varied from 20–80 in steps of 20 documents per topic. Then nd was varied using only the sentence data from 100–5000 documents per topic. For nd from 100–1000 steps of 100 documents per topic were used and steps of 1000 documents per topic were used beyond this point. Similar to the topic variation produced when changing nf, the topic results were roughly constant for nd greater than 200 documents per topic. The findings from this case study were similar to those provided by a white paper written by the developers of Anseri.

In a SumUp, LLC published technical report (Godbehere, 2016) it was found that when varying the nf and nd parameters, on a different corpus, the following held true. Figure 8 shows a graphical interpretation of the term/document matrix and the relationship of the topics identified and the relative size of the nf and nd parameters. When both nf and nd are small, Figure 8a shows that Anseri is able to find small clusters

of related words in the corpus. When nf is large and nd is small, Figure 8b shows that Anseri identifies larger groups of related words among similar documents. Finally, when nf is small and nd is large, Figure 8c shows that Anseri is able to identify more global patterns among the documents in the corpus (Godbehere, 2016). The results of Case Study 1 verify that Anseri is behaving in a way that is to be expected and that Anseri will provide reasonable results with the IS corpus.




The red boxes represent the selection of features and documents that make up a topic in the term/document matrix. (a): Both nf and nd small. (b): nf large, nd small. (c): nf small, nd large.

Figure 8. Graphical Representation of the Term/Document Matrix and the Effects of Different nf and nd Values. Source: Godbehere (2016)

Case Study 1 has tuned the nf and nd parameters and shown that the settings nf = 20 nd = 200 appear to yield consistent results from this corpus. Increasing the nf and nd parameters beyond this point does not improve the level of topic understanding and decreases the computing performance. Anseri will be run using these settings for the remainder of this study. Table 3 shows the performance of Anseri on the 2004 sentence level data using the recommended parameter settings of nf = 20 and nd = 200. Each column of the table represents the feature pattern that defines the topic. For the IS corpus it is much easier for the analyst to summarize the topics when using these settings.

Table 3. Anseri Performance on 2004 Sentence Level Data When Using the Recommended Settings of nf = 20 nd = 200.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| final, sale, commission, report, july, europe, north, america, september | jihad, qaida, bosnian, involvement, conflicts, afghan, network, laden, bin, allahu, akhbar, honor, messenger, evan, kohlmann, wing, military, mujahideen, committee, mesopotamia | allah, praise, blessing, iii, successful, blessings, operation, fate, oppressors, today, discover, escape, noon, single, accept, brother, place, power, strength, preached | globalterroralert, info, www, http, tawheed, wal, movement, zarqawi, musab, abu, communiqué, iraq, merciful, gracious, responsibility, august, apostates, october, benevolent, support | city, lions, islamic, nation, mosul, protect, heroic, accomplished, heroes, battle, fallujah, troops, baghdad, invading, outskirts, fifty, ramadi, killed, completely, forces | family, prayers, prophet, followers, friends, left, comfort, wealth, security, continue, announcement, fool, missed, judgement, british, arrow, arrows, message, hakim | vehicles, killing, destruction, inside, armored, fuel, american, destroyed, district, twelve, totally, turkish, tankers, containers, ghazaliya, convoy, americans, eighteen, afterward, consisting | enemy, losses, media, town, reported, international, suffer, observed, prison, arrived, battles, frontline, gentlemanly, enter, continuous, subject, airstrikes, dared, loss, acknowledged |

The first topic identified in Table 3 is representative of an advertisement for the sale of the September 11, 2001 commission that was printed on the documents in this directory. Topic 2 is about jihad and the organization's involvement in activities in Bosnia, Iraq, and Afghanistan. Topic 3 speaks to the narrative of IS fighting against oppressors in accordance with their religious beliefs. Topic 4 is a mix of web addresses and the introduction of Zarqawi who is a founder of IS. Topics 5 and 6 praise the fighters that have joined the movement and topics 7 and 8 discuss operations conducted during the year 2004. The operations discussed in topics 7 and 8 seem more current compared to the operations discussed in topic 2 implying that the documents that generated topics 7 and 8 could provide more recent information to the analyst.

This case study also revealed some interesting behavior regarding the singular values used to identify each topic. Next we will discuss the implications of the singular values that Anseri produces while it generates the topics. The singular values can provide insight into the prominence of topics within each corpus. The singular values can be used for a comparative assessment of topics within one directory.

### 1. Singular Values

The singular values differ when sentence level data is used compared to when the file level data is used to generate topics. This is a product of the normalization of the term/document matrix. When working with the data it was noted that the singular values were always greater when working with the sentence level data. Figure 9 shows a comparison between the singular values associated with the file level data and the sentence level data. According to A. Godbehere, a developer at SumUp LLC, this phenomena can be attributed to the numerical representation of the data in the term/document matrix (personal communication, March 30, 2016). Each row of the matrix is normalized to have the same overall weight and the normalized weights associated with the sentence level data are higher in general due to the size of the matrix. A similar effect is observed as the nd parameter is increased.


The singular values are consistently higher when analyzing the sentence level data versus the file level data. Each point represents one topic generated from the directory containing that year.

Figure 9. Singular Values Produced During the SVD Factorization of the File Level versus Sentence Level Term/Document Matrices.

The singular values give us an indication of how much a topic is discussed compared to the other topics returned by Anseri. Using the default settings, Figure 10 shows how the singular values change when discovering the top 50 topics in the entire corpus. Using the sentence level database for the entire corpus shows that while there exists some undulation, the general trend is a decrease in the singular values as the topics are discovered. The absolute magnitude of the singular values is a product of the size of the matrix whereas the relative difference between the singular values can indicate the degree to which topic 1 is more prominent than topic 2. The singular values do not decrease monotonically due to the deflation process. As each rank-1 approximation of the term/document matrix is calculated, the feature columns used for that topic are removed from the term/document matrix and renormalized.

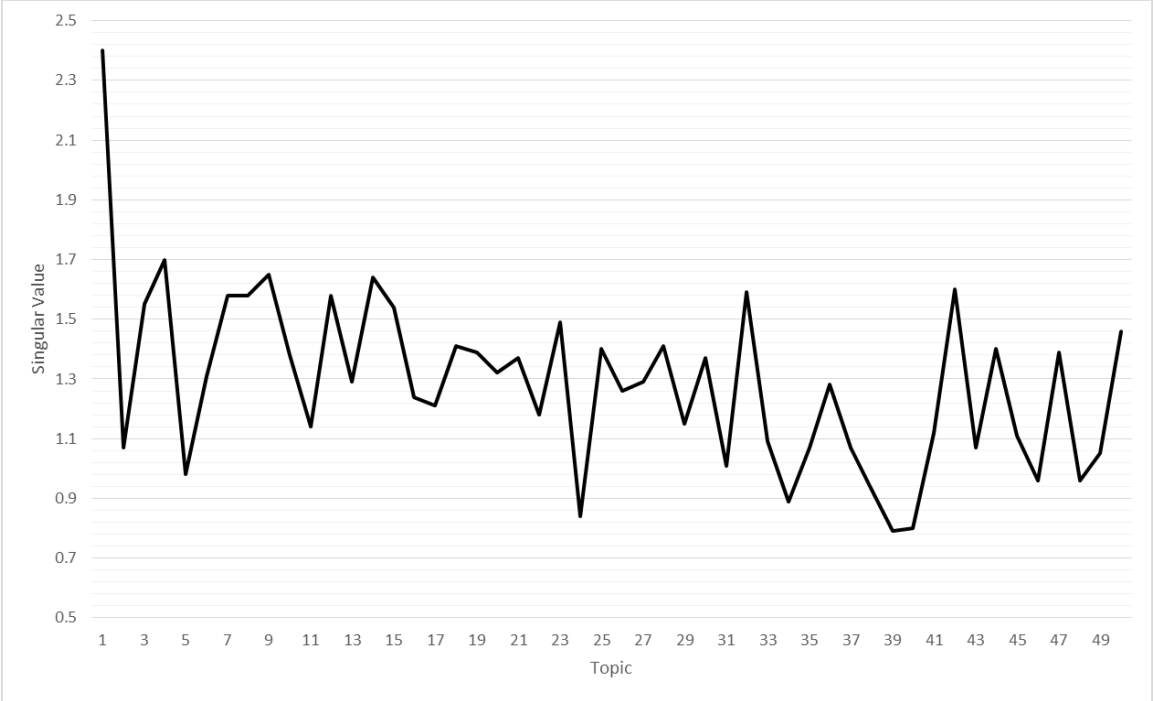



Figure 10. Singular Values Associated with First 50 Topics in the Entire Corpus.

To demonstrate the intuition gained from the relative changes in the singular values, consider the 2004 directory within the corpus. This corpus has a wide range in the singular values corresponding to the top eight topics and demonstrates a near monotonic decrease in singular values. Figure 11 shows the progression of singular values and one can infer that topic 1, with a singular value of 1.36, is much more prominent than topic 7 whose singular value is 0.35. The range of singular values indicates that the 2004 directory discusses a diverse range of topics. If there was not a great deal of topic variation the singular values would remain closer to constant.

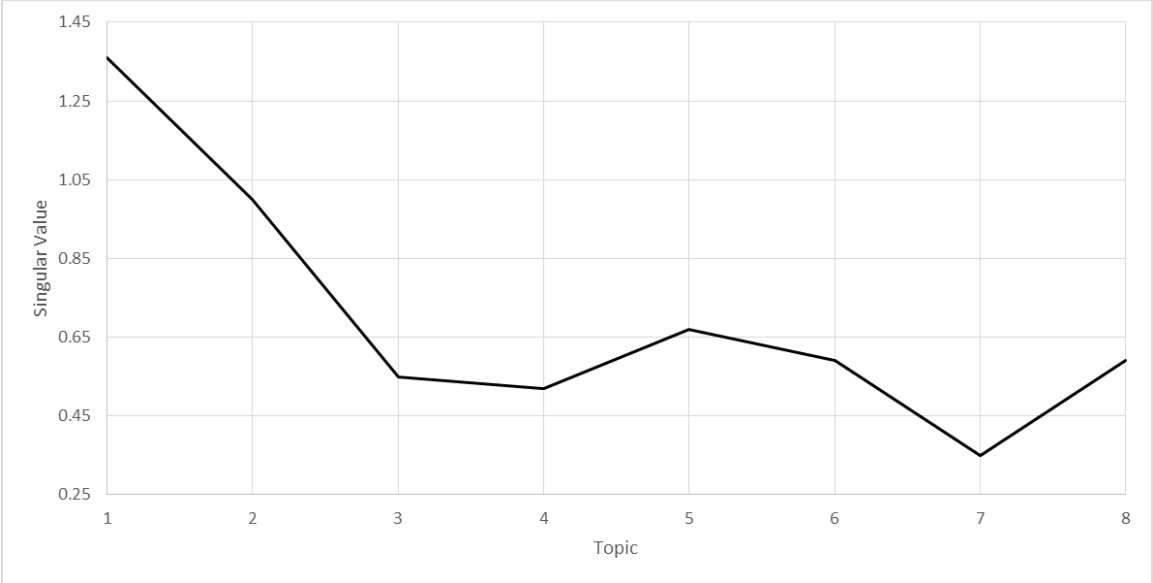


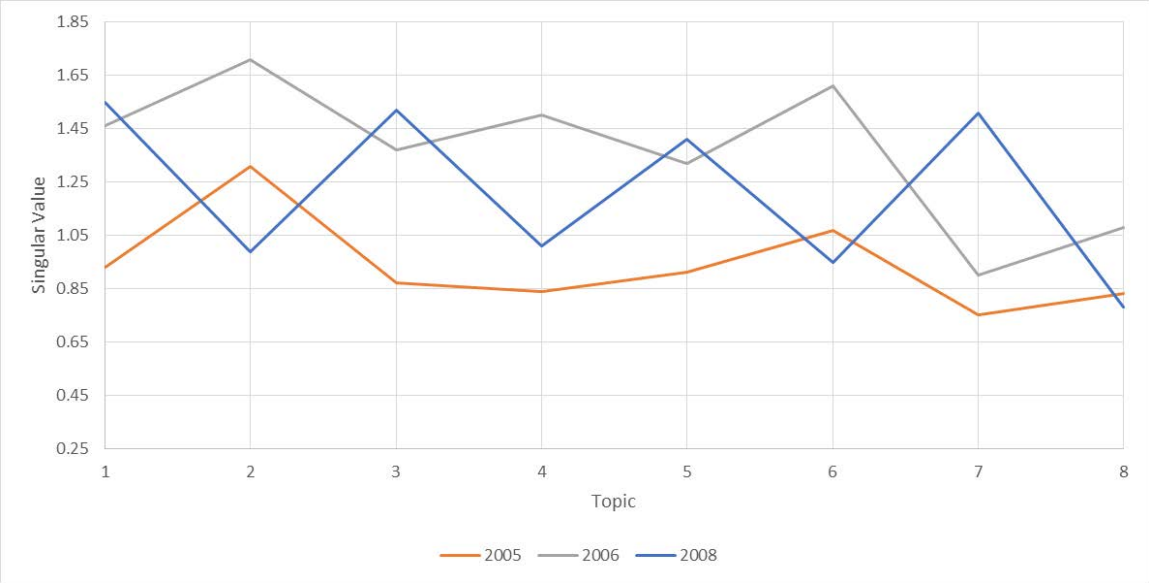

Figure 11. Singular Values Associated with the First Eight Topics from the 2004 Sentence Level Term/Document Matrix.

Table 4 shows features used to produce the top eight topics from 2004 that correspond to the series of topics graphed in Figure 11. Topic 1 is the content of an ad that is placed on most of the documents in the directory. Though the singular value for topic 1 is large it can be discarded as unhelpful by the analyst. Topic 2 appears to be a celebration of successful operations and with a singular value of 1.00 it is almost twice as prominent as topic 3, which appears to be focused on historical actions.

Table 4. First Eight Topics with Singular Values from the 2004 Sentence Level Term/Document Matrix.

| Topic | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Singular Value | 1.36 | 1.00 | 0.55 | 0.52 | 0.67 | 0.59 | 0.35 | 0.59 |
| 2004 | final, sale, commission, report, july, europe | praise, allah, blessing, iii, operation, successful, fate, oppressors | jihad, qaida, bosnian, involvement, conflicts, afghan, network, laden | globalterroralert, info, www, http, evan, kohlmann, iraq, movement | north, america | lions, protect, heroic, ramadi, mosul, minus, kill, managed | tawheed, wal, abu, kidnappers, purchased, italian, hostages, yesterday | enemy, losses, media, reported, international, town, suffer, observed |

To contrast a directory with a large variation of topics with some that do not vary greatly, we look at the singular values for three of the years that have a more monotonous topic theme. Figure 12 shows a graph of the singular values for three directories that have little variation in singular values. This could imply that the focus of the documents in these directories is similar. This could also show that no one topic is discussed more than any other topic. If every topic discussed in the directory received an equal amount of coverage than the range of singular values would be small.




The singular values for the top eight topics in a directory that has a relatively monotonous topic theme do not decrease in the same way the singular values decrease when there are a few prominent topics.

Figure 12. Singular Values Associated with the First Eight Topics from the 2005, 2006, and 2008 Sentence Level Term/Document Matrices.

The singular values can quantitatively rank the topics discovered in the corpus. The relative magnitude of the singular values among a set of topics offers insight into the prominence of the topics as well as the dominating nature of the topics. However, the singular values alone are not enough to fully understand the relevance of the topic. For an understanding of the relevance of the topic there needs to be further research by the analyst by conducting an in-depth reading of the material that produced each topic. After

the initial analysis is conducted by Anseri the analyst can conduct a more narrowly focused reading of the material vice attempting to read and analyze the entire corpus.

## E. SUMMARY

This case study verified that Anseri is performing in a manner that is to be expected with the corpus it was presented. The fact that results from Case Study 1 match those presented by the developer verifies that Anseri is interacting correctly with the IS corpus. This chapter has discussed the capabilities of Anseri on a technical level that is transparent to the end user. Once the system is tuned to the corpus properly the analyst can learn a lot about the subject matter included in the corpus without having to carefully read every document. The cases discussed in Chapter IV will offer a more qualitative discussion of how an analyst will use and benefit from Anseri.

# IV. INTELLIGENCE COMMUNITY IMPLICATIONS OF ANSERI

This chapter provides a qualitative discussion of the benefits of Anseri in the hands of a trained intelligence analyst. No software will replace the need to have a skilled analyst produce a comprehensive intelligence product for a decision maker. The nature of military operations requires the judgment of trained individuals to execute. Even as technology continues to reduce the need for placing people in harm's way, there remains a trained individual to ensure the machine is achieving the proper results. The results presented in Chapter III show that with no analyst to interpret the results Anseri is incapable of producing an intelligence product. The discussion provided in this chapter will connect the dots.

## A. WORKING WITH ANSERI AS AN ANALYST

An analyst that has Anseri may approach research in the following manner. After collecting a large amount of background materials and compiling a corpus the analyst must now attempt to understand it. First, the analyst should run the data through Anseri on the default settings. This will gain basic insights as to the content of the directory. In the case of the IS corpus we can see that there are some phrases that are used throughout the corpus that do not provide substantial information about the thematic nature of the corpus. The phrases "God is great" and "Final sale of the…" are what I will refer to as bumper stickers. The phrase "God is great" is used to start and end most communications between Muslim extremists and it does not tell us anything about the content of the speech. The skilled analyst will dismiss topics that are produced from bumper stickers and focus analysis on the more substantive results.

Once Anseri has been run using the default settings, tune the parameter settings as in Chapter III until the desired results are attained. For this particular corpus nf = 20 and nd = 200 are effective and efficient settings to gain an understanding of the documents. Once the analyst has reviewed the results of the initial topic analysis and properly tuned the parameters, they can now start to dig deeper into the topics that are relevant to the PIRs.

## B. CASE STUDIES: ANALYST BENEFITS OF ANSERI

Chapter III examined the technical aspects of Anseri and built an understanding of how this text analytics software is able to quickly and accurately understand a large corpus without requiring analysts to read every word on every page. The ability to read and understand large volume corpora is important in the effort to streamline the research and analysis portions of intelligence work. The case studies in this chapter will show the thought process of an analyst that is using the Anseri software and what insights can be gained from the results.

### 1. Case Study 2: Topics Changing over Time

An interesting result from background research is often a trend analysis of messaging over time. This case study gains insight into the analyst's ability to use Anseri to analyze the topic trends through the time dimension. An example of Anseri output is provided in Table 5 and the following discussion of trends will use this table as a reference. The Analyst starts with the corpus and minimal preprocessing to conduct this case study. Once the documents are collected into the corpus they are sorted and separated into directories according to the year of release. After the data is segregated properly, the directories are converted to term/document matrices and input to Anseri as discussed in Chapter III. Anseri is then run on the sentence level data to generate the top eight topics using 20 features per topic and 200 documents per topic.

The analyst now has the information needed to begin identifying trends in the corpus. The features associated with the topics produced each year reveals trends in IS messaging during the time period covered in this corpus. It is important to remember that this corpus is collected from IS sponsored sources and therefore do not have any outside perceptions about the events discussed. That is to say that these messages are a collection of raw data without prior analysis of their meaning. The views expressed in the corpus may differ from previous opinions about the impacts of events because of the time that has passed and analysis that has been conducted since these events took place. The intelligence analyst may have preconceived notions about the content of the corpus but it is important to understand that Anseri can only analyze the corpus at face value.

Table 5. Sentence Level Data Performance at Recommended Settings

| Topic | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Year | 2004 | final, sale, commission, report, july, europe, north, america, september | jihad, qaida, bosnian, involvement, conflicts, afghan, network, laden, bin, allahu, akhbar, honor, messenger, evan, kohlmann, wing, military, mujahideen, committee, mesopotamia | allah, praise, blessing, iii, successful, blessings, operation, fate, oppressors, today, discover, escape, noon, single, accept, brother, place, power, strength, preached | globalterroralert, info, www, http, tawheed, wal, movement, zarqawi, musab, abu, communiqué, iraq, merciful, gracious, responsibility, august, apostates, october, benevolent, support | city, lions, islamic, nation, mosul, protect, heroic, accomplished, heroes, battle, fallujah, troops, baghdad, invading, outskirts, fifty, ramadi, killed, completely, forces | family, prayers, prophet, followers, friends, left, comfort, wealth, security, continue, announcement, fool, missed, judgement, british, arrow, arrows, message, hakim | vehicles, killing, destruction, inside, armored, fuel, american, destroyed, district, twelve, totally, turkish, tankers, containers, ghazaliya, convoy, americans, eighteen, afterward, consisting | enemy, losses, media, town, reported, international, suffer, observed, prison, arrived, battles, frontline, gentlemanly, enter, continuous, subject, airstrikes, dared, loss, acknowledged |
| | 2005 | america, report, commission, sale, final, north, july, europe, talented, exceptionally, appetite, adjustments, reporting, stating, perpetrating, losing, misinformation, exaggeration, oppressive, unlawful | allah, praise, blessing, enemies, religion, bless, mercy, grant, victory, guide, accept, path, brothers, faith, martyrdom, bring, fight, good, harm, reward | jihad, qaida, bosnian, committee, mesopotamia, allahu, akhbar, afghan, network, conflicts, involvement, laden, bin, evan, kohlmann, honor, messenger, mujahideen, january, wing | abu, hamza, anas, musab, zubair, zarqawi, tarek, shaykh, ghraib, told, yousef, umair, brother, hariri, nour, martyr, iraqi, mujahid, protect, asked | globalterroralert, http, www, info, prayers, family, communiqué, gracious, merciful, mohammed, statement, followers, almighty, supporters, grants, behalf, responsibility, claiming, righteous, educated | attack, place, shaaban, enemy, september, friday, baghdad, launched, afar, tel, armed, campaign, revenge, response, sunnis, neighborhood, factions, symbols, initiate, morning | nation, islam, people, islamic, infidels, victorious, prepare, muslim, sacrifice, love, men, efforts, behold, continue, blood, order, hearts, desperately, felt, satisfaction | killed, result, wounded, soldiers, inside, humvee, americans, vehicle, peshmerga, destroyed, forget, dozens, street, crusaders, apostates, kill, damaged, haifa, replied, fighters |
| | 2006 | gratitude, praise, god, accurate, firing, annihilation, hits | commission, media, council, shura, mujahidin, iraq, attributed, statements | killing, operation, wounding, board, destroying, vehicle, destruction, hummer, complete, led, injuring, apostates, occupants, inside, vehicles, number, disabling, hummers, aboard, car | mahdi, army, dajal, ramadan, october, members, area, assassinated, liquidated, district, destroyed, killed, monday, thursday, member, sunday, eliminated, friday, tuesday, zamiyah | explosive, detonated, crusaders, september, sha'ban, region, patrol, crusader, device, brothers, foot, saturday, rabi, august, grace, armored, minesweeper, thani, police, wednesday | compassionate, merciful, translation, statement | hit, direct, precise, target, rockets, precision, great, scored, effective, attack, rocket, assistance, sustained, severe, precisely, scoring, strikes, bahani, confused, flames | november, dah, dhu, shawwal, sniped, mafriq, cross, soldier, worshipper, pagan, guard, fallujah, mim, national, apostate, policeman, ramadi, post, afternoon, taji |
| | 2007 | gratitude, praise, god | ministry, iraq, islamic, state, official, source, signed, informatio | june, awwal, jumada, tuesday, monday, morning, afternoon, place | thani, rabi, april, wednesday, saturday, sunday, noon, evening, happened | effective, hit, direct, earthquake, strong | hijjah, january, dhu, thursday, friday | national, pagan, guard, killed, soldier, soldiers, sniped, area, mujahidin, sniping, region, belonging, road, headquarters, sniper, checkpoint, fire, gunfire, launched, wounded | safar, march, occurred, diyyah, times, mustansariyyah, durah, highway |
| | 2008 | gratitude, praise, god, blessings, return | ministry, iraq, state, islamic, fajr, center, media, attributed, copied, statement, source, official, statements, lions, signed, dated, report | explosive, detonated, device, killing, patrol, police, area, vehicle, american, apostate, october, wounding, inside, hummer, neighborhood, destroying, injuring, foot, september, ramadan | dhu, december, hijjah, dah, saturday, wednesday, hijja, friday, sunday, occurred, november, monday, tuesday, thursday, published, january | awwal, jumada, rabia, march, june, april, rabi, biweekly, period, noon, fired, rockets, night, correspondent, green, zone, jarf, mukhtar, salim, sanad | peace, prophet, prayers, companions, muhammad, family, followers, lord, creation, house, hadith, entire, messenger, imam, mercy, praises, chosen, mohammad, asked, honorable | guard, national, pagan, sniped, member, killed, thani, soldiers, post, yarmuk, headquarters, tanak, destroyed, soldier, shot, base, july, liquidated, checkpoint, dawrat | side, left, received, mujahidin, peshmerga, mosul, detonating, disabled, city, rajab, intisar, element, quarter, sina, elements, mortars, firing, shelled, launched, attack |
| | 2009 | god, praise, gratitude, translation, compassionate, merciful, statement, grace, creation, lord, great, guided, message, audio, prophets, supplicate, unjust, aggression, poetic, verses | detonated, explosive, pagan, charge, national, neighborhood, guard, hummer, february, belonging, vehicle, patrol, march, guards, crusader, foot, april, police, area, july | media, center, fajr, islamic, ministry, iraq, state, source, attributed, copied, official, published, undated, furqan, establishment, signed, dated, disseminated, production, produced | wounded, killed, explosion, board, number, policemen, unidentified, result, destroyed, disabled, car, clashes, attack, undefined, totally, elements, completely, damaged, bodyguards, driver | killing, wounding, led, destruction, checkpoint, clash, apostates, members, army, onboard, october, shawwal, attacked, vehicles, weapons, light, destroying, inside, thermal, operation | koranic, verse, allah, believers, messenger, hypocrites, belongs, honor, munafiqun, partial, honour, almighty, nisa, tawbah, baqarah, glory, hujurat, anfal, verily, planners | prayers, peace, prophet, companions, muhammad, family, household, forget, mothers, mohammad, entire, champion, bereaved, mercy, thing, view, earth, click, epic, seal | apostate, assassination, rajab, assassinated, awakening, targeted, council, grenade, detachments, hand, november, dhu, august, bab, member, sha'ban, councils, sniped, policeman, squads |
| | 2010 | god, praise, gratitude, creation, lord, messenger, success, fear, aim, compassionate, great, merciful, prayers, granted, testify, guidance, religion, banner, muhammad, support | media, state, islamic, iraq, fajr, center, ministry, source, statement, disseminated, undated, attributed, signed, furqan, establishment, production, message, produced, copied, published | wounded, killed, vehicle, board, destroyed, explosion, soldiers, disabled, result, army, attack, completely, vehicles, members, number, personnel, instantly, unidentified, motorcade, hummer | killing, wounding, destroying, elements, led, police, thermal, checkpoint, officer, apostates, grenades, severely, neighborhood, patrol, clash, disabling, apostate, pagan, target, grenade | explosive, area, detonated, charges, charge, targeting, planted, jumada, thani, belonging, voting, detonating, simultaneously, december, centers, opening, gathering, awakening, hypocrisy, dajjal | koranic, verse, allah, partial, believers, belongs, hypocrites, munafiqun, honor, imran, honour, tawbah, nisa, baqarah, yusuf, anfal, verses, hath, hashr, raf | peace, prophet, companions, household, family, prayer, master, blessings, mercy, hadith, mujahidin, prophets, imam, mothers, leader, depart, promised, study, verily, ummah | text, onscreen, reads, dead, rising, abu, rejectionist, execution, screen, killers, muhajir, prisoner, remain, caption, finance, farewell, voice, rejectionists, allawi, muqtada |
| | 2011 | wounding, attack, killing, vehicle, destroying, board, number, members, soldiers, occupants, severely, disabling, checkpoint, escorts, policemen, unidentified, explosion, unspecified, driver, guards | god, praise, gratitude, lord, creation, compassionate, merciful, guidance, grace, almighty, aid, support, blessing, prayer, stand, fear, steadfast, great, satisfied, shari | explosive, device, detonated, army, area, belonging, patrol, pagan, jumada, thani, police, apostate, august, hijjah, safavid, ramadan, dhu, january, rajab, neighborhood | place, october, qida, september, shawwal, november, dah, muharram, qidah, shawwal, mahmudiyyah, sayyid, attaining, fields, proper, resources, wealth, wounde, glory, abdallah | state, islamic, iraq, media, ministry, center, fajr, source, statement, operations, military, documented, glorify, mujahidin, governorate, furqan, establishment, production, war, disseminated | killed, wounded, destroyed, instantly, completely, result, disabled, unknown, inside, commandos, member, spot, dozens, hummer, major, weapons, companions, building, force, tower | peace, prophet, muhammad, prayers, family, household, blessings, mercy, messenger, imam, entire, hadith, believers, master, affirmed, book, sunnah, honest, followers, truth | koranic, verse, allah, partial, belongs, munafiqun, hypocrites, honor, tawbah, baqarah, honour, imran, wait, nisa, idah, apostle, unbelievers, anfal, remember, restrain |
| | 2012 | number, wounding, killing, attack, servicemen, unspecified, unidentified, soldiers, large, indefinite, escorts, policemen, bodyguards | explosive, device, safavid, army, area, detonated, belonging, vehicle, detonating, patrol, hummer, mosul, police, ramadan, shawwal, september, july, august, neighborhood, azim | god, guide, supplicate, enemy, harming, punish, enemies, reward, harmful, attacks, harm, empower, aim, direct, witness, great, accurate, partner, bear, precise | place, june, rajab, dhu, hijjah, sha'aban, november, october, dah, rresponding, esponding, corr, imagine, thought, safe, friends | wounded, killed, destroyed, inside, soldier, completely, result, severely, members, companions, instantly, including, dozens, wretches, detail, protection, vehicles, seventy, twenty, driver | destroying, board, explosion, occupants, disabling, minesweeper, elements, totally, personnel, detonation, amputating, policeman, den, headquarters, site, leg, office, exploded, burning, injury | security, detachment, targeted, criminal, eliminated, weapons, silenced, dajjal, ministry, officer, member, spoils, interior, apostate, intelligence, carried, sha'ban, silencer, assassinated, called | muharram, december, safar, thi, element, capturing, eliminating, mber, dece, cember, decemb, correspondin, latifiyah, rocket, shelter, mahmudiyah, garage, humaydiyah, machinegun, marketplace |
| | 2013 | wounding, killing, attack, servicemen, number, board, unidentified, serviceman, personnel, destroying, scores, bodyguards, policemen, dozens, fourth, elements, tanker, hummer, volunteers, lahibi | detonated, explosive, device, safavid, safar, army, patrol, december, area, mosul, january, police, vehicle, rabi, awwal, neighborhood, foot, belonging, district, federal | killed, wounded, spot, result, instantly, destroyed, soldiers, inside, seized, unspecified, blast, spoils, people, war, incinerated, guards, unknown, large, twenty, gun | god, punish, guide, aim, precision, praise, enemies, reward, great, supplicate, gratitude, ordained, harmful, attacks, accurate, destined, survival, harm, guidance, discovered | place, dhu, hijjah, november, october, february, afar, scenario, anew, dah, published, occupants, orresponding, fierce, cowardice, lasted, reach, battle, day, homes | security, detachment, targeted, criminal, member, silenced, weapons, eliminated, detail, members, interior, ministry, sticky, detonating, apostate, muharram, hypocrisy, apostasy, awakening, intelligence | hit, direct, scored, shelling, precise, scoring, headquarters, mortar, missile, salvo, mortars, bombardment, shelled, inflicting, rising, smoke, columns, directly, shells, rounds | explosion, disabling, completely, protection, wounds, house, sections, hideout, officer, severe, injuries, experts, escort, motorcade, complex, vehicles, expert, escorting, temple, sever |

The first eight topics discovered by Anseri for each year directory in the corpus. The features in this spreadsheet allow an analyst to quickly build an understanding of the content of the corpus and focus reading efforts toward the directories that will provide the best answers for the PIRs.

In the early years of the corpus, the organization placed more emphasis on describing who they are and what they stand for. In 2004, the top eight topics appear to be focused on the explanation of IS in a celebratory nature. Topic 1 is a bumper sticker as discussed earlier but topics 2, 3, and 4 all have contain praise to Allah and the righteousness of their cause. The feature, "tawheed" (Tawheed wal Jihad) stands for monotheism in Jihad and which was one of the original names of IS going back to the 1990s (Whiteside, 2014). Topics 5 and 6 describe the brave fighters of their organization and thanks their families. It is not until topics 7 and 8 that there is mention of violence and operations against their enemies.

Chapter III, Figure 12 presented the singular values for the first eight topics in the years of 2005, 2006, and 2008. The conclusion was that the theme of these years was monotonous or the writers had given equal focus each topic in the first eight topics of these years. Each of these years shows a topic theme that is narrowly focused. The topics from 2005 continue to explain the organization and their relevance in the region. The topics for 2006 and 2008 begin to represent more operational activity and to discuss the results of attacks. In 2007 the term "Islamic State" starts to surface as a high level topic indicating that the name change is taking effect. That term begins to move lower on the list as the years pass because the organization becomes well known and the importance of self-definition falls behind claiming credit for destruction.

The later years of the corpus show a shift in messaging to nearly all operational summaries. From 2010–2013 the messaging is increasingly violent, but also straight forward and businesslike in nature of their description of operations. Chapter III, Figure 7 shows that the corpus is perceived to be overwhelmingly celebratory. It is not readily apparent by the results produced by Anseri that these documents are celebrating accomplishments, but the tone of the corpus can be determined through more expansive research.

Now that the analyst has evaluated the trends of topics as they pass through time returning to the corpus to continue research is essential to gain an understanding of the semantic nature of the writings. The insights gained from looking at the topics over time

is only a cursory understanding that is not in-depth enough to build an intelligence product. Anseri has provided the analyst with a great deal of focus for further studies but not a complete understanding of the content if the corpus.

### 2. Case Study 3: Author Sentiments

Finally, the corpus is segregated by those documents that have been labeled by author to better understand the thematic nature of each author's contribution to the corpus. This case study will also analyze the utility of the keyword search and file path display functions of Anseri. These functions can increase the efficiency of the analyst through a fast method to develop the understanding of the corpus. An analyst should bring some level of outside understanding to the analysis of the writings of these authors because having a rudimentary level understanding of background information for each author can scope the motivation for their writing.

There are four authors identified by the corpus: Abu Musab Zarqawi, Abu Omar al Baghdadi, Abu Hamza al Muhajir, and Ahman Zawahiri. Zarqawi was a militant Islamist from Jordan who formed al Tawhid wal Jihad in the 1990s which eventually became the Islamic State that is known today. Zarqawi is known as the leader that brought the insurgent fight against the United States in Iraq and was killed in 2006 (Whiteside, 2014). Zarqawi is the most prominent author in the corpus despite being killed early in the timeframe covered by the text. Baghdadi was selected to lead the Mujahideen Shura Council in 2006 and charged with the duty to build credibility for IS among Iraqis. Muhajir succeeded Zarqawi as the leader of IS after his death. Zawahiri succeeded Osama bin Laden after his death in 2011 (Whiteside, 2014). This small amount of biographical information shapes our expectations of what topics each author should cover in their directories. The first eight topics from each author's directory is provided in Table 6.

Table 6. Topic Summarization of Each Author Corpus.

| Topic | Baghdadi | Muhajir | Zarqawi | Zawahiri |
|---|---|---|---|---|
| 1 | god, praise, gratitude, creation, lord, grace, translation, compassionate, merciful, great, statement, fear, messenger, testify, muhammad, witness, banner, support, baghdadi, bear | praise, glory, god, creation, lord, incomplete, sentence, indistinct, satisfied, received, victories, aware, audio | allah, praise, blessing, religion, fear, victory, enemies, bless, grant, harm, almighty, mercy, reward, obey, swear, grace, good, path, fight, power | god, mercy, messenger, almighty, great, reward, good, praise, enemies, find, witness, bear, blessings, grace, seek, delay, muslims, guarantor, mothers, devotion |
| 2 | wounded, killed, explosion, board, vehicle, result, destroyed, number, unidentified, disabled, policemen, soldiers, guards, attack, army, completely, clashes, vehicles, undefined, members | peace, prophet, prayers, companions, muhammad, family, hadith, blessings, allah, messenger, household, prayer, master, mercy, followers, fight, pleased, muslims, sahih, mentioned | sale, commission, final, july, report, europe, north, america | state, islamic, iraq, video, history, narrator, media, accomplishments, formation, establishment, defend, announcing, talks, center, anniversary, established, clarify, concern, caliphate, cabinet |
| 3 | detonated, explosive, area, charge, pagan, guard, neighborhood, charges, police, patrol, national, march, apostate, april, july, hummer, targeting, rajab, foot, crusader | people, shortcomings, religion, human, islam, levant, patience, hide, advice, raise, thee, afraid, killed, living, fair, wise, weak, sound, disbelieving, revealed | god, lord, promise, praised, blessings, creation, seek, peace, support, sake, precious, compassionate, defeat, continue, chosen, render, mujahidin, rope, merciful, fast | people, mujahedeen, leads, scattering, scattered, afghan, justice, merit, denying, stances, kurdistan, representation, disengaged, ruling, result, islam, regime, governance, deceived, land |
| 4 | media, state, islamic, iraq, fajr, center, ministry, source, attributed, official, undated, published, copied, disseminated, establishment, furqan, production, produced, munafiqun, signed | state, islamic, iraq, ministry, media, source, fajr, center, official, statement, attributed, copied, banner, establishment, furqan, logo, production, graphic, harvest, signed | jihad, qaida, bosnian, allahu, akhbar, involvement, conflicts, network, afghan, laden, bin, mesopotamia, committee, evan, kohlmann, honor, messenger, mujahideen, wing, military | abu, umar, baghdadi, excerpt, osc, processed, audio, message, muhajir, bakr, shaykh, faraj, reads, hamzah, rasmi, contact, brothers, caption, statement, text |
| 5 | verse, koranic, allah, believers, partial, hypocrites, belongs, honor, tawbah, imran, baqarah, yusuf, affairs, anfal, full, raf, power, hashr, mankind, baqara | attack, killing, destroying, wounding, board, instantly, completely, elements, vehicle, damaging, bodyguards, detonated, explosive, burning, thani, inside, patrol, number, amputated, severely | abu, mus, yaman, anas, brothers, companions, musab, sheikh, brother, told, muhammad, afghanistan, ghraib, hamza, met, left, asked, ghadiyah, time, meeting | shia, attacks, ordinary, kill, leaders, ignorance, plans, forgiven, attack, subject, position, detailed, complicated, loss, befall, americans, matter, opinion, conflict, folk |
| 6 | people, war, acting, realize, percent, prevails, land, levant, jihad, elections, good, issue, time, knowledge, earth, rule, hurt, raise, call, thee | verse, koranic, partial, imran, tawbah, nisa, believers, munafiqun, almighty, belongs, hypocrites, honor, anfal, hashr, things, ahzab, faith, baqarah, power, hath | people, islam, laws, rule, sunni, follow, koran, defeated, beleaguered, truth, stable, meaning, order, blood, syrian, lived, quiet, remain, kind, saddam | place, meet, hope, public, battle, secure, residence, likewise, assistants, half, taking, battlefield, general, conducive, schism, behavior, penalty, grievous, harm, punish |
| 7 | wounding, killing, led, checkpoint, thermal, attacked, elements, grenades, clash, grenade, october, weapons, light, shawwal, safar, february, operation, destroying, apostates, destruction | text, onscreen, reads, abu, dead, rising, muhajir, umar, voice, prisoner, execution, remain, finance, zone, sisters, prisons, green, shaykh, appears, minister | enemy, fighting, muslims, stand, weakened, losses, weapons, reach, occupying, attack, wounded, numbers, attacking, find, ummah, envies, sea, control, lands, entered | koranic, verse, allah, partial, anfal, hajj, faith, verily, tawbah, planners, aid, remnants, hujurat, sincere, screenshot, zariyat, dhariyat, raf, decision, grandchildren |
| 8 | prophet, peace, prayers, companions, family, household, blessings, blessing, muslim, hadith, heard, pleased, leads, mercy, brother, unfair, lying, unfairly, medina, treated | gratitude, conquering, defeating, formation, announcement, express, saved, brink, stretches, pit, dissemination, brethren, fulfilled | iraq, eyes, jerusalem, invasion, zarqawi, awaken, interest, baghdadi, plan, land, rafidah, iraqis, experience, young, ahl, wanted, war, exist, northern, qaeda | support, popular, issue, strive, elaborate, strengthen, condition, lead, increase, striving, concession, laws, maintain, sharia, greate, hinting, traitor, fitnah, movement, technical |

Eight topics per author are produced using the sentence level data with 20 features per topic and 200 documents per topic.

The Baghdadi corpus has themes pertaining to attacks and people being wounded. The focus of his writing appears to be the people. He describes destruction and death, but he also writes with religious undertones. Knowing that he was appointed the head of the Mujahideen Shura Council and charged with building the credibility of IS with the Iraqis an analyst can draw the conclusion that his topic theme is to explain the use of violence from IS to the citizens affected by the violence. Without knowledge of the author these topics appear similar across each individual writer and it is essential for the analyst to read these topics with the understanding of the author and their intentions.

Muhajir and Zarqawi write in such a way as to call the people to arms. Zarqawi focuses on the enemy and the injustices endured by the people of the area. There is mention of Abu Ghraib, weakened Muslims, and Jerusalem. The features of these topics are inflammatory to the Muslim audience and show a strong base for recruiting other extremists. Muhajir continues this message as Zarqawi's successor and continues to recruit through displaying the strengths of IS. The language used in these two corpora appear to be rousing propaganda statements that carry the extremist message unlike Baghdadi.

The topics from Zawahiri are vaguer about the relation to IS. The position of Zawahiri during the timeframe of the corpus offers a unique insight to the thematic nature of his writings. As the second in command to al Qaeda at the time, he was used as the mouthpiece for the organization. Osama bin Laden was notoriously careful about his electronic communication while he was in hiding and Zawahiri had to assume the role of lead communicator. With this historical knowledge an analyst can read the topics produced from the Zawahiri corpus to be coming from a position of authority. There is a level of encouragement for the actions taken by IS from Zawahiri and yet he seems less concerned with recruiting others to their cause or building their credibility. There was a rift between Zarqawi's group of believers and Zawahiri's leadership that eventually manifested in the severing of ties between the two groups shortly after the collection of this corpus (Whiteside, 2014).

Anseri has provided the analyst with an overview of the thematic nature of each author and some of the key features that determine the topics in the corpus. Now the

analyst has an understanding of what the most prominent topics are for each author and more focused search of the key terms in the directory will build the understanding of the author's sentiment regarding various topics. The keyword search function of Anseri will provide the analyst a method for which to match the author's opinion to the various PIR that need to be answered for the decision maker.

### *a. Keyword Study*

Anseri was run on various author directories to determine their association to several keywords and keyword groups. Anseri returns a list of topics prioritized by the keywords searched by adding a greater weight to the features that match the keywords in the term/document matrix. The increased weight increases the singular values and effectively shuffles those associated topics to the top of the list. The keyword allows the analyst to focus on those documents that have information pertaining to the PIRs. The goal of Anseri in the hands of the analyst is to produce a greater understanding of the corpus with respect to the assigned intelligence requirements. The use of keyword search is not recommended for a naïve approach to topic discovery and analysis. Once there is an initial understanding of the thematic nature of the corpus the keyword search can refine the analysis.

To illustrate the futility of the naïve keyword search each of the author directories were analyzed using the search terms in Table 7. The first eight topics returned by Anseri are vaguely related to the keyword prescribed in the input parameters. Table 7 does not include any search terms that are tuned to the biographical information discussed about the authors and while the results show that the author is discussing these topics there is not a focus toward an analytical end.

Table 7. Common Naïve Keyword Search Terms that Yield Results from Each Author in the Corpus.

| Baghdadi | Muhajir | Zarqawi | Zawahiri |
|---|---|---|---|
| America american killed | America american killed | America american killed | America american killed |
| bin laden | bin laden | bin laden | bin laden |
| bush | bush | bush | bush |
| death jihad | death jihad | death jihad | death jihad |
| europe | fatwa | europe | europe |
| islam infidel | islam infidel | fatwa | islam infidel |
| islam koran | islam koran | islam infidel | islam koran |
| islamic state leadership | islamic state leadership | islam koran | islamic state leadership |
| jihad | jihad | islamic state leadership | jihad |
| kill enemy | kill enemy | jihad | kill enemy |
| muhajir | baghdadi | kill enemy | baghdadi |
| zarqawi | zarqawi | peace love | muhajir |
| | zawahiri | baghdadi | zarqawi |
| | | muhajir | |
| | | zawahiri | |

The knowledge of the author's topic analysis can lead us to more precise keyword searches. The Zarqawi directory holds the greatest number of documents so for the remainder of this keyword exploration we will continue to analyze the output from this directory. Table 8 provides a sample of the topics discovered in response to the keyword "jihad" applied to the Zarqawi directory. Jihad was one of the terms in Table 7 used during the naïve exploration of this functionality across each author's corpus. The results of the keyword search show each topic has a religious theme, portraying the jihadist as the victor and the Jews and Christians as infidels. Zarqawi appears to use this religious message to recruit fighters to his cause similar to the results observed in Table 6. This unfocused use of the keyword function does not increase the analyst's understanding of the author's opinions or overarching message.

Table 8. Naïve Keyword Search Results from the Zarqawi Sentence Level Data.

| Search Term: jihad | |
|---|---|
| Topic | Features |
| 1 | qaida, bosnian, europe, akhbar, allahu, involvement, conflicts, network, laden, afghan, bin, evan, kohlmann, honor, messenger, mujahideen, mesopotamia, committee, january, august |
| 2 | allah, religion, fear, enemies, continue, brothers, dedicated, attacks, duty, job, fight, bless, almighty, land, blessing, congratulate, suri, deal, ready, word |
| 3 | abu, yaman, good, afghanistan, hamzah, iraq, ummah, insist, russians, hamza, birth, doctrine, qur, errors, clear, sunnah, witnessed, killed, sheikh, setback |
| 4 | movement, wal, tawheed, globalterroralert, info, http, www, musab, communiqué, zarqawi, september, martyrs, merciful, lions, gracious, praise, forward, friends, support, brigade |
| 5 | god, sake, carry, shaykh, islam, highly, jews, people, uniting, shaking, efforts, brother, obedience, increasing, hands, disavowed, unite, shari, defend, unbeliever |
| 6 | islamic, nation, enable, leadership, uproot, chase, threatened, caliphate, society, reestablish, experience, hijaz, issues, action, fields, long, favored, behalf, heroes, sun |
| 7 | military, commander, idat, adl, saif, wing, biography, division, declare, leader, lord, original, qualified, enforcement, agencies, law, translation, slaughtering, document, jihadist |
| 8 | infidels, working, christianity, implant, bury, life, fronts, restrictions, situation, nature, lived, solidarity, times, unity, permission, met, order, thorn, anbar, apostates |

Jihad is a term that is commonly used throughout the corpus and using it as a keyword search term does not increase the analyst's understanding of the material included in the corpus.

When used correctly, the keyword functionality will allow analysts to compare the topics in the corpus with the PIR set by the commander. The correct use of this function leverages the knowledge gained from the initial runs of Anseri described previously. A keyword search conducted without prior understanding of the thematic nature of the corpus yields results that may not be helpful. Table 9 shows the kind of results that can be produced by an intelligent keyword search. The Zarqawi corpus largely appeals to the extremist themes of IS propaganda and leveraging that knowledge

about the corpus an additional list of keywords was tailored to the Zarqawi corpus. Table 9 shows the topic analysis results of a keyword search using the terms "west, american, english, aggressors." These results will focus the analysis of the Zarqawi corpus based on his attitudes towards the western nations operating in Iraq.

Table 9. Focused Keyword Search Results from the Zarqawi Sentence Level Data.

| Search Term: west american english aggressors | |
|---|---|
| Topic | Features |
| 1 | jihad, resist, time, iraq, qaida, tawheed, mesopotamia, committee, globalterroralert, wal, evan, kohlmann, bosnian, europe, movement, enemy, lion, october, communiqué, info |
| 2 | vehicle, armored, killing, bomb, detonated, infidels, destroy, killed, causing, allah, operation, targeted, board, destroyed, january, inside, roadside, passing, taamim, set |
| 3 | baghdad, forces, street, haifa, clashes, house, raid, pagan, captured, guard, attack, executed, convoy, martyrdom, targeting, capital, located, ghraib, occurred, airport |
| 4 | district, attacked, infantry, rpgs, muatasim, shuhadaa, martyrs, organization, badr, ghadriyah, troops, falcon, grouping, friday, agents, monday, battalion, samarra, morning, lions |
| 5 | abu, anas, god, spoke, late, decreed, fluent, zarqawi, musab, commander, commenced, claiming, shaykh, execution, hostage, today, assistant, current, crusaders, arrested |
| 6 | army, bankrupt, liar, accursed, rare, begun, disappear, presence, cities, secret, worst, shi, lines, rear, proxy, pass, brigades, spies, fighting, mercenaries |
| 7 | base, city, ramadi, siege, nearby, revolted, mujahideen, bombarded, mortars, azamiyah, hawn, mosul, battles, fought, support, heroes, apostates, spread, strikes, nowadays |
| 8 | ahl, sunnah, started, tanks, rafidah, pact, empty, area, difference, bait, east, shiites, reasons, enemies, north, moving, islamic, south, establishment, protecting |

This is a more focused table of the keyword results based on what we know about the author and what we would like to gain from them.

The topics produced from this keyword search are centered on the fighting and maintain the propagandized versions of the events that are taking place. The next step for

the analyst is to return to the source documents in the corpus to investigate the true nature of these topics.

### *b. File Path Display*

The overall goal for Anseri is to improve the workflow of the analyst. If text analytics is going to reduce the amount of time spent on "Processing and Exploitation" then it needs to provide a simple method to retrieve the source material that was analyzed to produce the topic analysis. Without the source material the analyst cannot build a comprehensive picture of the thematic nature of the corpus and the decision maker cannot make decisions based on the output of Anseri alone.

The output of Anseri can be manipulated to display the file path associated with the content that generates the topics. With this functionality the analyst can return to the original text that produced the topics and begin the in-depth research required to produce an intelligence product. The benefit of this file path display is the enhanced speed of the research process. Anseri is able to read the entire corpus and offer a basic understanding of the topic themes in seconds. Then the analyst can conduct focused analysis of the documents that are relevant to the PIRs without wasting effort on irrelevant documents.

Anseri increases the speed in which an analyst can process raw data into an intelligible product. The most important job of an intelligence analyst is communicating the results of their analysis to the decision maker. The decision makers rely on the intelligence provided to them to make sound decisions in an uncertain environment. The more time that can be committed to analyzing, synthesizing, and presenting that valuable intelligence to the decision maker improves the quality of the product.

## C. SUMMARY

This chapter has discussed the results of two case studies conducted from the perspective of the intelligence analyst. The addition of text analytics tools such as Anseri could increase the quality and timeliness of intelligence provided to the consumer. Anseri is able to analyze and provide a cursory understanding of the content within a corpus

within seconds. With additional research a deeper understanding of the material can be gained in a fraction of time it would take without Anseri.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. CONCLUSION

The previous chapters discussed the nature of the intelligence process, current text analytics methods, and potential benefits of the Anseri software that can quickly process text. Effective intelligence support to operations is heavily reliant on timeliness and accuracy. The collection and processing and exploitation stages of the intelligence cycle are time-consuming and automation tools that can decrease the amount of time spent focused on these stages is highly beneficial to the intelligence analyst.

The recent technology push to digitally store media has made access to large amounts of documents easy and has raised the expectation of the amount of research one can conduct on many topics. Software tools, such as Anseri, that use text analytics can conduct topic analysis and reduce the amount of reading that is currently being done by intelligence analysts. Text analytics tools are able to summarize the topic themes within a corpus of over 3,500 pages of text in seconds and direct the analyst to the documents that need further investigation. In the current changing threat environment the United States military leadership needs fast, reliable intelligence to make sound decisions. Text analytics software would be a useful addition to the intelligence analyst's toolset.

## A. RECOMMENDATIONS FOR FUTURE WORK

This thesis does not fully explore the functionality of Anseri as an intelligence tool. There are other aspects of intelligence analysis and production that were purposely neglected to showcase the benefits of Anseri in one specific aspect of the intelligence cycle. Further study should be conducted in the following areas to better understand the capabilities of Anseri as an intelligence tool: corpus diversity, visualization, and user interface.

The text in the IS corpus has been previously identified to be overwhelmingly celebratory. Whiteside (2014) read each of the documents for his dissertation and made an analytical determination as to the nature of each document. Having a high percentage of documents with the same identification label limits the ability of the analyst to verify the performance of the topic analysis. The ability to distinguish between a diverse set of

semantics is not fully tested when analyzing a corpus with a monotonous sentiment. A more diverse corpus or multiple corpora could be used to evaluate the indexing abilities of Anseri.

Presentation is an important aspect of a good intelligence product. The decision maker needs a clearly communicated product to assist in the decision making process. Topic summarization of a corpus is a small portion of the analysis required to produce a comprehensive intelligence product; but, in some cases, it might be helpful to display the information gained from topic analysis for the decision maker. The output files from Anseri are not easily converted into a presentable format and it would be nice to be able to show the results of Anseri to show decision makers for explanation purposes.

A graphical user interface (GUI) would improve the analyst interaction with Anseri. The current interaction with Anseri is through a command line interface. While analysts do not necessarily have a baseline knowledge of command line interfaces, most have some experience and can more easily be trained to use a GUI. A GUI that allows the user to seamlessly compare the topic analysis produced by Anseri and the PIRs provided by the commander reduces the amount of training required to effectively run a text analysis software package.

Finally, exploring the indexing capabilities of text analytics could change the way that the military stores their own information. Intelligence collection and reporting utilizes a uniform formatting that could be read and classified by a machine. Autonomous classification of documents based on their thematic nature offers a unique method for improving search performance by accessing documents based on the thematic nature of document rather than the appearance of keywords in the title and abstract.

# LIST OF REFERENCES


Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Calafiore, G., & El Ghaoui, L. (2014). *Optimization Models*. Cambridge, UK: Cambridge University Press.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391.

El Ghaoui, L., Pham, V., Li, G. C., Duong, V. A., Srivastava, A., & Bhaduri, K. (2013). Understanding large text corpora via sparse machine learning. *Statistical Analysis and Data Mining, 6*(3), 221–242.

Godbehere, A. B. (2015). Fast and effective approximations for summarization and categorization of very large text corpora. Doctoral dissertation, University of California, Berkeley, CA.

Godbehere, A. B. SumUp LLC (2016). *Technical report: Anseri arguments*. Berkeley, CA: SumUp LLC.

Hofmann, T. (1999, July). Probabilistic latent semantic analysis. *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 289–296). Morgan Kaufmann Publishers Inc.

Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). *Taming text: how to find, organize, and manipulate it*. Shelter Island, NY: Manning.

Introducing JSON. (n.d.). Retrieved May 6, 2016, from http://www.json.org/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: Springer.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2013). *Handbook of latent semantic analysis*. Psychology Press, Hove, United Kingdom.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34(1),* 1–47.

SumUp LLC (2016). *SumUp corporate fact sheet.* Berkeley, CA.

U.S. Department of the Army (1994). *Collection management and synchronization planning, Field Manual 34–2*. Washington, DC: Department of the Army, March 8, 1994.

U.S. Joint Chiefs of Staff. (2013). *Joint intelligence*. Joint Publication 2–0. Washington, DC: U.S. Joint Chiefs of Staff, October 22, 2013.

Vinodhini, G., & Chandrasekaran, R. M. (2012). *Sentiment analysis and opinion mining: a survey*. *International Journal*, *2*(6).

What is text analytics? (n.d.). Retrieved January 7, 2016, from http://www.clarabridge.com/text-analytics/

Whiteside, C. A. (2014). The smiling, scented men: the political worldview of the Islamic State of Iraq, 2003–2013. Doctoral dissertation, Washington State University, Pullman, WA.

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California