CYBERSPACE AND REAL-WORLD BEHAVIORAL RELATIONSHIPS: TOWARDS THE APPLICATION OF INTERNET SEARCH QUERIES TO IDENTIFY INDIVIDUALS AT-RISK FOR SUICIDE

THESIS

Casey C. Miller, Captain, USAF

AFIT/GCE/ENG/12-08

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT/GCE/ENG/12-08

CYBERSPACE AND REAL-WORLD BEHAVIORAL RELATIONSHIPS: TOWARDS THE APPLICATION OF INTERNET SEARCH QUERIES TO IDENTIFY INDIVIDUALS AT-RISK FOR SUICIDE

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Engineering

Casey C. Miller, B.S. Computer Engineering

Captain, USAF

June 2012

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT/GCE/ENG/12-08

CYBERSPACE AND REAL-WORLD BEHAVIORAL RELATIONSHIPS: TOWARDS THE APPLICATION OF INTERNET SEARCH QUERIES TO IDENTIFY INDIVIDUALS AT-RISK FOR SUICIDE

Casey C. Miller, B.S. Computer Engineering

Captain, USAF

Approved:

_//Signed_____	___30 May 2012__
Maj Jonathan W. Butts, PhD (Chairman)	Date
_//Signed_____	___30 May 2012__
Dr. Robert F. Mills (Member)	Date
_//Signed_____	___30 May 2012__
Juan Lopez Jr. (Member)	Date

Abstract

The Internet has become an integral and pervasive aspect of society. Not surprisingly, the growth of ecommerce has led to focused research on identifying relationships between user behavior in cyberspace and the real world – retailers are tracking items customers are viewing and purchasing in order to recommend additional products and to better direct advertising. As the relationship between online search patterns and real-world behavior becomes more understood, the practice is likely to expand to other applications. Indeed, Google Flu Trends has implemented an algorithm that accurately charts the relationship between the number of people searching for flu-related topics on the Internet, and the number of people who actually have flu symptoms in that region. Because the results are real-time, studies show Google Flu Trends estimates are typically two weeks ahead of the Center for Disease Control.

The Air Force has devoted considerable resources to suicide awareness and prevention. Despite these efforts, suicide rates have remained largely unaffected. The Air Force Suicide Prevention Program assists family, friends, and co-workers of airmen in recognizing and discussing behavioral changes with at-risk individuals. Based on other successes in correlating behaviors in cyberspace and the real world, is it possible to leverage online activities to help identify individuals that exhibit suicidal or depression-related symptoms?

This research explores the notion of using Internet search queries to classify individuals with common search patterns. Text mining was performed on user search

histories for a one-month period from nine Air Force installations. The search histories were clustered based on search term probabilities, providing the ability to identify relationships between individuals searching for common terms. Analysis was then performed to identify relationships between individuals searching for key terms associated with suicide, anxiety, and post-traumatic stress disorder. Findings based on the calculated χ^2-test statistic demonstrate a strong correlation between the individuals who searched for the key terms. The results demonstrate the utility of clustering individuals who exhibit similar search patterns and provide the foundation for future efforts to bridge the gap between cyberspace and real-world situational awareness for identifying at-risk individuals.

Acknowledgments

This thesis was only possible through the support of my family, the guidance of Major Jon Butts and Dr. Rob Mills, the expertise of Jimmy Okolica and Dr. Deanne Otto, and the counsel and tenacity of Juan Lopez. I am indebted to you all. Thank you.

Casey C. Miller

Table of Contents

Page

Abstract iv
Acknowledgments vi
Table of Contents vii
List of Figures x
List of Tables xi
I. Introduction 1
1.1 Background 1
1.2 Motivation 2
1.3 Problem Statement 3
1.4 Approach 4
1.5 Contributions 4
1.6 Assumptions and Limitations 5
1.7 Thesis Organization 6
II. Literature Review 7
2.1 Overview 7
2.2 Behavioral Modeling 8
2.2.1 Theory of Planned Behavior (TPB) 10
2.3 Web Usage Mining 12
2.4 Text Mining and Cluster Analysis 13
2.4.1 Text Mining Examples 16
2.4.2 Latent Semantic Analysis (LSA) 17
2.4.3 Probabilistic Latent Semantic Analysis (PLSA) 18
2.4.4 Latent Dirichlet Allocation (LDA) 18
2.4.5 Self-Organizing Maps (SOMs) 21
2.4.6 Market Basket Analysis (MBA) 22
2.5 Summary 27
III. Data Attributes and Pre-Processing 28
3.1 Problem Definition 28
3.2 Data Attributes 29

3.3 Data Acquisition 29
3.4 Sampling Strategy 30
3.4.1 Sampling Frame 30
3.4.2 Sampling Selection 30
3.5 Raw Data File Format 31
3.6 Data Pre-Processing/Cleaning 32
3.6.1 Phase 1 32
3.6.2 Phase 2 33
3.6.3 Phase 3 35
3.7 Base Analysis 37
3.8 LDASOM 38
3.8.1 Topic Analysis 39
3.8.2 Cluster Analysis 40
3.9 Summary 44
VI. Determining Relationships 45
4.1 Air Force Suicide Prevention 45
4.2 Cyber Indicators 47
4.3 Disorder Determination 47
4.3.1 Anxiety Disorder 48
4.3.2 Post-Traumatic Stress Disorder (PTSD) 49
4.3.3 Suicide 50
4.4 Disorder-Related Searches 53
4.5 Contingency Tables 54
4.6 Analysis and Evaluation Technique 57
4.7 Combining Clusters 58
4.8 Cramer's V 67
4.9 Patterns in Clusters 70
4.10 Evaluation 72
V. Conclusions and Future Work 73
5.1 Conclusions 73
5.2 Future Work 73
5.3 Relevance of Work 74

Appendix A. BlueCoat Proxy Categories 76
Appendix B. Perl Scripts 77
B.1 Pre-processing - Original-to-SearchLog Filter 77
B.2 Pre-processing - Combine daily SearchLogs into monthly SearchLog 79
B.3 Pre-processing – Unique IP Counter 81
B.4 Pre-processing - IP Filter 83
B.5 Disorder-Related Search Histories 86
B.6 Cluster Statistics 89
Appendix C. LDASOM Topics 90
Works Cited 96

List of Figures

Page

Figure 1. Google Flu Trends estimate vs. CDC Data 9

Figure 2. Theory of Planned Behavior..... 11

Figure 3. Graphical model representation of LDA 20

Figure 4. Proxy log data is uploaded to the I-NOSC for centralized storage 29

Figure 5. Data flow from AFB logs to individual search histories by IP Address 31

Figure 6. Phase 1 process to create Search Logs 33

Figure 7. Phase 2 filter, which writes the actual searches from the daily Search Logs to a single file representing the searches from a given base for the entire month .. 34

Figure 8. Phase 3 filter, creating a search history for each IP address 35

Figure 9. Process to filter and clean the actual search queries from the URL..... 36

Figure 10. Example Search History..... 36

Figure 11. Top-level flowchart for LDASOM..... 39

Figure 12. SOM generated by LDASOM for +24,000 valid search histories 41

Figure 13. Determining search histories which contain disorder-related searches..... 54

Figure 14. Process to find disorder-related search histories within the SOM clusters 56

Figure 15. Clusters and their top three topic probabilities..... 61

Figure 16. SOM with lines showing where clusters were combined..... 63

Figure 17. SOM with lines showing updated cluster combinations 64

Figure 18. Air Force suicide totals since 2003 69

Figure 19. The six clusters containing the most disorder-related searches..... 71

Figure 20. The 11 clusters containing the most disorder-related searches 72

List of Tables

Table Page

1. Data attributes for each of the nine AFBs 38

2. Cluster analysis 43

3. Dictionary of words associated with anxiety disorders 49

4. Dictionary of words associated with PTSD 50

5. Dictionary of words associated with suicide 52

6. Observed Frequency with 3 disorders and 17 clusters 57

7. Observed frequency with clusters 11, 22, and 24 removed 58

8. Expected frequency with clusters 11, 22, and 24 removed 59

9. Updated observed frequency table 63

10. Updated expected frequency table 64

11. Updated observed frequency table 65

12. Updated expected frequency table 65

13. χ^2-test statistic 66

14. χ^2-test table with alpha of .05 67

15. Cramer's V characterizations 68

16. Clusters from high to low based on number of disorder-related searches 70

AN EXPLORATORY STUDY OF CYBER-BASED SEARCH QUERY ATTRIBUTES TO IDENTIFY BEHAVIORAL INTENT

I. Introduction

1.1 Background

The frequency at which society looks to the Internet for answers inspired John Battelle, the founder of Federated Media Publishing and *Wired Magazine*, to develop and define the Database of Intentions:

> "The aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result... This information represents, in aggregate form, a place holder for the intentions of humankind - a massive database of desires, needs, wants, and likes that can be discovered, subpoenaed, archived, tracked, and exploited to all sorts of ends. Such a beast has never before existed in the history of culture, but is almost guaranteed to grow exponentially from this day forward. This artifact can tell us extraordinary things about who we are and what we want as a culture. And it has the potential to be abused in equally extraordinary fashion [1]."

The underlying question with the Database of Intentions is "What, in the end, might search tell us about ourselves and the global culture we are creating online?" Online search is a means to an end – an attempt to achieve an underlying goal of discovering what a user believes exists on the Internet.

The Internet has become the first place many people look for real-world answers. Ginsberg *et al.* [2] described how monitoring health-seeking behavior in the form of online search queries can improve the early detection of both seasonal and pandemic

influenza. Researchers from West Point and Princeton University determined that drastic changes in online search queries in Egypt at the end of 2010 into 2011 provided strong indicators to the upcoming uprising in Egypt [3]. These two cases provide an initial framework to develop and implement an automated information system capable of quantifying relationships between cyber and real-world behavior.

1.2 Motivation

According to the August 2010 final report of the Department of Defense Task Force on Prevention of Suicide by Members of the Armed Forces, from 2005 to 2010, service members committed suicide at a rate of approximately one every 36 hours [4]. Last September during a presentation to the Subcommittee on Military Personnel Committee on Armed Services, United States House of Representatives, Lieutenant General Darrell Jones stated that "Despite our prevention efforts, suicide rates remain a concern… So far this year, 56 Total Force Airmen and Civilians have taken their own lives, which equates to a suicide rate of 14 suicides per 100,000 Airmen." [5]

The Department of Defense has devoted significant resources to suicide awareness and prevention, including over $67 million in research [4]. These efforts are primarily focused on providing family, friends, and co-workers with information on identifying at-risk personnel. According to the United States Air Force Suicide Prevention Program (AFSPP), these three groups are in the best position to recognize behavioral changes, discuss these changes with the at-risk individual, and provide care and support [6]. Although behavioral change in itself does not imply someone will become suicidal, the indicators may identify individuals that warrant close monitoring.

Despite Air Force efforts, the suicide rates have remained largely unaffected. The 7 May edition of the Air Force times pointed out "more airmen killed themselves in the first 3 months of this year than in any other first quarter in the past decade." [7]

This research explores the possibility of leveraging a different indicator of behavioral change – online search queries. As society becomes more dependent on the Internet for answers to life's questions [8], it seems logical that individuals considering hurting themselves would search for answers as well. If so, is it possible these individuals may exhibit behaviors within cyberspace that indicate they are at-risk?

1.3 Problem Statement

Traditional suicide research is limited by several factors [9]. First, the variables selected for analysis are decided by the researchers, making them inherently subjective. Furthermore, the studies are retrospective since the suicide victims have already passed. For example, the analysis of suicide notes after an incident is often the primary means for insight into the individual's motivations [9]. Finally, the data sets are unrepresentative convenience samples since researchers are typically limited by their sources. Search query analysis has the potential to minimize these limitations and provide early indicators of someone considering suicide or suffering from a depressive state.

At a macro-level, this research effort explores the possibility of a relationship between cyberspace and real-world behavior. This is accomplished by identifying, classifying, and analyzing the online search query histories of individuals in order to cluster users who exhibit similar search patterns. Once these search histories are classified and clustered, topics associated with suicide, anxiety and post-traumatic stress

disorder (PTSD) are targeted to determine if relational attributes exist that provide insight into real-world characteristics.

1.4 Approach

In order to examine the relationship between cyber and real-world behavior, search query data is parsed from the proxy logs of nine Air Mobility Command Air Force bases. Search queries are attributed to the originating internet protocol (IP) address, and a search history is saved for each IP. Text mining is then performed on these search histories to determine the most probabilistic topics based on the search query terms. The search histories are then clustered based on these probabilities, making it possible to mathematically determine the relationship between different search histories.

Once search histories are classified and clustered, those which contained search queries with terms associated with suicide, anxiety and PTSD are evaluated to determine if relationships exist with the clustered data. Indeed, a correlation demonstrates the ability to group search histories based on search queries – a first step in connecting real-world and cyber-indicators for identifying at-risk personnel.

1.5 Contributions

Previous research provides an initial framework for the development and implementation of an automated information system capable of quantifying relationships between cyber and real-world behavior [2] [3]. A system focused on the application of Internet search queries to identify individuals at-risk for suicide could have significant impact across the Department of Defense (DOD).

Despite the one-dimensionality, search query analysis could have a significant impact on the most significant limitation in traditional suicide research – a dependence on previously collected data [9]. However, an even greater impact may be the addition of a second dimension – clustering the online search histories of users and identifying relationships with at-risk individuals who have searched for topics relating to suicide, anxiety and PTSD. By determining the heuristics and taxonomy required to model online search patterns, this research lays the foundation for future efforts to bridge the gap between cyber and real-world situational awareness.

1.6 Assumptions and Limitations

Internet search logs are obtained from Air Force bases and, as such, it is assumed the users are either in the military, employed by the government, or accustomed to military culture. Therefore, the results of this research may not apply to the general population. Although future research should include a larger data-set, both spatially and temporally, this would guarantee an improvement in results.

Additionally, the Air Force associates IP address to end-user computer systems via associated dynamic leases. Although the possibility exists that multiple users are associated with the same IP address, this scenario is not consistent with Air Force implementation processes. Indeed, the Integrated Network Operations Security Center, responsible for network management, has deemed that it is sufficient to assume that one IP address is associated to one specific user according to their operating practices. In keeping with maintaining user anonymity, there is no personally verifiable information

within the logs – IP address rather than user names are used to delineate unique search queries.

Regarding the disorder dictionaries, although the terms included in each are mutually exclusive, the symptoms are not – as the disorders studied share many. Furthermore, if other terms are used in the dictionaries, the output will likely be very different.

1.7 Thesis Organization

Chapter 2 describes previous and related work in the fields of behavioral modeling, data mining, and pattern recognition. Chapter 3 details the methodology used to pre-process the proxy log data and discusses associated data attributes. Chapter 4 examines relationships and observations by clustering individuals whose search histories were associated with suicide, anxiety and PTSD search queries. Finally, Chapter 5 concludes with a discussion on the implications of the research and ideas for future research.

II. Literature Review

This chapter focuses on the major influences in behavioral modeling and prediction. Although literature relating to this specific topic as it applies to cyberspace is extremely limited, a great deal of supportive literature does exist. This chapter examines pertinent research in these areas.

2.1 Overview

There is an innumerable amount of information available online. The challenges and benefits of modeling online search patterns have grown immensely over the last several years [10]. The process of extracting these patterns is a relatively young, interdisciplinary field within computer science termed data mining [11]. This research focuses on web mining, a subcategory of data mining. Web mining seeks to discover patterns in web data and use the patterns to develop realizations about the people who produced them [12].

Before web mining can take place, a suitable data-set must be identified and collected. The suitability of a data-set depends on the goals of the research effort. Regardless, the initial step involves preprocessing the data to allow for the practical application of data mining techniques and statistical analysis [12]. These initial steps are critical, as they set the foundation for the research effort. If the collected data does not include the correct demographic and type of information, any successive research will be difficult or impossible. Therefore, a heavy emphasis is placed on the targeting, planning, and collection of web usage data.

Web usage data can be partitioned into four categories: Usage, Content, Structure, and User [13]. Usage data is typically used for web mining and pattern extraction, since it includes search queries and represents the navigational behavior of the user [11]. Every HTTP request will typically generate an entry in a server log, which would include information such as the time and day of the request, client IP address, resources requested, status of the request, and if the client is a returning visitor - most likely with a client-side cookie.

After collection, the data often requires a substantial amount of data preparation. Pre-processing the original data, integrating data from multiple sources, and transforming the integrated data into a form suitable for input into specific data mining operations must be accomplished before any analysis can be performed [11].

2.2 Behavioral Modeling

For the purpose of this research, cyber behavior is defined as the set of observable online activities (e.g., search queries) and the statistical characterization that accompanies it. Internet search queries have the potential to provide insight into human behavior. For example, a user browsing web sites for cars, financing, and dealerships could be characterized at the most basic level as simply "web browsing." Note that this characterization alone does not specify the underlying cyber mediated behavior associated with buying a car.

Several research efforts have demonstrated valid relationships between user behavior in cyberspace and the real world [2] [3] [14]. Retailers are tracking the items their customers are viewing and buying, and using this information to determine other

items to recommend and advertise to them [10]. By doing this, retailers are using trends in cyber behavior to make real-world dollars. Indeed, these relationships are already being exploited. Google Flu Trends accurately charts the relationship between the number of people searching for flu-related topics on the Internet, and the number of people who actually have flu symptoms in that region [2]. Figure 1 shows a comparison of the Google Flu Trend estimate and Center for Disease Control (CDC) truth data from 2004 through the present [2]. Because Google Flu tracks search queries in real-time, the estimates are typically two weeks ahead of CDC, which relies on post-processed data.

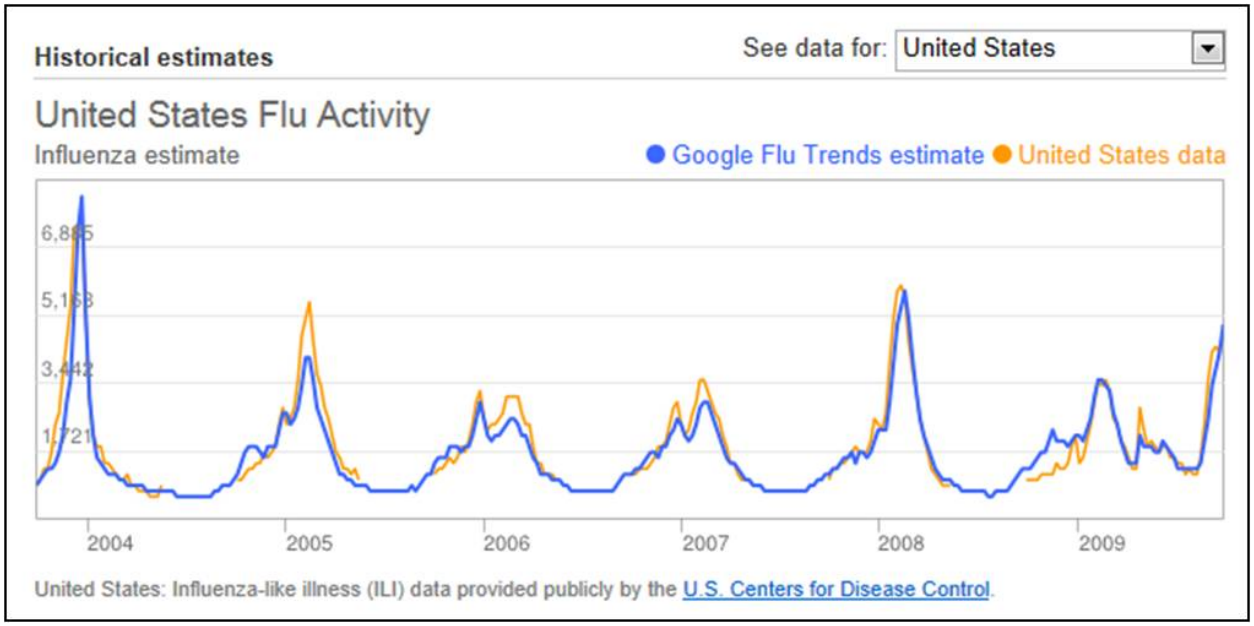

Figure 1. Google Flu Trends estimate vs. CDC Data

Researchers from West Point and Princeton University found drastic changes in online search queries in Egypt at the end of 2010 into 2011. They determined these changes were significant enough to have hinted to the upcoming uprising in Egypt, if the changes had been monitored [3]. For decades, researchers in psychology have studied

behavior and the process people go through before a behavior is established [15]. There has been minimal formal research at this juncture to extend these theories to cyberspace.

2.2.1 Theory of Planned Behavior (TPB)

Extensive research has been conducted in psychology regarding Icek Ajzen's Theory of Planned Behavior (TPB) which explores the link between attitudes and behavior [16]. Outside of psychology, TPB has been used to explain and exploit behavior in many fields including advertising, public relations, and healthcare [16] [17] [18]. The theory states that an individual's behavioral intentions and behaviors themselves are directly related to his/her attitude toward the behavior, subjective norms, and perceived behavioral control [15].

Attitude is defined as an individual's positive or negative feelings about performing the behavior in question. It is calculated by assessing two things: the individual's beliefs regarding the consequences of the behavior and the desirability of these consequences. *Subjective norm* is an individual's perception of whether other people (specifically people close to the individual) think the behavior should be performed. Finally, *perceived behavioral control* is defined as an individual's perception of how difficult it would be to perform the behavior.

These three factors feed into each other, with the relationships depicted by the solid lines in Figure 2. The arrows describe the direction of the links. According to Ajzen, the combination of *attitude, subjective norm,* and *perceived behavioral control* determine *intent* [15]. *Intent* is an indication of an individual's readiness to perform the behavior. The dotted line signifies that *perceived behavioral control* serves as a proxy

for predicting behavior, since performance of a behavior depends on both favorable intent and a sufficient level of behavioral control [15].

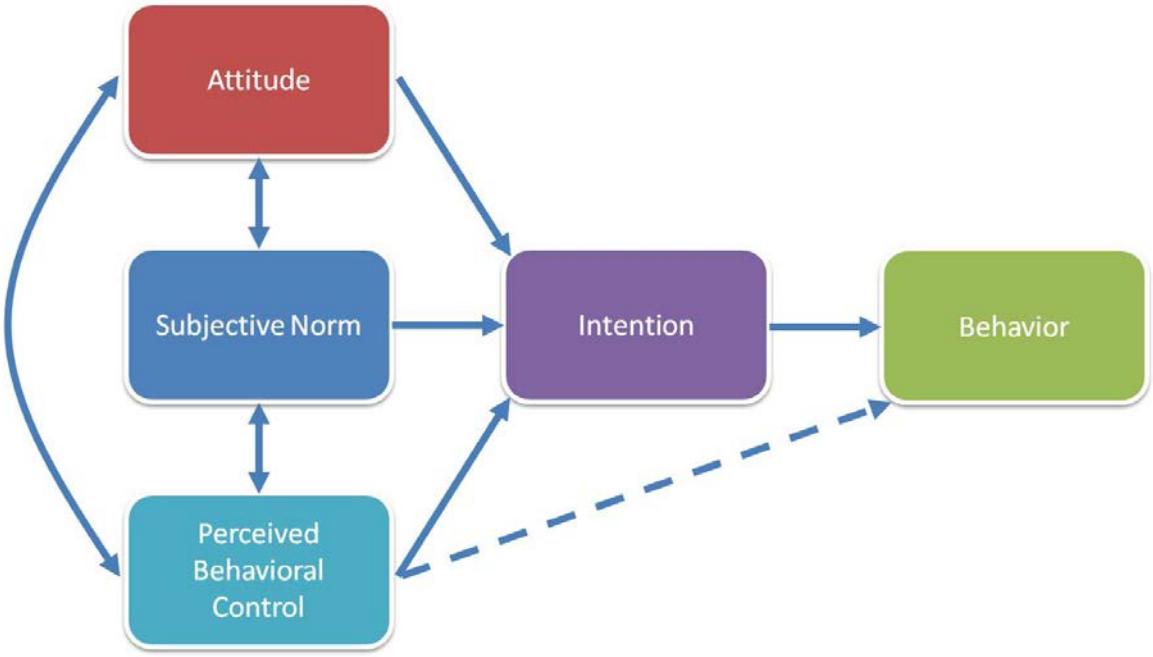

Figure 2. Theory of Planned Behavior

Previously, the only way to measure these factors was actively; by talking with the individual, completing surveys, etc. This research explores the possibility of passively gathering information regarding these factors through search queries. For example, if an individual was unsure of public perception regarding smoking, it would not be unusual for him/her to submit a search query about it. Furthermore, if the individual was attempting to quit smoking, he/she may look to the Internet to provide answers on the best way to accomplish it (perceived behavioral control).

2.3 Web Usage Mining

Proxy logs are data-rich, but information-poor. Web mining seeks to discover patterns in web data and use these patterns to develop knowledge about the individuals or groups who produced them [11]. In general, Web Mining is broken out into three core areas: web structure mining, web content mining, and web usage mining [11]. Web usage mining, the focus of this research, is by far the biggest growth area [13].

While web structure and content mining are primarily concerned with web sites and web data, web usage mining is concerned with the end user. The focus of web usage mining is to extract user access patterns by analyzing and interpreting information from users' Internet browsing patterns. One of the fundamental analysis techniques used in web usage mining involves tracking clickstreams - historical descriptions of what a user did and where a user went while browsing a particular web site [19].

Clickstreams contain many sources of data including the timestamp and source of the request, the destination host, an assortment of information about the browser, and the Uniform Resource Identifier (URI). The URI is a critical piece of information in a clickstream. It represents the global address of documents and resources present on the World Wide Web, with web page addresses being the most common [19].

Although research in web usage mining has been conducted in a number of fields including web page pre-fetch/cache, web site optimization, and recommender systems – e-commerce is generating the most interest and revenue [13]. Continued increases in online shopping have spurred a growing interest in profiling and analyzing online shoppers to better target sales [20]. Retail and marketing firms are taking advantage of

user profiling by aggregating data on the purchase history of individuals (on and offline), finance records, magazine subscriptions, supermarket savings cards, surveys, sweepstakes entries, and many other sources [19]. In order to make this information meaningful, it is pre-processed, organized, and analyzed using a number of statistical and data mining techniques. The final product is a basic shopping profile of an individual. When aggregated, these profiles are used for targeted ad campaigns, personalize shopping experiences, and making recommendations for additional product purchases – all based on a user's purchase history.

2.4 Text Mining and Cluster Analysis

Another subset of data mining is text mining – where the words within a document become the target set for identifying and extracting patterns. The theory behind text mining has been defined and explored since the inception of data mining. At the most basic level, text mining analyzes a set of documents for meaningful information and patterns within the text. The size of the sample set can range from as small as hundreds, to as large as millions, and the differing structures of the documents dictates the amount of standardization required to improve the results [21]. The complexity of mining words instead of numbers means more computational processing power is needed for analysis. Compared to data mining, text mining requires more extensive cleaning and standardizing before the data can be analyzed [21]. The immense growth of technology over the last several decades has made research in text mining a reality.

Most automated topic modeling/clustering techniques extract and generate labels from data sets based on the keywords and phrases within the text, eliminating the rigidity

associated with a pre-determined set of labels [22] [23]. Although this approach is ideal from a human perspective, it severely complicates the normalization of the data and has the potential to produce a distinct label for each distinct object. Therefore, this research effort opts to define a static number of topics to map the input data. The challenge with this approach is finding a representative set of category labels with the necessary depth and flexibility. This process is described in Chapter 3.

One significant disadvantage with machine categorization is its inability to interpret polysemy and homonyms – terms having multiple meanings depending on the context [19]. For example, the word "bow" can reference a kind of tied ribbon or a weapon that shoots arrows. Therefore, machine generated categorization should involve some manual, human verification. When this is not possible or must be avoided, machine-made datasets are the only real alternative, and can either be supervised or unsupervised.

The objective of supervised text categorization is to learn classifiers from examples or training sets. The three most widely studied and effective algorithms are k nearest neighbor (k-NN), Naive Bayes, and support vector machines (SVM); all three rely on pre-categorized training data. Although some research has been performed on unlabeled training data, results are not comparable [24] [25].

In unsupervised learning, the machine receives inputs only without supervised target outputs. A number of well studied algorithms exist in the realm of text-based categorization, but most pertinent to this research is hierarchical text clustering [26]. Clustering, the process of finding natural groups in unlabeled data, is a well-documented

form of unsupervised learning relative to text categorization. The goal behind clustering is to characterize groups of individuals and maximize intra-cluster similarity while minimizing inter-cluster similarity [19]. This type of analysis helps determine the most prevalent characteristics within a group, which can then be used to customize individual profiles.

Within clustering is another unsupervised learning technique called Topic Models. Given a topic is a probability distribution over words, Topic Models provide a simple mechanism to analyze and label copious volumes of text by treating documents as a mixture of topics. There are four different pieces within a document that are accounted for: characters, words, terms, and concepts [21]. Characters describe letters, numbers, and symbols within the text. Words describe a combination of characters with a space or punctuation on either side. It is computationally less expensive to mine characters and words than terms and concepts, but the information and predictive ability are severely limited. Terms are a combination of two or more words based on occurrences within the text. Terms offer more information to the analyst, but if the amount of data is limited, there may be too few occurrences. The last piece is concepts, which are derived from hybrid categorization methodologies and by cross-referencing phrases and words to determine what the text is truly describing, even if specific words or phrases are not included [21]. Topic Mining uses contextual clues to connect words with similar meanings and distinguish words with multiple meanings, and have been implemented successfully in multiple research efforts as a means to effectively and efficiently extract key concepts from text [27] [28] [29].

Changes in the frequency of key words or phrases within a document or over a period of time can provide valuable information to analysts. Unfortunately, prepositions and other *stop-words* such as *the*, *and*, and *is* can make it difficult for the algorithm to focus on words that have meaning. Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) each have a different process to extract predictive and useful information from text, and will be described in detail in this section.

2.4.1 Text Mining Examples

The medical field was one of the first to seriously utilize text mining to aid research [16]. The National Centre for Text Mining published an article describing attempts to improve the organization of bacteria classes, since the names and descriptions of many bacteria vary with location [30]. Another research effort applied text mining to descriptions of biological activity and the target of the biological activity (i.e., gene, protein, cell, or microorganism) to predict and understand the effects of natural substances [31]. A similar effort sought to automatically extract the microorganisms and habitats [32].

Marketing and knowledge management fields have also used text mining to improve their practices. Researchers in marketing have mined large amounts of data in an attempt to more effectively reach customers and encourage them to purchase particular products [33]. The convenience store 7-Eleven used text mining to determine where and how to implement an iced coffee product by text mining social media sites to gain insight into peoples' thoughts about their products and flavors [34]. Going a step further, text

mining was conducted on social media sites to determine customer sentiment toward companies, and then applied to trends in the stock market - suggesting that text mining may allow researchers to detect economic trends more quickly [35].

These medicine and marketing examples are just scratching the surface of what research in text mining can bring in the future. For years, the Department of Defense (DOD) and other government agencies have utilized text mining to track possible threats to the nation's defense, predict/detect terrorist activities, and find/trace viruses [36] [37] [38]. This research effort describes a different application of text mining for the DOD – analyzing online search histories of users in order to classify the heuristics for online search patterns.

2.4.2 Latent Semantic Analysis (LSA)

LSA multiplies a series of three matrices (i.e., document eigenvector, eigenvalue, and term eigenvector) to approximate an original matrix to describe the document. The size of the document eigenvector is determined by the number of documents (n) multiplied by the unique dimensions of the sample set (r). The eigenvalue matrix defines the unique dimensions (n x r) and the term eigenvector matrix is the number of unique terms (m) multiplied by the number of unique dimensions (m x r).

LSA is limited in that the words in one topic have little relation to other topics, and the words in one topic cannot occur in other topics. This means words with multiple meanings cannot be classified under two different topics. LSA works best on documents with similar writing styles, but its functionality is limited by the fact that the reduction of

the document matrix does not use robust probability theory and an adequate number of topics cannot be determined statistically [39].

2.4.3 Probabilistic Latent Semantic Analysis (PLSA)

PLSA expands on the LSA methodology and improves results by calculating several probabilities: the document within the sample set to be $P(d)$, a topic to be $P(z|d)$, and a word $P(w|z)$; where d, z, and w stand for document, topic, and word respectively. $P(z|d)$ describes the probability of a topic given a document. $P(w|z)$ describes the probability of a word given a topic. PLSA finds general themes or trends in documents by expressing the results in terms of probabilities of these three occurrences [39]. PLSA does allow words to occur in different topics but does not fully reflect the generative process at a document level [39].

2.4.4 Latent Dirichlet Allocation (LDA)

LDA was developed to extend the applications of LSA and PLSA. It determines the number of words in a document by sampling with a Poisson distribution, creates a distribution for the topics using the Dirichlet distribution, and generates topics and words for the topics based upon a document's distribution [39]. LDA performs well on lengthy documents that have multiple topics and includes spatial statistics to provide a relational factor to words that occur near each other often.

By adjusting the analysis, LDA ensures the number of occurrences of one word or phrase does not overshadow the power or significance of other words [40]. A hierarchical Bayesian model enables LDA to determine topics, filters out insignificant stop-words, and categorize the words from the texts into the topic(s) [41]. LDA was the

best fit for this research because of its superior ability over LSA and other text mining methods to separate words and create topics that have a probability distribution of how often words occur within the documents. After each iteration through the sample set of documents, LDA sorts each word based upon the probability distributions of the words given the topics, then re-calculates the distributions for the topics. The total number of iterations depends on the application, but eventually the words given the topics gain a specific distribution which describes the topic. At this point, the topics become well defined and distinct. Each iteration calculates an updated conditional probability distribution of words given a topic, and each word is then allocated into topics based on this distribution.

Figure 3 provides a graphical representation of the LDA model. The shaded node represents observed variables while all others are latent variables. Arrows represent dependencies and boxes represent repeated sampling operations. α and β are the dirichlet priors used to parameterize these distributions. The boxes are "plates" representing replicas. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document [37]. Symbols represented in Figure 3 are as follows:

M: number of documents in the data set

N: number of words per document

T: number of topics

z: topic from which a particular word *w* is drawn

θ: per-document multinomial *topic* distributions

ϕ: per-topic multinomial *word* distributions

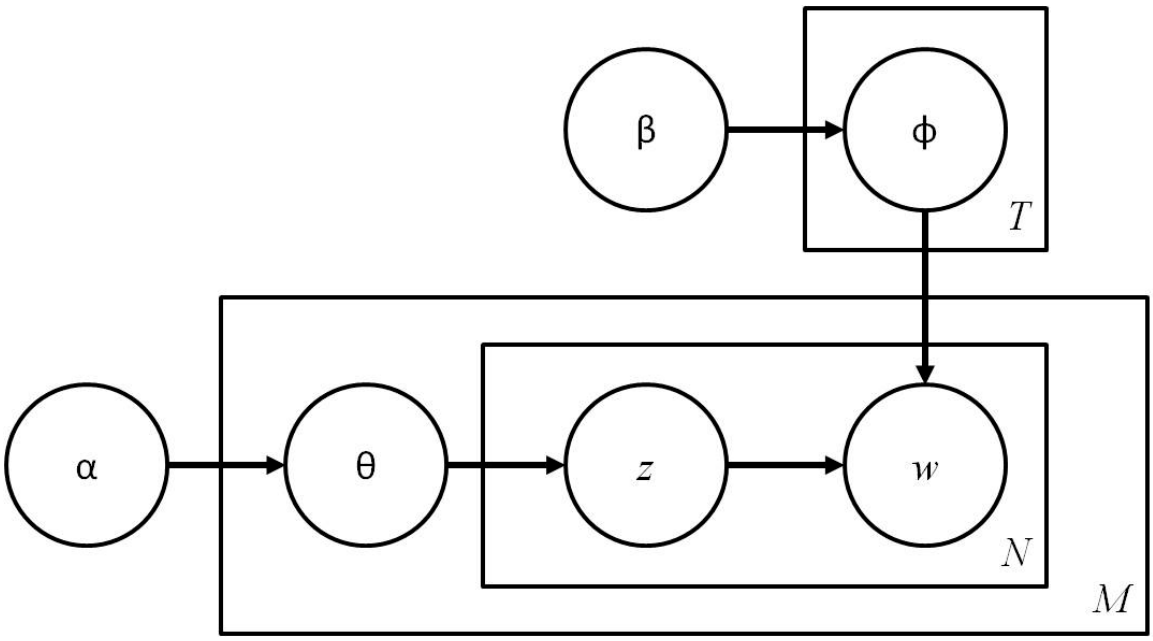

Figure 3. Graphical model representation of LDA

LDA is a straightforward process:

1. Choose values for the hyper-parameters α and β and the number of topics *T*. The values of α and β depend on *T* and the vocabulary size. Recommended choices are $\alpha = 50/T$ and $\beta = 0.01$ [29].
2. For each document:
    a. Choose the number of words *N*.
    b. For each word:
        i. Sample *z* from $\theta^{(j)}$, where *j* is the current document index.
        ii. Sample *w* from $\phi^{(z)}$.

In order to perform document clustering, LDA finds $P(z|w)$ for fixed α, β, and *T*. Once LDA calculates $P(z|w)$, the distributions ϕ and θ are estimated for each topic and document. The topic distributions (θ) form the basis of the clustering method.

Miller provides a simple example to describe the LDA process [42]. Given millions of pennies, quarters, dimes, nickels, half dollars, and dollar coins, LDA randomly separates them into topics. LDA then creates a distribution of each topic of

coins. While one topic may contain thirty pennies, five dimes, twelve quarters, two nickels, one half dollar, and no dollar coins, another topic may have a high density of nickels. After the next iteration through the sorting process, LDA uses the distribution of the coins within that topic to determine which topic best describes each coin. In this case, LDA will allocate more pennies to the first topic and more nickels to the second. After repeating this process a few hundred times, topics dominated by one coin or another will emerge and the topics will clearly represent a particular coin.

2.4.5 Self-Organizing Maps (SOMs)

SOMs are a form of unsupervised neural network consisting of a fixed lattice of processing elements, which is typically 2-dimensional [43]. Each processing element has an associated prototype vector, which initially is random. Learning takes place in a competitive fashion. For each input, the processing element with the shortest Euclidean distance is identified as the Best Matching Unit. The prototype vector for all other elements within a neighborhood is updated according to:

$$w_j(t+1) = w_j(t) + \alpha(t) h_{ji}(t)(x_m - w_j) \qquad \text{(Eq. 1)}$$

where w_j is the prototype vector associated with the jth processing element, $\alpha(t)$ is a monotonically decreasing learning rate, $h_{ji}(t)$ is a time-decreasing neighborhood function (typically Gaussian), and x_m is the input sample. Over time, the original input space converges to a low-dimensional representation. Self-organizing maps naturally cluster the input data so that inputs with similar features are mapped to the same or neighboring processing elements, while preserving the topology of the original high-dimensional input space on the lattice [44]. These properties, coupled with the relationship

preservation between samples in the high-dimensional input space on the low dimensional mapping, make SOMs an ideal tool for visualizing high-dimensional data in 2-dimensional space [43] [45].

Once the elements are mapped onto the SOM lattice, the lattice itself is clustered using *k*-means. In order to gain a simple and intuitive view of the document collection, the lattice is plotted along with cluster boundaries. Since SOMs preserve the topology, the user can visually identify related groups of documents. Neighboring clusters often have one or more topics in common, based on the mean topic distribution for each cluster [44].

2.4.6 Market Basket Analysis (MBA)

Market-Basket Analysis (MBA) is a modeling technique based upon the theory that if an individual buys a certain group of items, he/she is more/less likely to buy another group of items. Although this research project is not focused on shopping, there are some universal concepts used by MBA to determine trends that may also occur in online search patterns. Just as a user is more likely to buy a particular group of items based on the ones he is already planning to purchase, a user may be more likely to search for a particular term based on the ones he has already searched for.

Researchers have found that in retail, most purchases are made on impulse. MBA provides clues as to what a customer might have bought had they thought of it ahead of time. Therefore, as a first step it can be used to decide where goods should be located within a store. A set of items in a supermarket domain may be:

$$I = \{Milk, Napkins, Video\ Games, Candy\}$$

If it has been observed that people who purchase video games are more likely to buy candy, it would become a rule within this example:

$$\{Video\ Games\} => \{Candy\}$$

Therefore, candy should be placed near the video game display. This way customers shopping for video games, who would have bought candy had they thought of it, will now be tempted to do so.

A major difficulty in MBA is the large number of rule sets. Although the volume of data has been reduced, there is still the problem of asking the user to find a needle in a haystack. In order to avoid missing any exploitable results, MBA requires each rule to have a high minimum support level and high confidence level risks.

The next level of analysis, Differential MBA (DMBA), is a partial solution to this problem. DMBA deals with finding interesting results and eliminating problems with potentially high levels of trivial results [19]. It compares results between different stores, different demographic groups, different days of the week, different seasons, etc. If researchers observe that a rule holds in one store, but not in any other (or vice versa), then they know there is something interesting about that store. Investigating what makes this store unique, from the clientele to the organization of the store itself, is worth exploring in hopes of improving sales.

Most approaches to association discovery are based on the Apriori algorithm, which finds groups (i.e., itemsets) of items or pageviews occurring frequently together

across multiple transactions [46]. Once there exists a set of frequent itemsets, researchers can apply constraints on measures of significance and interest to generate interesting association rules to satisfy a minimum confidence threshold. An association rule is an expression of the form X→Y [sup, conf], where:

X and Y are itemsets.

sup is the support of the itemset X ∪Y - the probability that X and Y occur together in a transaction.

conf is the confidence of the rule, defined by sup(X ∪Y) / sup(X); the conditional probability that Y occurs in a transaction given that X has already occurred in that transaction.

Let X be an itemset and Y the multiset of all applicable transactions. The absolute support of the itemset X is the number of transactions in Y that contain X. The relative support of X is the percentage of the transactions in Y which contain X. The support of the rule X→Y is computed as follows:

$$support = \frac{(X \cup Y).count}{n} \quad \text{(Eq. 2)}$$

The confidence of a rule, X→Y, is the percentage of transactions P that contain X and also contain Y. It is computed as follows:

$$confidence = \frac{(X \cup Y).count}{X.count} \quad \text{(Eq. 3)}$$

Confidence determines the quality and predictability of the rule. If the confidence of the rule is low, one cannot reliably infer or predict Y from X, which drastically limits its use.

Lin *et al.* [47] proposed collaborative recommendation - a mining algorithm which finds an appropriate number of rules for each target user by automatically selecting the minimum support. It generates association rules among users as well as among items. In the case that a user minimum support is greater than a threshold, the system generates recommendations based on user associations. Otherwise, it uses item associations.

A problem with using a single minimum support threshold in association rule mining is that the discovered patterns will not include "rare" but important items which may not occur frequently in the transaction data. Thus, for more effective mining, it is important to capture patterns and generate recommendations that contain these items. Liu *et al.* [46] proposed a mining method based on multiple minimum supports which allow users to specify different support values for different items. In this method, the support of an itemset is defined as the minimum support of all items contained in the itemset. The specification multiple minimum support thus allows frequent itemsets to potentially contain rare items which are deemed important. Most online vendors now include recommendations for users based on items they view or purchase, and a substantial amount of research on association analysis is underway for these types of recommendation systems.

Forte *et al.* [48] extended MBA by conducting an experiment investigating whether it was possible to determine profiles of online shoppers based solely on outside observation by learning repeatable patterns of behavior through data mining techniques. Their result suggest that statistical models applied to the actions taken and time spent on each decision may enable the discernment of demographic information, as well the

inference of connections between shopper and recipients. These results bring optimism to the idea that shopper identity could eventually be determined by examining the action itself, as well as the target of the actions.

Forte and his colleagues believe the methods they employed would translate well to other online environments - determining identities in online fraud cases or other issues related to cyber security. They hypothesized that individuals as well as groups with similar demographic backgrounds will maintain consistent patterns. Their test data included type and sequences of actions, time between actions, and a confidence rating for each participant and profile.

The experiment used two metrics - general data trends and prediction model accuracy. The general data trends included gender, college degree, confidence and profile. Once demographic groups were distinguished, they were analyzed on time and actions in a search for statistical differences leading to trends. Significant differences were determined by two sample t-tests and a single factor analysis test at a 95% confidence level.

Their analysis involved significance testing, regression modeling, and weights of evidence (WOE). WOE uses a scoring system to test the ability of the model to accurately predict a demographic characteristic of a user. The scoring system is created by splitting characteristics such as time into separate ranges, finding the probability of a user's action falling into one of the ranges conditional on an observation, and calculating a logodds score. The WOE technique has been successful in credit scoring to predict behavioral patterns for lenders.

2.5 Summary

This chapter discussed some of the major influences in behavioral modeling and prediction. It also outlined techniques to determine applicable data sets, mine useful information from text, and determine relationships between documents. The remaining chapters describe the methodology required to identify, characterize, and cluster search histories, as well as the analysis of the results.

III. Data Attributes and Pre-Processing

3.1 Problem Definition

The average American spends nearly 3 hours a day on the Internet [8] - checking the weather, fantasy football stats, and countless other things. The Internet has become part of the daily routine, the communication and information medium of choice, because it puts answers at our fingertips to questions we have not yet asked.

What are people searching for on the Internet? What questions are they asking? Is it possible to categorize or classify searches to gain knowledge about behavioral intent? In other words, is it possible to determine real-world events based strictly on cyber behavior? This exploratory research was conducted over a 12 month period to lay the groundwork for a system capable of clustering the online search histories of users and identifying individuals at-risk for suicide.

This chapter discusses the approach taken to explore patterns in online search queries across nine Air Mobility Command (AMC) Air Force Bases in North America over the course of 32 days, from 7 November to 7 December 2011. At a macro-level, this research effort explores the relationship between cyber and real-world behavior by identifying, classifying, and analyzing the online search query histories of individual users in order to cluster users who exhibit similar search patterns. Once these search histories are classified and clustered, specific topics are targeted to determine if relationships exist and what knowledge can be gained about the real world from these cyber-indicators. It is important to note this research respects the privacy of the users by associating search histories to IP addresses instead of specific users.

3.2 Data Attributes

The AMC bases from which the proxy logs were acquired are responsible for the worldwide cargo and passenger delivery, air refueling, and aero-medical evacuation. Although the specific mission varies by installation, the AMC mission is to provide global air mobility – right effects, right place, right time. The users generating the search queries represent roughly a quarter of the more than 134,000 active-duty, Air National Guard, Air Force Reserve, and DOD civilians responsible for making AMC's rapid global mobility operations possible.

3.3 Data Acquisition

Air Force bases are responsible for logging the traffic on their servers [49]. This data is stored in one of two centralized locations called Integrated Network Operations and Security Centers (I-NOSCs), as depicted in Figure 4. The I-NOSCs store several different logs, including: Aruba, Cisco, firewall, ironmail, proxy, and proxy system logs. This research focuses on the proxy logs, as they contain the outgoing HTTP GET messages for every user on the base.

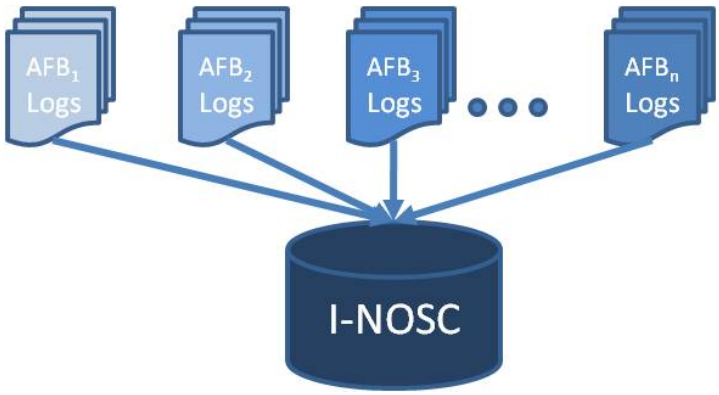

Figure 4. Proxy log data is uploaded to the I-NOSC for centralized storage

3.4 Sampling Strategy

Since specific Air Force bases are used in this research, a non-probability purposive sampling strategy is employed. Specifically, the convenience data-set used for analysis was derived from a non-proportional quota sampling procedure. In this method, a minimum number of search histories are collected for each Air Force base.

3.4.1 Sampling Frame

The category for collection is a single entity that acts as a gateway for all ingress and egress Internet traffic for a collection of IP's. The conditions of having numbers that match the proportions in the population are not paramount. Instead, there needs to be enough samples from each category to assure that relative representativeness for even small groups in the population. This method is the non-probabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in the sample.

3.4.2 Sampling Selection

In order to have an adequate sample set, the server logs from nine Air Force Bases were collected for a contiguous 32 day period. The bases represented in the sample are: Charleston, Dover, Grand Forks, Little Rock, MacDill, McConnell, McGuire, Scott, and Travis.

The steps required to obtain the historical data logs for this research are outlined in Figure 5. Each base uploads its logs daily to the I-NOSC. The logs are downloaded from the I-NOSC to a personal computer for storage. The logs are then filtered into the

desired format by creating new text files containing the data necessary for analysis. Finally, these new text files are used to create a search history for each unique IP address.

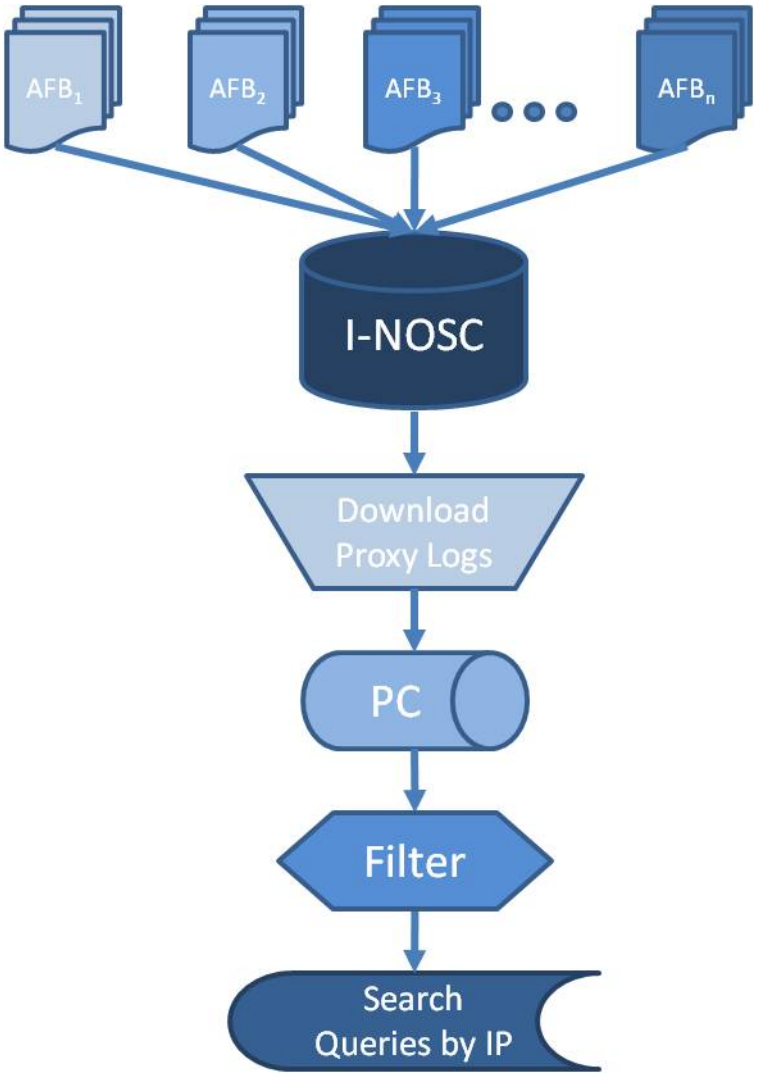

Figure 5. Data flow from AFB logs to individual search histories by IP Address

3.5 Raw Data File Format

The logs downloaded from the I-NOSC consist of a one line entry for each outgoing HTTP GET message; including information on the date, time, IP address, website, and browser. The logs are saved by date. The daily logs for each location were too large for

a single file, requiring multiple files per day by location to be stored. Each GET message is categorized according to the BlueCoat Proxy into one or several of the 84 categories listed in Appendix A. One function of the BlueCoat proxy is to categorize billions of web pages into useful categories that can easily be managed by IT administrators. Note that since this research focuses on changes in online search patterns, it is only concerned with messages that are categorized as Search Engines/Portals.

3.6 Data Pre-Processing/Cleaning

3.6.1 Phase 1

Since this research focuses on changes in online search patterns, the only relevant GET messages are those categorized as Search Engines/Portals. Therefore, each log is parsed by a Perl script which saves those messages to a new file. This script is included in Appendix B.1. This new file uses the name of the original (i.e., pre-parsed file) with "_Searches" appended to the end (hereafter referred to as Search Logs). This process is depicted in Figure 6. In order to ensure the quality of the parsed data, the Perl script was tested on a proxy log with fewer than 100 lines so the output could be visually verified.

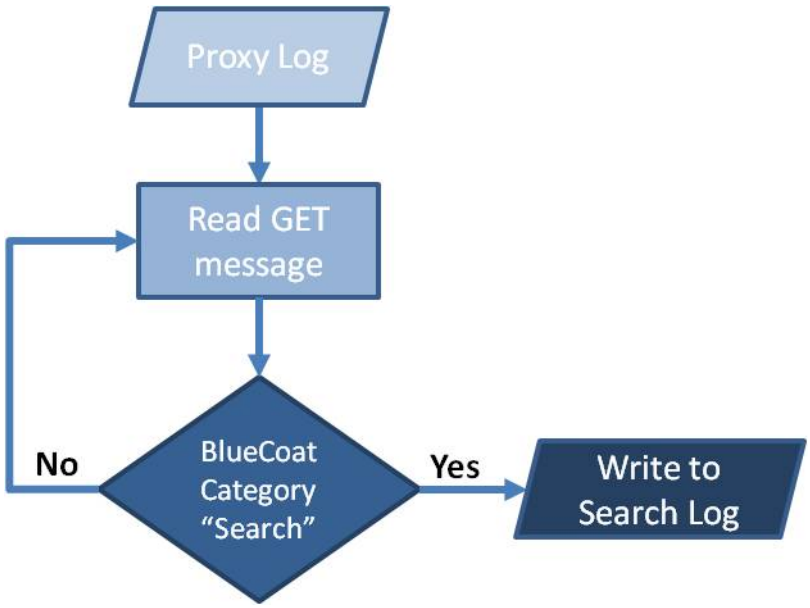

Figure 6. Phase 1 process to create Search Logs

3.6.2 Phase 2

Each of the Search Logs is processed through another filter written in Perl, which separates the actual searches from every GET message categorized by the BlueCoat proxy as "Search." This filter is included in Appendix B.2. Other metadata of interest includes the date, client IP, client-server category, the client-server referrer, the router-server content type, the client-server URI-path, and the client-server URI-query. The client-server referrer identifies which search engine is being queried. The content type distinguishes between web-pages and other web-content. The client-server URI-path ensures that it is a search, and the client-server URI-query is the search-query itself.

Figure 7 shows this process in detail. The filter traverses each Search Log in the directory, reading the GET message and performing several checks to ensure the message represents an actual search. Actual searches have the following attributes: content type of HTML, URI path labeled search, and referrer with the name of a search engine. If any of

the checks fail, the filter moves to the next GET message. If the GET message has each of the required attributes, it is written to a new file. This aggregate file contains the online search history for the given location for the entire period of the data-set in chronological order.

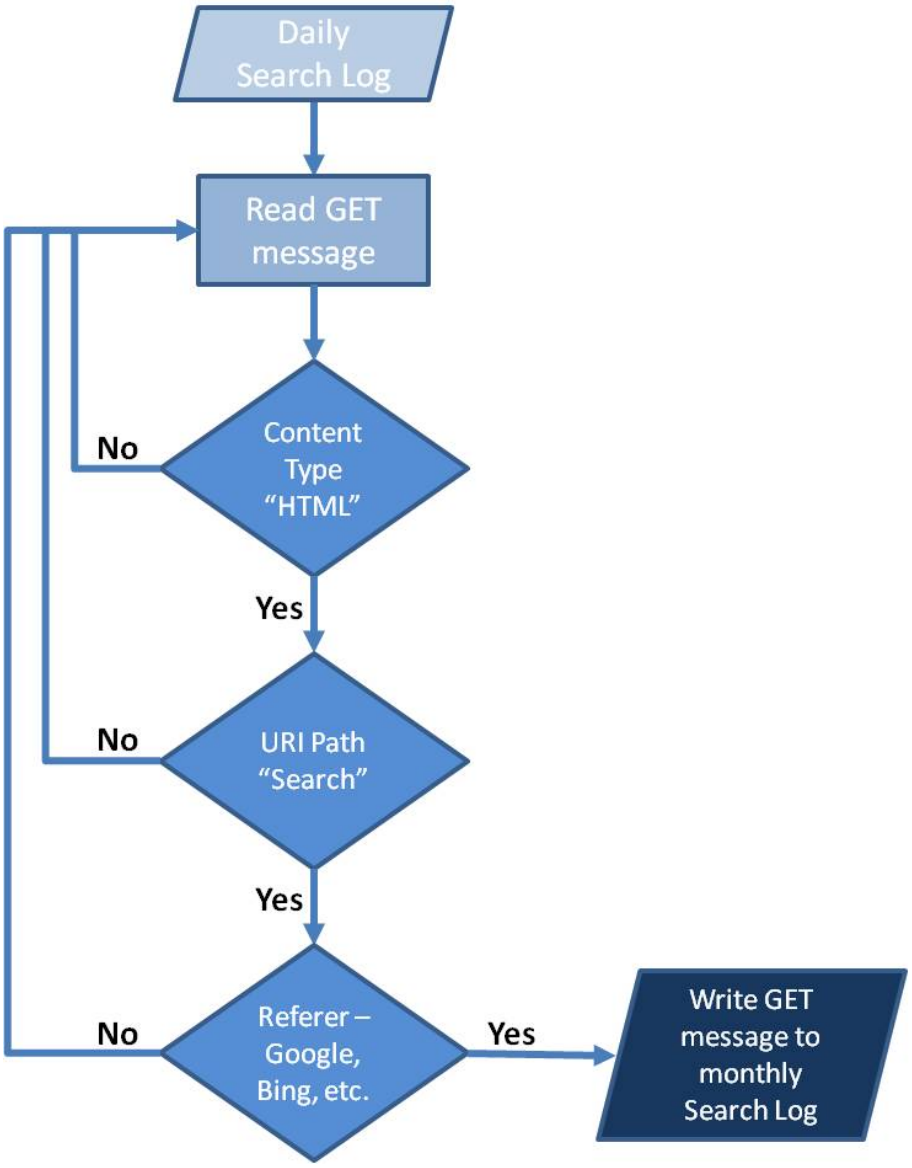

Figure 7. Phase 2 filter, which writes the actual searches from the daily Search Logs to a single file representing the searches from a given base for the entire month

After this process is performed for each location, there are nine Search Logs containing all the data required to complete the remainder of the research. This time, the Perl script was tested on a Search Log with fewer than 100 lines categorized as "Search" so the output could be visually verified.

3.6.3 Phase 3

Each of the logs are then processed by another Perl script, which outputs a file for each IP Address, containing the online search history for that IP throughout the entire month. This code is included in Appendix B.3. Figure 8 describes this flow.

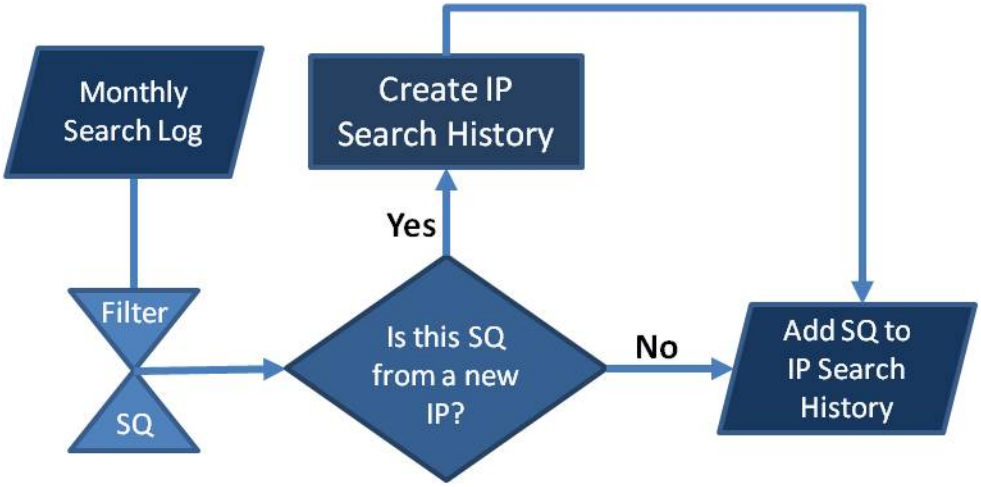

Figure 8. Phase 3 filter, creating a search history for each IP address

The search query is contained within the cs_uri_query field of each GET message. In order to parse the search query from the cs_uri_query field, the filter progresses through several steps which are outlined in Figure 9. Depending on the referrer, the filter parses different portions of the cs_uri_query. Regardless, the string passed to the subroutine responsible for cleaning and printing the actual search query contains an extra character on the back-end, which the filter removes. This extra character represents the end of the search query. In order to standardize all search queries, the filter converts the text to

lower-case, hex to ASCII, and removes punctuation and extra whitespace on the front and back-ends. In order to ensure the search queries are filtered correctly, the filter was tested on multiple URI queries from each search engine found within the data set.

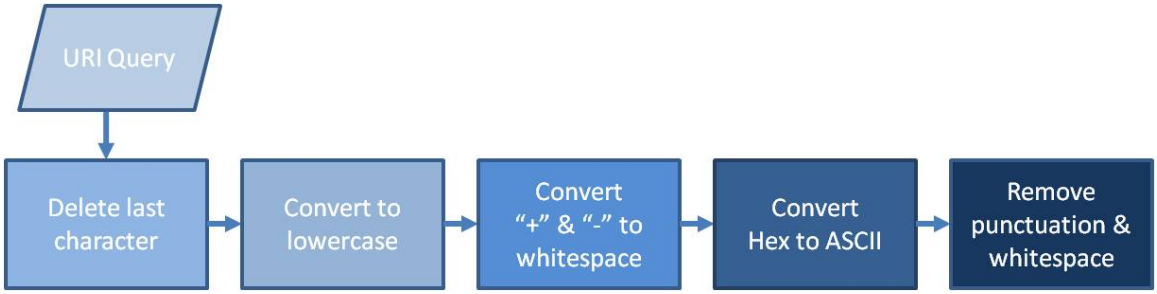

Figure 9. Process to filter and clean the actual search queries from the URL

Figure 10 provides an example search history file output by the filter. When the process is complete, a unique search history file is saved for each unique IP address, and contains the search queries for the user.

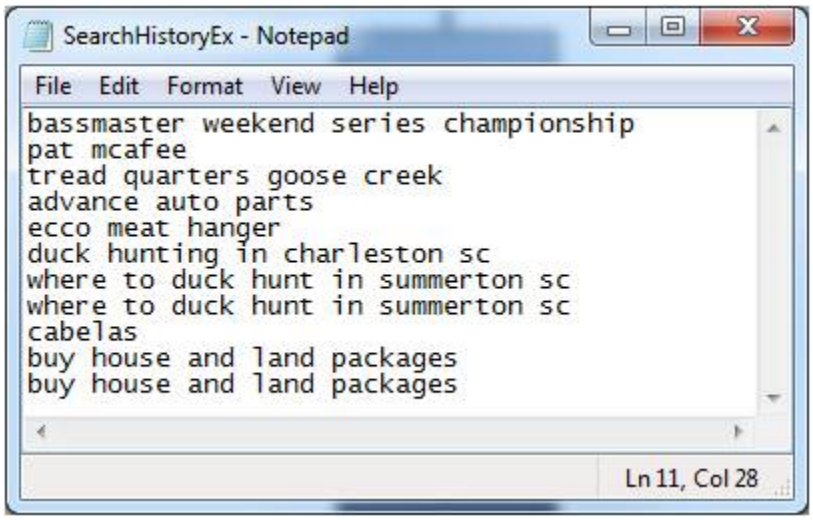

Figure 10. Example Search History

In order to ensure the quality and consistency of data, there are several criteria each search history document must meet before being included in the analysis. First, the document must contain a minimum of three lines representing three valid searches. Less

than three searches is not sufficient information to derive behavior. In order for a search to be considered valid, it must contain English word(s). Searches made up exclusively of numbers or words from another language are also excluded, as they do not provide a sufficient representation to derive behavior.

3.7 Base Analysis

Once the data is in the desired format, diagnostics are conducted to gain an understanding of what is included in the data. Several features are calculated, including the total number of searches, average number of searches per day, the total number of unique IP addresses, the average number of searches per IP, the total number of search terms, and the average number of search term per search. Table 1 provides the data attributes, per base, for the period of the data set.

For example, the Charleston AFB had 32 days worth of proxy search logs for analysis. The total number of searches was 60,690 throughout the 32 days, which averaged out to 1,897 searches per day. There were 2,655 unique IP address responsible for generating the total number of searches, thus the average number of searches per IP address computed to 23. The total number of search terms was 214,430. Dividing the number of search terms by the total number of searches provides an average of four words per search for Charleston.

Table 1. Data attributes for each of the nine AFBs

AFB	Days	Searches	Avg Searches/Day	IPs	Avg Searches/IP	Search Terms	Avg Words/Search
Charleston	32	60690	1897	2655	23	214430	4
Dover	32	50716	1585	2321	22	185004	4
Grand Forks	31	17842	576	858	21	66766	4
Little Rock	32	46627	1457	2343	20	161019	3
MacDill	29	54132	1867	2523	21	185040	3
McConnell	32	27470	858	1577	17	97940	4
McGuire	29	96158	3316	3973	24	338392	4
Scott	30	56893	1896	4152	14	196838	3
Travis	31	73228	2362	3971	18	252613	3
Totals	NA	483756	1757	24373	20	1698042	4

3.8 LDASOM

Retrieving information from large collections of text documents can be aided through clustering and visualization. Previous research has been conducted on a clustering and visualization method based on Latent Dirichlet Allocation and self-organizing maps (LDASOM) [42]. The data is in the necessary format for evaluation – one month-long search history for each of the over 24K unique IP addresses.

Figure 11 provides a top-level depiction of how LDASOM works. The search histories are all saved in a single folder, which is read by LDASOM. LDASOM implements a probabilistic topic model to cluster documents, rendering them in an intuitive graphical two and three-dimensional format. In order to maintain the 10:1 dimensionality recommended for LDA, the number of topics is set to 300 [50].

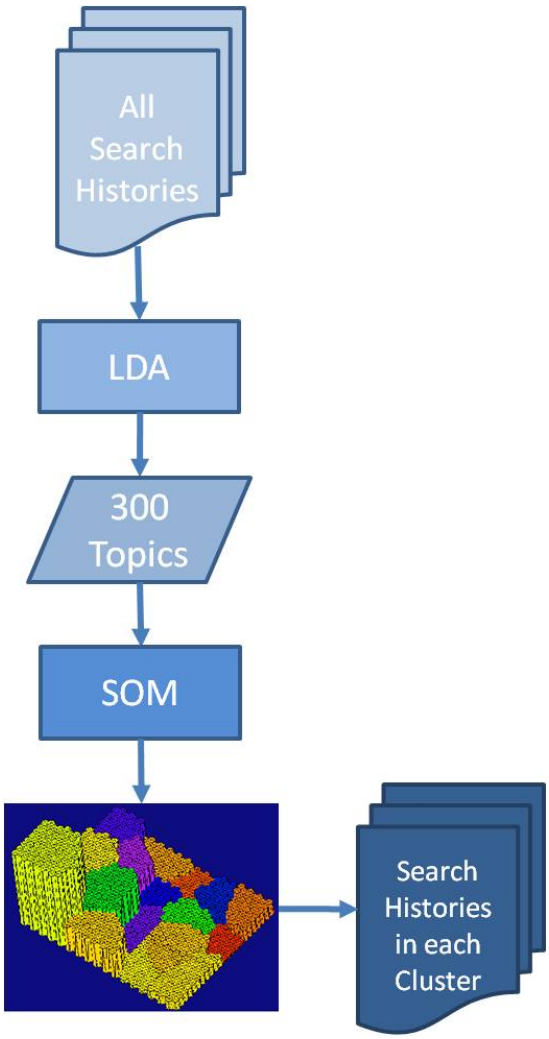

Figure 11. Top-level flowchart for LDASOM

3.8.1 Topic Analysis

The 300 topics determined by LDASOM are included in Appendix C. These topics describe the words most likely to be found in the same search history by calculating the probability of their occurrence throughout all the search histories in the sample set. They are represented similar to definitions in a dictionary and are simply a string of the words best representing the words most often found together through the search histories.

For example, topic 22 is defined as: “calendar santa holidays rose candy julian boot nuclear claus dec.” This topic describes search histories which included search queries about the holidays. Because the topic list is saved as a text file, it can quickly be searched for interesting terms. For example, a search for the word *divorce* produces a hit in topic 5, “park child divorce support laws coloring masks suicide knives comic.” D*ivorce* is most likely found in search histories which also contain searches for the other words in the topic.

3.8.2 Cluster Analysis

Once LDA is finished computing the probabilities for all the search histories relative to the topics, LDASOM clusters the search histories using self-organizing maps. The topology preserving properties of SOMs ensures that documents with similar topic distributions are clustered near one another. While each of the 300 topics computed by LDA includes probabilities for each of the 24K documents, the SOM clusters with a 1:1 matching of document to cluster, ensuring mutual exclusion between clusters.

SOMs are a form of unsupervised neural network capable of grouping the input data so that those with similar features are mapped to the same or neighboring clusters. The SOM provides a visual approach to represent search histories the algorithm determined are related, without any human interaction. Each cluster shows the top three topics, based on the topic distribution for each document within the cluster. It is possible to see exactly which search histories are included within a cluster by clicking on it. Code was added to LDASOM which made it possible to export the names of the files within the cluster to a separate text file. This process is completed for each cluster, and these

text files are saved to the same folder. At this point, the folder contains a text file for each cluster, containing the names of the search histories included in it. Together, these cluster files contain the names of every one of the 24K IPs.

The SOM in Figure 12 provides a 3D visual representation of how the search histories were clustered based on topic probability. The number on each cluster is the ID number for each cluster and is provided as output from LDASOM. Note that the ID number is for cluster identification only and does not imply any attributes associated with classifying the data. In addition to the ID number, clusters can also be distinguished by color. The elevation of a cluster represents the number of search histories included in that cluster – the higher the elevation, the greater the number of search histories. Each search history is associated with only in one cluster in the SOM landscape.

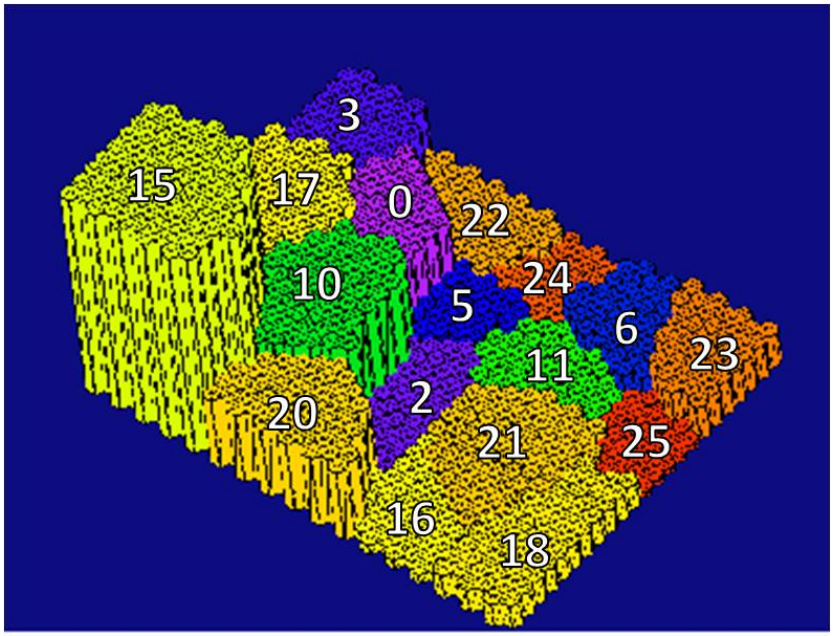

Figure 12. SOM generated by LDASOM for +24,000 valid search histories

The next step involves analyzing the search histories in each cluster. The total number of search histories, searches, and average number of searches per document are

calculated and shown in Table 2. For example, ClusterID0 contained 2,058 unique search histories. Those search histories contained 46,917 searches, for an average of 23 searches per search history. Descriptive analysis is also conducted to determine how each geographical location is represented within each cluster. Within ClusterID0, Charleston AFB had 275 search histories, Dover had 153, Grand Forks had 66, Little Rock had 224, MacDill had 294, McConnell had 134, McGuire had 371, Scott had 333, and Travis had 208.

There are a few interesting observations worth noting. For example, ClusterID11 is dominated by search histories from Travis AFB in California. The top three topics based on probability for that cluster were topic 200 (53.48%), topic 229 (9.10%), and topic 274 (8.88%). Topic 200 has a significantly higher probability than the next two highest topics. Topic 200 is "ca vacaville fairfield sacramento vallejo solano suisun kaiser walnut Roseville." A majority of search histories in ClusterID16 are also from Travis, and once again, the highest topic probability is topic 200 (43.48%).

Table 2. Cluster analysis

ClusterID	Docs	Searches	Avg Searches/Doc	Docs per Air Force Base								
				Charleston	Dover	Grand Forks	Little Rock	MacDill	McConnell	McGuire	Scott	Travis
ClusterID0	2058	46917	23	275	153	66	224	294	134	371	333	208
ClusterID10	2962	53438	18	373	305	155	282	286	245	456	561	299
ClusterID11	113	1969	17	2	2	0	2	0	1	2	2	102
ClusterID15	7474	129264	17	819	764	278	698	773	546	1210	1521	865
ClusterID16	317	6320	20	14	14	5	27	21	5	19	37	175
ClusterID17	2582	48785	19	328	327	91	284	261	140	506	405	240
ClusterID18	845	13658	16	46	58	17	72	61	50	126	96	319
ClusterID2	101	1851	18	5	0	1	2	2	0	6	2	83
ClusterID20	2547	52412	21	303	222	101	263	251	177	406	506	318
ClusterID21	384	6599	17	2	1	0	2	0	1	3	6	369
ClusterID22	498	7262	15	23	26	13	33	35	21	33	64	250
ClusterID23	1182	29557	25	109	127	39	126	138	73	238	156	176
ClusterID24	157	4043	26	10	9	4	4	10	6	11	13	90
ClusterID25	178	4009	23	14	22	3	7	11	6	39	12	64
ClusterID3	1974	37292	19	209	201	68	228	247	126	325	298	272
ClusterID5	269	6227	23	35	15	3	19	32	12	51	52	50
ClusterID6	731	34135	47	88	75	14	70	101	34	171	88	90
Totals	24372	483738	21	2655	2321	858	2343	2523	1577	3973	4152	3970

3.9 Summary

The methodology described in this chapter outlines the approach used to classify search histories and group users with similar behaviors. LDASOM makes it possible to categorize and cluster these search histories to gain knowledge about users. The remainder of this thesis describes the knowledge gained as a result of analyzing the clustering of these search histories.

VI. Determining Relationships

Chapter 3 outlined the general concepts and approach used to classify search histories and group users with similar behaviors. Relationships do exist between the search histories of individuals – evidenced by the SOM clustering. Chapter 4 describes the methodology, results, and analysis of determining relationships based on online search histories.

4.1 Air Force Suicide Prevention

The Air Force has devoted a considerable amount of resources on suicide awareness and prevention, spearheaded by the United States Air Force Suicide Prevention Program (AFSPP). A significant portion of these resources are devoted to providing family, friends, and co-workers with information on identifying at-risk personnel. AFSPP states that these three groups are in the best position to recognize behavioral changes, discuss these changes with the at-risk individual, and provide care and support [6]. Behavioral change in itself does not imply someone will become suicidal. However, according to AFSPP, individuals exhibiting changes in one or more of the following may warrant close monitoring [6]:

- Mood
- Concentration
- Sleep pattern
- Energy
- Appetite
- Substance use
- Impulse Control
- Reduced capacity for enjoyment
- Helplessness or hopelessness
- Peer Relations
- Work Performance

- Military bearing
- Personal hygiene and grooming
- Ineffective problem-solving

Last year, a committee made up of members of the Uniformed Services University of the Health Sciences submitted their findings in the Journal of Affective Disorders in an article titled "Suicide in the United States Air Force: Risk factors communicated before and at death." In this article, Cox *et al.* [9] described how their project aimed at describing and evaluating the communications (i.e., verbally and in suicide notes) of 13 suicide risk factors in the suicide death investigation files of 98 active duty airmen. Their findings support the USAF emphasizing certain risk factors over other suicide prevention efforts. They concluded that interpersonal risk factors appeared to be more salient than intrapsychic risk factors in the minds of the decedents. The last section of the article describes the limitations of the research – the most significant being that researchers must depend on previously collected information for analysis [9].

A recent article in the LA Times [51] discusses how social media serves as a lifeline for many members of military families – connecting them to supportive communities to help cope with specific strains and stresses. Facebook, in conjunction with the Department of Veterans Affairs and Blue Star Families, has unveiled a lifeline from within their website that includes informational and response tools customized for service members and their families. This new tool enables those connected to veterans or active-duty military and their families, to obtain specific information about crisis services

tailored to the military. These services include the Veteran's Crisis Line, which can respond over the phone, through online chat, or by text message.

4.2 Cyber Indicators

The initiatives on analyzing both real-world and cyber indicators of suicide raise a question worth answering. Is it possible that similar factors exist within cyberspace that do not necessarily mean an individual is suicidal, but may warrant close monitoring? If so, what are these factors and how can they be monitored? The simplest method is a dirty-word search – where individuals who search n-times for blacklisted term(s) are put on watch. Unfortunately, this type of system is one-dimensional. Doctors in Taiwan conducted a research effort along these lines, when they examined search query data from 2004 through 2009 for trends in searches related to suicide. They compared the frequency to truth data provided by the State and found a relationship between increases in suicide related searches and failed/successful suicides [14].

Despite its one-dimensionality, search query analysis could have a significant impact on the most significant limitation in traditional suicide research – a dependence on previously collected data [9] – if conducted in real-time. However, an even greater impact may be made by adding a second dimension – clustering the online search histories of users and identifying relationships with at-risk individuals.

4.3 Disorder Determination

In order to determine which search histories should be targeted, a professional psychologist was asked to provide a list of suicide-related disorders. This list consisted

of anxiety, depression, PTSD, stress, and suicide. A dictionary was constructed for each disorder, made up of words associated with each. In order to make the dictionaries more sufficiently exclusive, terms associated with Stress Disorders were combined with PTSD and terms associated with Depression Disorders were combined with Suicide. The dictionaries were sufficiently exclusive, though not exhaustive. The following sections describe each disorder and the words included in each dictionary.

4.3.1 Anxiety Disorder

Anxiety disorder is a general term used for many different forms of abnormal or pathological anxieties, fears, and phobias. Although *fear*, *anxiety* and *phobia* are often used interchangeably in conversation, clinically they have different meanings. Anxiety is described as an unpleasant emotional state [52]. Unfortunately, the causes can be difficult to identify. In order to receive effective treatment and a better prognosis, it is important to distinguish between different anxiety disorders. Anxiety disorders are the most common psychiatric illness affecting both children and adults and are frequently accompanied by physiological symptoms that may lead to fatigue or even exhaustion [52]. The dictionary provided in Table 3 shows which words and compound words were used to filter search queries related to anxiety.

Table 3. Dictionary of words associated with anxiety disorders

anxiety	irritability	suffering
angst	misery	tension
anxiousness	nervousness	uncertainty
apprehension	obsessive compulsive	unease
apprehensiveness	panic	uneasiness
desperation	phobia	worry
distress	restlessness	worries
dread	solicitude	worriment
jitters		

4.3.2 Post-Traumatic Stress Disorder (PTSD)

Stress is the body's normal response to anything that disturbs its natural physical, emotional, or mental balance. Everyday stressors can be managed with healthy stress management behaviors, but untreated chronic stress can result in health conditions including anxiety, insomnia, muscle pain, high blood pressure and a weakened immune system [53]. Chronic stress is known to contribute to anxiety and depression, which greatly increases the risk for heart disease [54]. Additionally, people exposed to chronic stress are at a heightened risk of developing a drug addiction [55]. If a high stress level continues for a long period of time, it is important to reach out to a licensed mental health professional [56].

Post-Traumatic Stress Disorder affects thousands of veterans by causing the brain to sense stress and danger regardless of the situation and at unexpected times [57]. Symptoms include:

- Flashbacks
- Bad dreams
- Frightening thoughts
- Avoiding places, events or objects that are reminders of the traumatic event
- Feeling emotionally numb

- Feeling strong guilt, depression or worry
- Losing interest in activities that were enjoyable in the past
- Having trouble remembering the traumatic event
- Being easily startled
- Feeling tense or on edge
- Having difficulty sleeping

According to the National Institute of Mental Health, a person must have all of the following symptoms for at least one month in order to be diagnosed with PTSD [57]:

- At least one re-experiencing symptom
- At least three avoidance symptoms
- At least two hyperarousal symptoms
- Symptoms that make it hard to go about daily life, go to school or work, be with friends or family, and take care of important tasks

The dictionary provided in Table 4 shows which words and compound words were used to filter search queries related to PTSD.

Table 4. Dictionary of words associated with PTSD

PTSD	operational exhaustion	prison
Stress	shell shock	rape
combat disorder	burden	flashback
combat fatigue	traumatize	nightmare
combat neurosis	assault	hyper vigilance
complete exhaustion	domestic abuse	survivor guilt

4.3.3 Suicide

Depression is a serious medical illness where long-lasting feelings of sadness, anger, loss, or frustration get in the way of life. Clinical depression is categorized from mild to severe and symptoms include [58]:

- Difficulty sleeping or too much sleeping
- Fatigue and lack of energy

- A drastic change in appetite, accompanied by weight gain or loss
- Self-loathing and low self-worth
- Trouble focusing
- Feeling exasperated and helpless
- Anxiety
- Restlessness
- Agitation
- Inactivity and detachment
- Frequent thoughts of suicide

Unfortunately, many people with a depressive illness never seek treatment [59]. Depression is considered a common but serious illness, and most people who experience depression require treatment in order to get better. Research indicates that depressive illnesses are disorders of the brain, and are most likely caused by a combination of genetic, biological, environmental, and psychological factors [59].

Many symptoms of depression are indicators for individuals at-risk for suicide [6]. Suicide is the second leading cause of death, after accidents, among active duty U.S. military members [60]. The Department of Veterans Affairs provides the following suicide warning signs:

- Talking about wanting to hurt or kill oneself
- Trying to get pills, guns, or other ways to harm oneself
- Talking or writing about death, dying, or suicide
- Hopelessness
- Rage, uncontrolled anger, seeking revenge
- Acting in a reckless or risky way
- Feeling trapped, like there's no way out
- Saying or feeling there's no reason for living
- Calling old friends, particularly military friends, to say goodbye
- Cleaning a weapon that they may have as a souvenir
- Visits to graveyards
- Obsessed with news coverage of the war, the military channel

- Wearing their uniform or part of their uniform, boots, etc
- Talking about how honorable it is to be a soldier
- Sleeping more (sometimes the decision to commit suicide brings a sense of peace of mind, and they sleep more to withdraw)
- Becoming overprotective of children
- Standing guard of the house, perhaps while everyone is asleep staying up to "watch over" the house, obsessively locking doors, windows
- If they are on medication, stopping medication and/or hoarding medication
- Hoarding alcohol -- not necessarily hard alcohol, could be wine
- Spending spree, buying gifts for family members and friends "to remember them by"
- Defensive speech, "you wouldn't understand"
- Stop making eye contact or speaking with others

The dictionary provided in Table 5 shows which words and compound words were used to filter search queries related to suicide.

Table 5. Dictionary of words associated with suicide

suicide	repeater	bleakness
abuse	schizophrenia	dejection
asphyxiation	self deliverance	desolation
autocide	self destruction	despondency
bipolar	self harm	disconsolation
divorce	self immolation	discouragement
end it all	self murder	dispiritedness
ideation	self mutilation	dreariness
insomnia	self slaughter	gloom
kill yourself	suicidal	hopelessness
new normal	suidecide	melancholy
noose	suisad	misery
overdose	take your own life	mortification
penacide	top yourself	sorrow
perturbation	wrist slitting	unhappiness
pre disposing risk factor	Depression	woefulness
reasons for dying	despair	

4.4 Disorder-Related Searches

After building the dictionaries, a Perl script is used to filter each of the 24K search histories for search queries which include a term from one or more of the dictionaries. The filter begins by loading a search history and analyzing each search query. If the query contains a word(s) from a disorder dictionary, the name of the search history is written to a text file saving the names of all search histories that searched for that disorder. These text files are named AnxietyIPs.txt, PTSDIPs.txt, and SuicideIPs.txt. In addition to the name of the search history, the number of search queries containing disorder-related terms is saved as well.

If a search history includes terms from multiple disorders, the name of the search history is saved to the disorder with the majority of hits. In the case that terms from multiple disorders are searched for the same number of times, the name of the search history is randomly assigned to one of the disorders tied for the highest term hit count. This process is outlined in Figure 13 and the Perl script used to execute this filter is included in Appendix B.5.

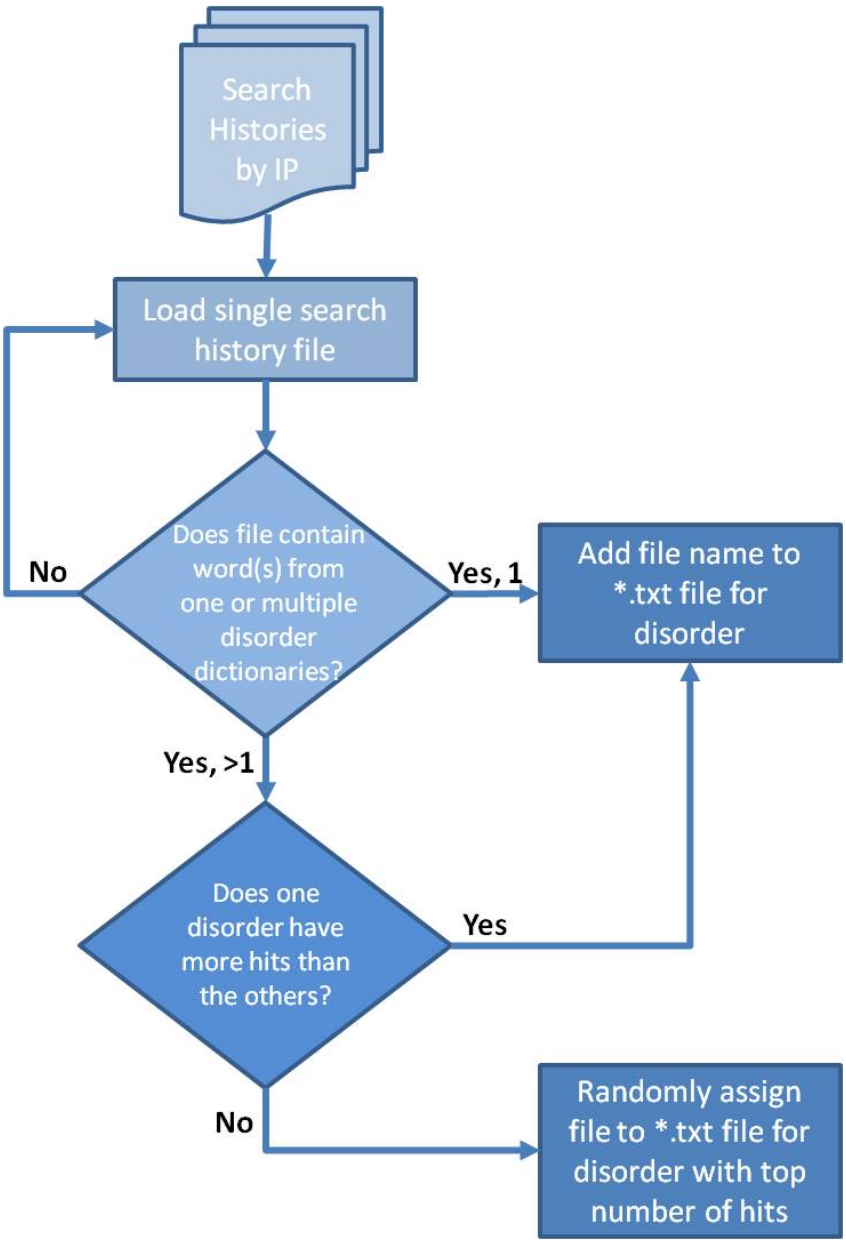

Figure 13. Determining search histories which contain disorder-related searches

4.5 Contingency Tables

The disorder IP text files provide the first dimension of analysis by saving the IPs which have performed searches related to the given disorders. In order to get to the second dimension, it is necessary to determine where these search histories are clustered within the SOM. This is accomplished through contingency tables where the rows represent the

clusters, while the columns represent the searches for each disorder. These tables are used to complete the analysis and evaluation portion of the research, and allow for the calculation of a chi-square (χ^2) statistic.

Figure 14 describes how the Perl script accomplishes the tasks necessary to build the contingency tables. The names and counts of the search histories are read from the disorder IP files, and loaded into an array for each disorder. The program then opens a cluster text file, which contains the name of every search history included in the cluster. If the name of the search history in the cluster text file is also included in one of the disorder IP files, the number of disorder-related hits associated with that search history is added to the total number of disorder-related hits for that cluster. This process continues until every search history in every cluster has been checked against the search histories with disorder-related searches. Once completed, the program outputs a comma-delimited-file containing the counts for the disorder-related searches for each cluster. The Perl script responsible for performing this analysis is in Appendix B.6.

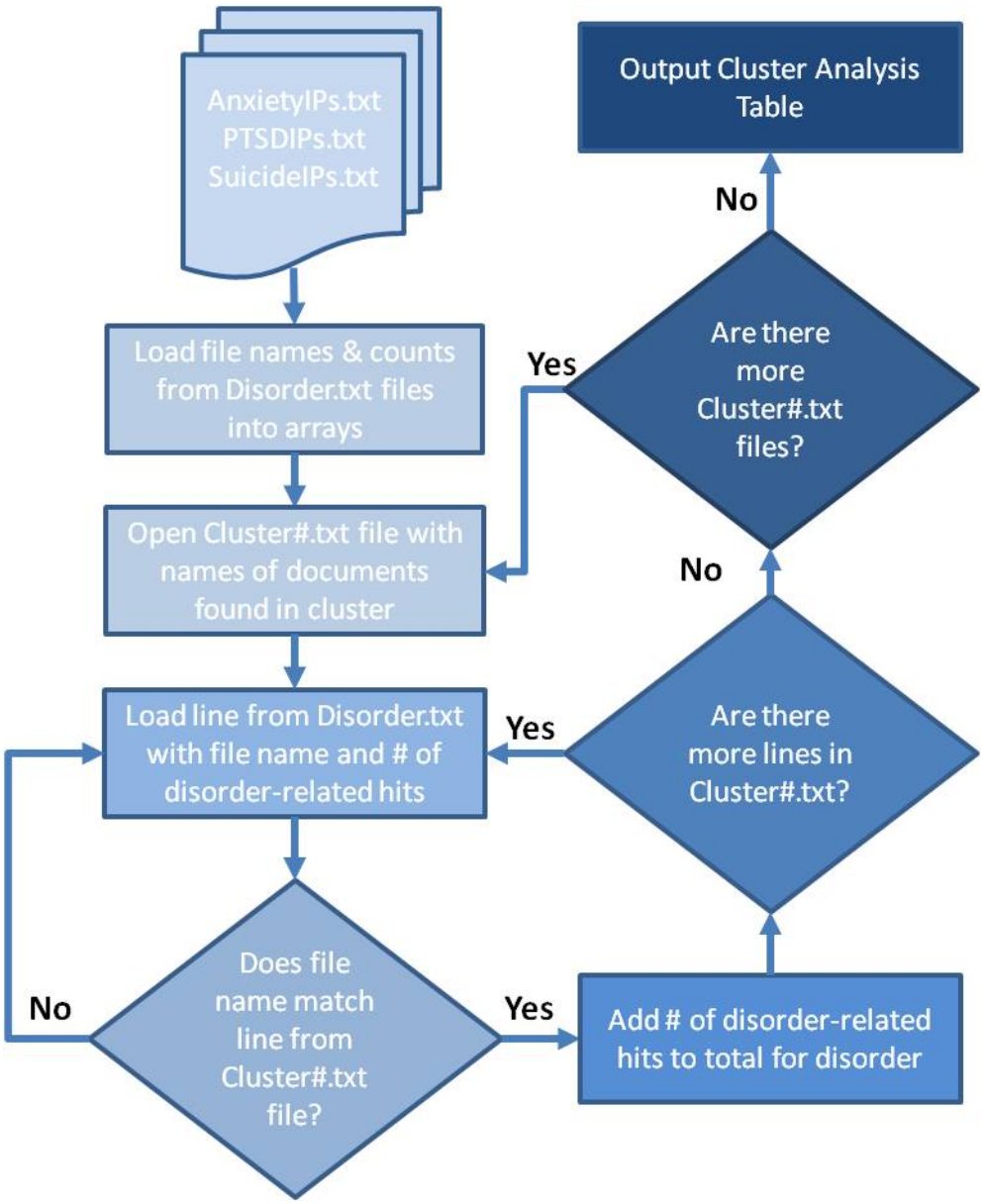

Figure 14. Process to find disorder-related search histories within the SOM clusters

Table 3 is the output from the cluster analysis just described. The numbers in the columns represent the number of searches for the given disorder within the given cluster. The totals on the right are the total number of disorder related searches within the given cluster. The total in the last row are the total number of searches for each disorder, across all clusters.

Table 6. Observed Frequency with 3 disorders and 17 clusters

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID0	1	19	14	34
ClusterID10	4	16	14	34
ClusterID11	0	0	0	0
ClusterID15	15	25	64	104
ClusterID16	0	0	1	1
ClusterID17	2	9	9	20
ClusterID18	3	3	3	9
ClusterID2	0	0	2	2
ClusterID20	6	8	7	21
ClusterID21	1	1	4	6
ClusterID22	0	0	0	0
ClusterID23	1	3	8	12
ClusterID24	0	0	0	0
ClusterID25	0	3	1	4
ClusterID3	2	4	26	32
ClusterID5	0	0	2	2
ClusterID6	2	8	6	16
Totals	37	99	161	297

4.6 Analysis and Evaluation Technique

The primary analysis technique is the Chi-square statistic. This is a non-parametric technique. A chi-square (χ^2) statistic is used to investigate whether distributions of categorical variables differ from one another by testing for independence and not a goodness-of-fit. The observed and expected frequencies play an important role in the χ^2 calculation. The results of the independent test state whether a relationship exists, but not how strong the relationship is. The hypothesis for this study is:

- H_0: The clusters and disorder categories are independent (i.e., there is no relationship between the clustering and the disorder-related searches)
- H_a: The clusters and disorder categories are not independent (i.e., there is a relationship between the clustering and the disorder-related searches)

4.7 Combining Clusters

The first step after generating the observed frequency table is to eliminate any rows with straight zeros, since those clusters contain no search queries with disorder-related terms. Within Table 3, clusters 11, 22, and 24 all contain zero disorder-related searches. Table 4 shows the reduced observed frequency table with those clusters removed.

Table 7. Observed frequency with clusters 11, 22, and 24 removed

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID0	1	19	14	34
ClusterID10	4	16	14	34
ClusterID15	15	25	64	104
ClusterID16	0	0	1	1
ClusterID17	2	9	9	20
ClusterID18	3	3	3	9
ClusterID2	0	0	2	2
ClusterID20	6	8	7	21
ClusterID21	1	1	4	6
ClusterID23	1	3	8	12
ClusterID25	0	3	1	4
ClusterID3	2	4	26	32
ClusterID5	0	0	2	2
ClusterID6	2	8	6	16
Totals	37	99	161	297

At this juncture, the expected frequency table can be computed. Each cell in the expected frequency table (E_{ij}) is calculated by multiplying the marginal row total (r_i) with the marginal column total (c_j), and dividing the result by the sum of all disorder-related searches (n):

$$E_{ij} = \frac{r_i \times c_j}{n} \qquad \text{(Eq. 4)}$$

This process is continued until every cell in the contingency table is calculated. Table 5 represents the expected frequency table calculated using the observed values from Table 4.

Table 8. Expected frequency with clusters 11, 22, and 24 removed

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID0	4	11	18	34
ClusterID10	4	11	18	34
ClusterID15	13	35	56	104
ClusterID16	0	0	1	1
ClusterID17	2	7	11	20
ClusterID18	1	3	5	9
ClusterID2	0	1	1	2
ClusterID20	3	7	11	21
ClusterID21	1	2	3	6
ClusterID23	1	4	7	12
ClusterID25	0	1	2	4
ClusterID3	4	11	17	32
ClusterID5	0	1	1	2
ClusterID6	2	5	9	16
Totals	37	99	161	297

Once the expected frequency table is in place, a few diagnostics must be made about the values within the cells before moving forward to analysis. First, in order to rely on the results of the χ^2, no more than 20% of the values in contingency table can be less than 5. Second, no cell can have a value less than 1 since it will be used as a denominator in the next calculation. Currently 62% of the values are below 5, much higher than the 20% maximum required to ensure a reliable χ^2-test statistic. Additionaly, five of the values in the contingency table are less than 1. Two options are available to remedy these problems – either combine disorders along the x-axis or cluster's along the y-axis.

With only three disorders along the x-axis, it is unreasonable to combine disorders any further. Therefore, it was necessary to begin combining clusters in order to get the expected values less than 5 below the 20% ceiling. Every cluster was compared to its surrounding clusters by comparing the top three topic probabilities for each of the clusters. Figure 15 shows these comparisons. The left-most column contains the Cluster ID, as well as the number of disorder-related hits within that cluster. The next three columns represent the top three topic probabilities for that cluster. The remaining columns are for top three topic probabilities of the clusters that share a side with the cluster in the first column. Highlighted topic numbers represent a match. The cluster in the first column was combined with the cluster(s) highlighted its top three topic probabilities.

Since the goal of combining clusters is to minimize the number of values less than 5, clusters with the fewest number of disorder-related searches were targeted. For example, ClusterID2 only had two disorder-related searches. Its top three topics were Topic 200 (35.3%), Topic 138 (9.35%), and Topic 289 (7.89%). ClusterID2 shares a side with clusters 5, 10, 16, 20, and 21, as seen in Figure 16. Although ClusterID2 also shares a side with ClusterID11, it was excluded since it did not contain any disorder-related searches. The probabilities for the top 3 topics for each of the other clusters were compared to those for ClusterID2. The top 3 topics for ClusterID10 matched those of ClusterID2, so they were combined.

Top 3 Topic Probabilities																		
ClusterID0	1	274	2	184	3	134	17	275	169	10	138	200	289	5	200	274	229	3	274	260	127
34		15.98		8.91		6.98		12.25	11.37	7.39	11.51	8.57	6.65		24.96	18.74	9.42		13.09	12.27	11.72
ClusterID10	1	138	2	200	3	289	15	284	52	270	20	200	114	102	2	200	138	289	5	274	229	0	274	184	134	17	184	275	169
34		11.51		8.57		6.65		6.51	6.28	5.76	12.98	12.74	8.26		35.3	9.35	7.89		24.96	18.74	9.42		15.98	8.91	6.98		12.25	11.37	7.39
ClusterID15	1	284	2	52	3	270	20	200	114	102	10	138	200	289	17	184	275	169											
104		6.51		6.28		5.76		12.98	12.74	8.26		11.51	8.57	6.65		12.25	11.37	7.39											
ClusterID16	1	200	2	191	3	230	20	200	114	102	18	200	254	230	21	200	254	60	2	200	138	289							
1		43.48		7.21		7.02		12.98	12.74	8.26		39.83	33.19	18.7		63	12.07	6.48		35.3	9.35	7.89							
ClusterID17	1	184	2	275	3	169	15	284	52	270	10	138	200	289	0	274	184	134	3	274	260	127							
20		12.25		11.37		7.39		6.51	6.28	5.76		11.51	8.57	6.65		15.98	8.91	6.98		13.09	12.27	11.72							
ClusterID18	1	200	2	254	3	230	16	200	191	230	25	200	254	261	21	200	254	60											
9		39.83		33.19		18.7		43.48	7.21	7.02		40.23	23.92	13.56		63	12.07	6.48											
ClusterID2	1	200	2	138	3	289	20	200	114	102	16	200	191	230	21	200	254	60	5	200	274	229	10	138	200	289			
2		35.3		9.35		7.89		12.98	12.74	8.26		43.48	7.21	7.02		63	12.07	6.48		24.96	18.74	9.42		11.51	8.57	6.65			
ClusterID20	1	200	2	114	3	102	15	284	52	270	16	200	191	230	2	200	138	289	10	138	200	289							
21		12.98		12.74		8.26		6.51	6.28	5.76		43.48	7.21	7.02		35.3	9.35	7.89		11.51	8.57	6.65							
ClusterID21	1	200	2	254	3	60	16	200	191	230	18	200	254	230	25	200	254	261	2	138	289								
6		63		12.07		6.48		43.48	7.21	7.02		39.83	33.19	18.7		40.23	23.92	13.56		35.3	9.35	7.89							

Figure 15. Clusters and their top three topic probabilities

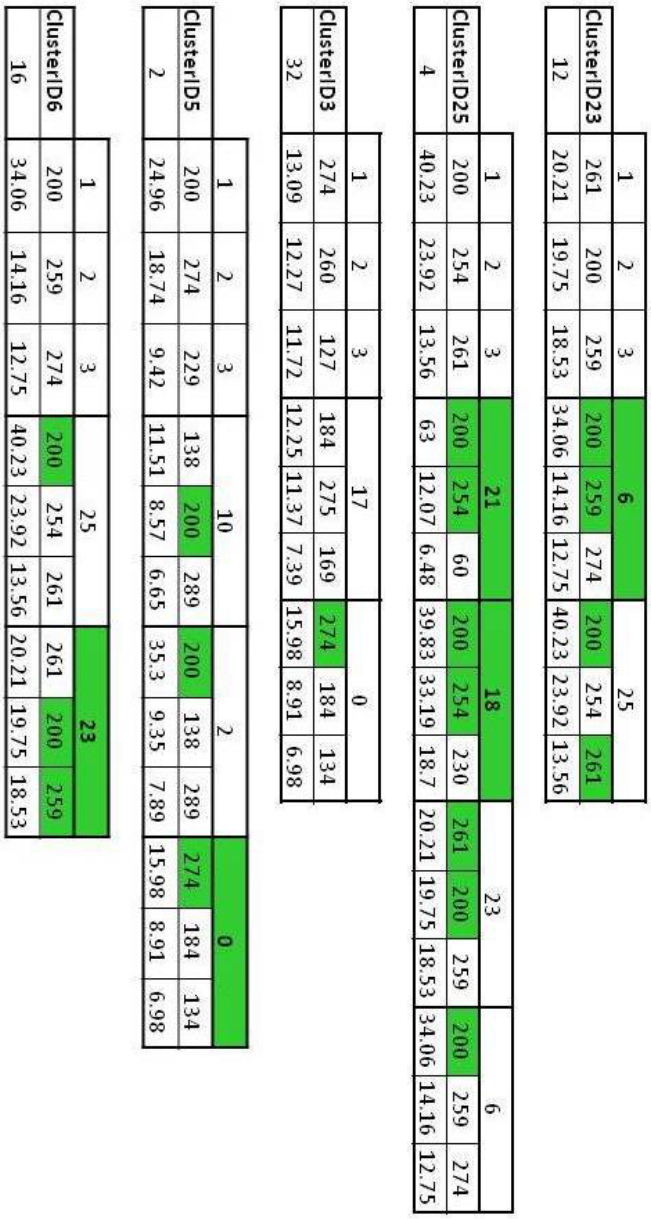

Figure 15. (continued) Clusters and their top three topic probabilities

The white lines outlining the clusters in Figure 16 show which clusters where combined once this first round of combinations was complete. Cluster 5 was combined with 0 (ClusterID0_5), clusters 2 and 20 with 10 (ClusterID10_2_20), clusters 16, 25,

and 21 with 18 (ClusterID18_16_25_21), and cluster 23 with 6 (ClusterID6_23). Table 6 provides the updated observed frequency table.

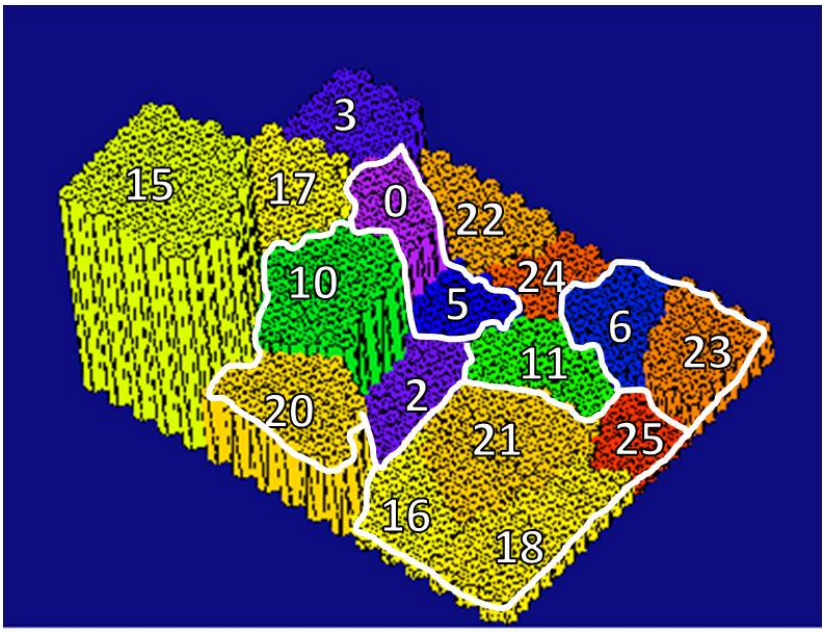

Figure 16. SOM with lines showing where clusters were combined

Table 9. Updated observed frequency table

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID0_5	1	19	16	36
ClusterID10_2_20	10	24	23	57
ClusterID15	15	25	64	104
ClusterID17	2	9	9	20
ClusterID18_16_25_21	4	7	9	20
ClusterID3	2	4	26	32
ClusterID6_23	3	11	14	28
Totals	37	99	161	297

The values from Table 6 are used to compute the updated expected frequency table, provided in Table 7. This process of combining clusters and calculated new

expected frequency tables continues until no more than 20% of the values are below 5, and no values are less than 1. At this point, 24% of the values are less than 5.

Table 10. Updated expected frequency table

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID0_5	4	12	20	36
ClusterID10_2_20	7	19	31	57
ClusterID15	13	35	56	104
ClusterID17	2	7	11	20
ClusterID18_16_25_21	2	7	11	20
ClusterID3	4	11	17	32
ClusterID6_23	3	9	15	28
Totals	37	99	161	297

Figure 17 shows the updated SOM cluster after the next round of combinations, where ClusterID0_5 is combined with ClusterID17 (ClusterID17_0_5). Tables 8 and 9 depict the updated observed and expected frequency tables.

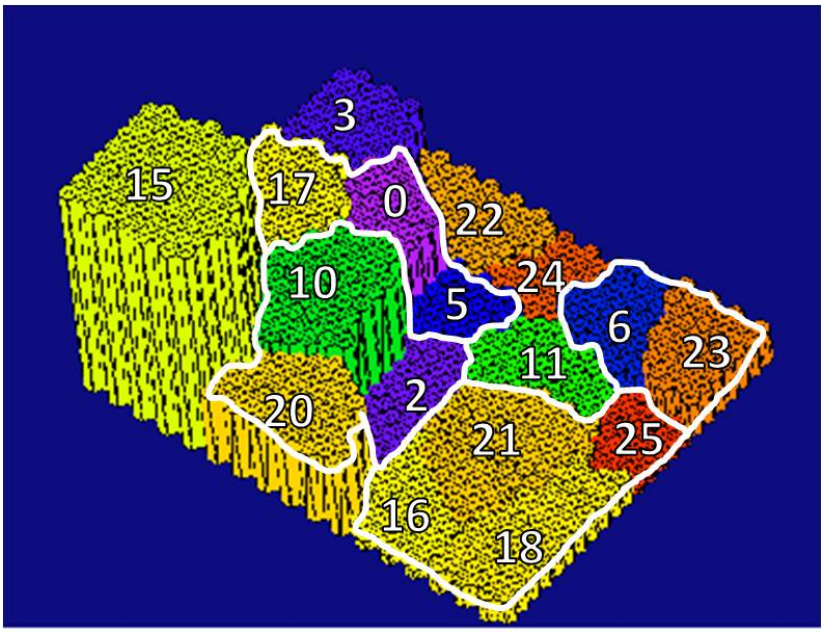

Figure 17. SOM with lines showing updated cluster combinations

Table 11. Updated observed frequency table

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID10_2_20	10	24	23	57
ClusterID15	15	25	64	104
ClusterID17_0_5	3	28	25	56
ClusterID18_16_25_21	4	7	9	20
ClusterID3	2	4	26	32
ClusterID6_23	3	11	14	28
Totals	37	99	161	297

Table 12. Updated expected frequency table

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID10_2_20	7	19	31	57
ClusterID15	13	35	56	104
ClusterID17_0_5	7	19	30	56
ClusterID18_16_25_21	2	7	11	20
ClusterID3	4	11	17	32
ClusterID6_23	3	9	15	28
Totals	37	99	161	297

Once again, the number of cells less than 5 in the expected frequency table is counted. At this point, 14% of the values are less than 5, below the 20% limit, and it is now possible to calculate the χ^2 test statistic to determine whether a relationship exists between the clusters and the IPs which searched for the disorders.

Each cell in the χ^2 test statistic is calculated by taking the square root of the observed value (o_i) minus the expected value (e_i) over the expected value. These values are then summed together to get the observed χ^2 value.

$$\frac{(o_i - e_i)^2}{e_i} \quad \text{(Eq. 5)}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \quad \text{(Eq. 6)}$$

Table 10 shows the table for the χ^2 test statistic. When the value from each cell is summed together, the χ^2 value is calculated at 28, as seen in the bottom-right cell of the table.

Table 13. χ^2-test statistic

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID10_2_20	1	1	2	5
ClusterID15	0	3	1	4
ClusterID17_0_5	2	5	1	8
ClusterID18_16_25_21	1	0	0	1
ClusterID3	1	4	4	9
ClusterID6_23	0	0	0	0
Totals	6	13	9	28

In order to determine whether a relationship exists between the clustering produced by LDASOM and the disorder-related searches, a few more steps must be executed. The degrees of freedom (df) are computed by taking one less than the number of rows times one less than the number of columns. With six row and three columns, the df is calculated at 10. In order to find the critical value, a χ^2-test lookup table is referenced with an α of .05 and 10 df. The alpha is equated with the p-value, which is the

probability of calculating a test statistic at least as extreme as the observed, assuming the null hypothesis is true. Table 11 shows a partial χ^2-test lookup table [61].

Table 14. χ^2-test table with alpha of .05

df	α, 0.05
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31

As stated in section 4.6, the hypothesis of this research is:

- H_0: The clusters and disorder categories are independent (i.e., there is no relationship between the clustering and the disorder-related searches)
- H_a: The clusters and disorder categories are not independent (i.e., there is a relationship between the clustering and the disorder-related searches)

If the calculated χ^2 value is less than the critical χ^2, than the null hypothesis is accepted - the clusters and documents are independent and no relationship exists. However, the calculated χ^2 (28) is greater than the critical (18.31) as referenced in the table. Thus the null hypothesis is rejected and it can be assumed with a greater than 95% confidence (α = .05) that a relationship does exist between the clustering and the targeted search queries.

4.8 Cramer's V

Cramer's V is useful for comparing multiple χ^2 test statistics and is generalizable across contingency tables of various sizes. It is useful in situations where suspicions exist that a

statistically significant χ^2 was the result of large sample size instead of any substantive relationship between the variables, since it is not affected by sample size. Cramer's V is a measure of the relative strength of an association between two variables, with a coefficient between 0 to 1, where 1 represents perfect association.

$$V = \sqrt{\frac{\chi^2}{n(q-1)}} \quad where\ q = smaller\ \#\ of\ rows\ or\ columns \qquad \text{(Eq. 7)}$$

A Cramer's V greater than .5 suggests a high association, .3-.5 moderate, .1-.3 low, 0-.1 little if any. These coefficients represent the general case, as in practice a Cramer's V of .10 may provide a good minimum threshold for suggesting there is a substantive relationship between two variables [62]. The calculated Cramer's V derived from the calculated χ^2 was .22, implying the strength of the association is low according to Table 12.

Table 15. Cramer's V characterizations

Value	Association
> .5	High
.3 to .5	Moderate
.1 to .3	Low
0 to .1	Little if any

However, an article in the 7 May 2012 edition of the Air Force Times highlighted the recent spike in the number of suicides and included statistics on the number of suicides per Air Force Base since 2003 [63]. The suicide totals per base are graphed in Figure 18. The suicide totals for the bases used in this research occur around the middle of the data set. Furthermore, the average number of suicides across all bases was 5.75,

while the average number of suicides across the nine bases used in this study was 5.86. These facts provided additional confidence that the data used in this research is representative of the population.

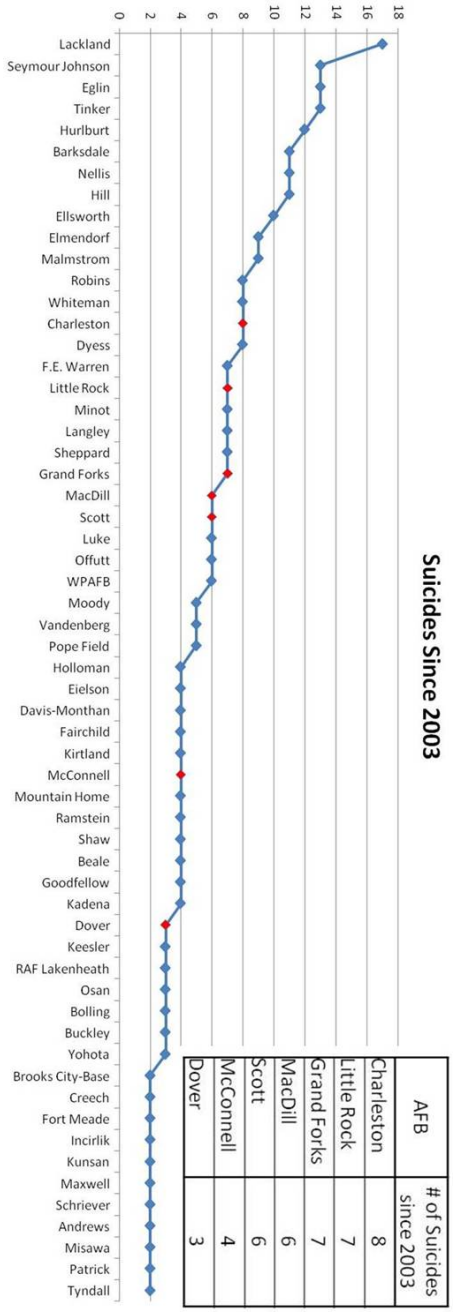

Figure 18. Air Force suicide totals since 2003

4.9 Patterns in Clusters

After analyzing the documents within the clusters, the next step is analyzing where the search histories that contained disorder-related searches fell across the SOM. Table 13 shows the clusters ranked from highest number of disorder-related searches to lowest. When those values are attributed to the clusters in which they occur, it is clear that most of the documents containing disorder-related searches were clustered in the top left of the SOM. The white line in Figure 19 outlines the six clusters with the highest number of disorder-related searches.

Table 16. Clusters from high to low based on number of disorder-related searches

ClusterID	Anxiety	PTSD	Suicide	Sum Total
ClusterID15	15	25	64	104
ClusterID0	1	19	14	34
ClusterID10	4	16	14	34
ClusterID3	2	4	26	32
ClusterID20	6	8	7	21
ClusterID17	2	9	9	20
ClusterID6	2	8	6	16
ClusterID23	1	3	8	12
ClusterID18	3	3	3	9
ClusterID21	1	1	4	6
ClusterID25	0	3	1	4
ClusterID2	0	0	2	2
ClusterID5	0	0	2	2
ClusterID16	0	0	1	1
Totals	37	99	161	297

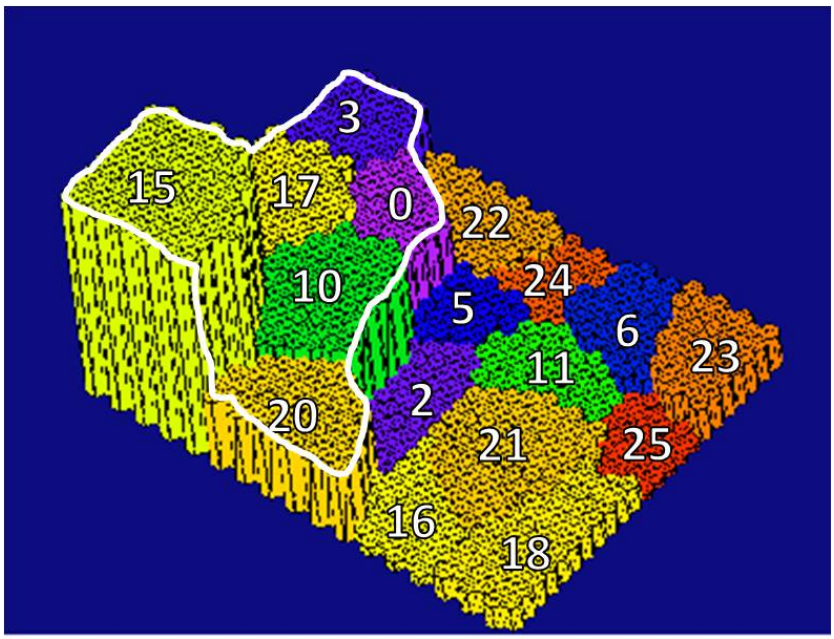

Figure 19. The six clusters containing the most disorder-related searches

Clusters 0, 3, 10, 15, 17, and 20 account for 76% of the total searches, and contain 82% of the disorder-related searches. The next five clusters containing the most disorder-related searches are located in the bottom-right of the SOM – clusters 6, 18, 21, 23, and 25. Once those values are included, these 11 clusters account for 94% of the total searches, and 98% of the disorder-related searches. The white line in Figure 20 outlines the two mega-clusters responsible for a majority of the disorder-related searches.

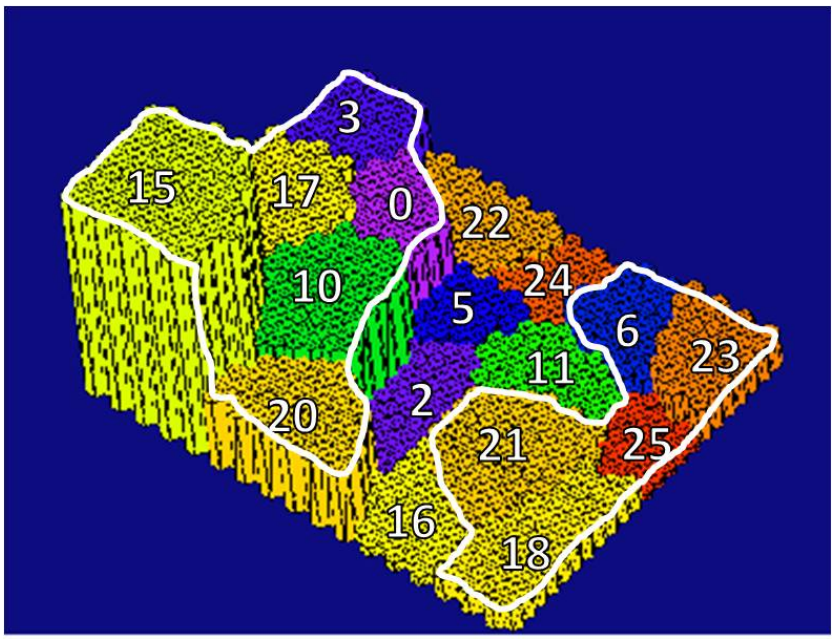

Figure 20. The 11 clusters containing the most disorder-related searches

4.10 Evaluation

The rejection of the null hypothesis in the χ^2-test statistic is a significant result – it is possible to state with greater than 95% confidence that a relationship does exist between the clustering and the disorder-related searches. This is supported further by the analysis of SOM itself, and the location of the majority of the disorder-related searches in such a small number of clusters.

V. Conclusions and Future Work

5.1 Conclusions

Extensive research has been conducted on predicting future actions based on real-world behavioral changes [15]. These indicators provide clues to the attitudes and perceived behavioral control of an individual. As society has become more dependent on the Internet for information [8], it is logical to assume that cyber-behavior can provide indicators as well – and not just for future actions in cyberspace. If Google Flu Trends can accurately correlate search queries related to flu symptoms to the number of people going to the doctor with the flu, what other cyber-indicators can be exploited?

This research determined that it is possible for software to sufficiently identify relationships between individuals searching for terms related to three different disorders – anxiety, PTSD, and suicide. The calculated χ^2-test statistic exceeded the χ^2 critical value at $\alpha = .05$, thus rejecting the null hypothesis that the clustering produced by the LDASOM was independent of the targeted search queries.

5.2 Future Work

Significant insight could be provided by working with the Air Force's Surgeon General's office to obtain the search histories of individuals who have been diagnosed with these disorders or who have committed suicide. The trends found in those search histories could greatly increase the reliability and predictability of an "Early Warning Radar" based on cyber indicators.

Furthermore, a larger data set covering more geographical locations and a longer time period would provide much more robust search histories. It would also better represent the actual population. An extended time period (e.g., 18 months) would make trend analysis a possibility.

Significant work should be done to ensure the disorder dictionaries are as mutually exclusive as possible and should be coordinated with a professional team of psychologists. This also includes concerns about false positives/negatives. False positives would include individuals searching online for information or help for a friend, family member, etc. False negatives would include users with or at-risk for the disorders, but who's search history did not include any disorder-related search queries.

Finally, the accuracy of the system could be bolstered by having a panel of psychologists review the search histories of those individuals who searched for the terms related to the disorders, and have them sorted into piles of *concerned*, *not concerned,* and *highly concerned*. After performing some inter-rater reliability tests, examine how LDASOM clustered those documents compared to the professionals.

5.3 Relevance of Work

Traditional research for these disorders, especially suicide, is conducted with a limited number of cases on data gathered after the fact and requires a substantial amount of human involvement [9]. This research lays the foundation for a system capable of predicting groups of people at a higher risk for suicide or other serious disorders by analyzing search histories – an indicator not yet examined. This is significant because this analysis can be conducted in real-time and is scalable to incredibly large populations

of users. Furthermore, the clustering produced by LDASOM identifies relationships between users who are conducting disorder-related searches and users who are not. As research progresses on this topic, it may be possible to develop a system capable of identifying high-risk groups before tragedy hits.

Appendix A. BlueCoat Proxy Categories

Abortion	Government/Legal	Phishing
Adult/Mature Content	Greeting Cards	Placeholders
Alcohol	Hacking	Political/Activists Groups
Alternative Sexuality/Lifestyle	Health	Pornography
Alternative Spirituality/Belief	Humor/Jokes	Potentially Unwanted Software
Art/Culture	Illegal Drugs	Proxy Avoidance
Auctions	Informational	Radio/Audio Streams
Audio/Video Clips	Internet Telephony	Real Estate
Blogs/Personal Pages	Intimate Apparel/Swimsuit	Reference
Brokerage/Trading	Job Search/Careers	Religion
Business/Economy	Lesbian/Gay/Bi-sexual/Transgender	Remote Access Tools
Charitable Organizations	Malicious Outbound Data/Botnets	Restaurants/Dining/ Food
Chat/Instant Messaging	Malicious Sources	Scam/Questionable/ Illegal
Child Pornography	Media Sharing	Search Engines/Portals
Computers/Internet	Military	Sex Educations
Content Servers	News/Media	Shopping
Dynamic DNS Host	Newsgroups/Forums	Social Networking
Education	Non-Viewable	Society Daily Living
Email	Nudity	Software Downloads
Entertainment	Online Meetings	Spam
Extreme	Online Storage	Sports/Recreation
Financial Services	Open/Mixed Content	Suspicious
For Kids	Pay to Surf	Tobacco
Gambling	Peer-to-Peer	Translation
Games	Personals/Dating	Travel
TV/Video Streams	Vehicles	Web Advertisements
Uncategorized	Violence/Hate/Racism	Web Applications
User Defined	Weapons	Web Hosting

Appendix B. Perl Scripts

B.1 Pre-processing - Original-to-SearchLog Filter

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

#*********************************************
# The file will filter all logs in a directory, only
# keeping those GET messages the BlueCoat Proxy
# classified as Searches.  It is run from the
# directory we're interested in.
#*********************************************

#*********************************************
# Declarations to make before running program
#*********************************************
my $base = "Dover";
my $dirname = "F:\\INOSCLogs\\${base}\\11_November";
my $writefile;

#*********************************************
# Open directory with files to be read/written
#*********************************************
opendir DIR,$dirname or die "open failed : $!\n";
for(readdir DIR) {
  if (/$base/i) {
    if (/Searches/i) {}
      else{
        open FILE, $_ or die $!;
        $writefile = substr($_,0,-4);
        open FILE2, ">" . $writefile . "_Searches" or die $!;

        #*********************************************
        # Declarations to make before running program
        #*********************************************
        my $line;  # string representing an entire line within the log

        #*********************************************
        # loop to analyze every line in file, printing out those
        # categorized as search engines/portals
        #*********************************************
        while ($line = <FILE>){
          my $w = "(.+?)";
          $line =~ m/^$w $w $w $w $w $w $w $w "$w" $w $w $w $w $w $w $w
                      $w $w $w $w $w $w $w $w $w "$w"/;

            #*********************************************
            # log fields
            #*********************************************
            my $date = $1;
```

```
            my $time = $2;
            my $timetaken = $3;
            my $c_ip = $4;
            my $cs_username = $5;
            my $cs_authgroup = $6;
            my $x_exceptionID = $7;
            my $sc_filter_result = $8;
            my $cs_categories = $9;
            my $cs_referer = $10;
            my $sc_status = $11;
            my $s_action = $12;
            my $cs_method = $13;
            my $rs_content_type = $14;
            my $cs_uri_scheme = $15;
            my $cs_host = $16;
            my $cs_uri_port = $17;
            my $cs_uri_path = $18;
            my $cs_uri_query = $19;
            my $cs_uri_extension = $20;
            my $s_ip = $21;
            my $sc_bytes = $22;
            my $cs_bytes = $23;
            my $x_virus_id = $24;
            my $r_ip = $25;
            my $cs_user_agent = $26;

            #*********************************************
            # line-by-line processing, printing searches to new file
            #*********************************************
            if ( $cs_categories =~ m/Portals/i ) {
                print FILE2 "${line}";}}

        #*********************************************
        # close files
        #*********************************************
        close(FILE2);
        close(FILE);}}}

closedir DIR or die "close failed : $!\n";
exit 0;
```

B.2 Pre-processing - Combine daily SearchLogs into monthly SearchLog

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

#*********************************************
# Combines a week's worth of searches into one file
# The file is the name of the location, with
# "_SearchLog.txt" appended to the end
#*********************************************

my $location = "";       #insert the name of the location
open FILE_Month, ">>".$location."_SearchLog.txt" or die $!;

my $dirname = "F:\\INOSCLogs\\$location\\2_DailySearchLogs";
my $line;

opendir DIR,$dirname or die "open failed : $!\n";
for(readdir DIR) {
  if (/Searches/i) {  #only open daily searchlogs
    open FILE, $_ or die $!;

    #*********************************************
    # loop to analyze every line in file, printing out those categorized
    # as search engines/portals
    #*********************************************
      while ($line = <FILE>){
        my $w = "(.+?)";
        $line =~ m/^$w $w $w $w $w $w $w $w "$w" $w $w $w $w $w $w $w $w
                   $w $w $w $w $w $w $w $w "$w"/;

        #*************************************************
        # log fields
        #*************************************************
        my $date = $1;
        my $time = $2;
        my $timetaken = $3;
        my $c_ip = $4;
        my $cs_username = $5;
        my $cs_authgroup = $6;
        my $x_exceptionID = $7;
        my $sc_filter_result = $8;
        my $cs_categories = $9;
        my $cs_referer = $10;
        my $sc_status = $11;
        my $s_action = $12;
        my $cs_method = $13;
        my $rs_content_type = $14;
        my $cs_uri_scheme = $15;
        my $cs_host = $16;
        my $cs_uri_port = $17;
        my $cs_uri_path = $18;
        my $cs_uri_query = $19;
```

```
        my $cs_uri_extension = $20;
        my $s_ip = $21;
        my $sc_bytes = $22;
        my $cs_bytes = $23;
        my $x_virus_id = $24;
        my $r_ip = $25;
        my $cs_user_agent = $26;

        my $ContentType = $rs_content_type;
        my $status = $sc_status;
        my $path = $cs_uri_path;
        my $query = $cs_uri_query;
        my $referer = $cs_referer;
        my $cIP = $c_ip;

        if ($ContentType =~ m/HTML/i && $path =~ m/search/i){
          if ( $referer =~ m/google/i ) {
            if ( $query =~ m/((&q=).*?&)/ ) {print FILE_Month $line;}}
            elsif ( $referer =~ m/bing/i ) {
              if ( $query =~ m/((\?q=).*?&)/ ) {
                if ( $referer =~ m/handlers/ ) {}
                else {print FILE_Month $line;}}}
          elsif ( $referer =~ m/ask/i ) {
            if ( $query =~ m/((\?q=).*?&)/ ) {print FILE_Month $line;}}
          elsif ( $referer =~ m/yahoo/i ) {
            if ( $query =~ m/((\?p=).*?&)/ ) {
                if ( $referer =~ m/handlers/) {}
                else {print FILE_Month $line;}}}
          elsif ( $referer =~ m/lycos/i ) {
            if ($query =~ m/((&query=).*?&)/) {print FILE_Month $line;}}
          elsif (($referer =~ m/youtube/i) && ($referer =~ m/query/i)) {
            if ( $query =~ m/(=.*?&)/ ) {print FILE_Month $line;}}
          elsif ( $referer =~ m/(forecast.weather.gov)/i ) {
            if ( $query =~ m/(=.*?&)/ ) {print FILE_Month $line;}}}}}

close(FILE);
if (/Searches/i) {print "file complete\n";}}

closedir DIR or die "close failed : $!\n";
exit 0;
```

B.3 Pre-processing – Unique IP Counter

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

#*************************************************
# This program counts the number of unique IP addresses
# for a location, and also outputs each IP address
#*************************************************

my $location = "";       #insert the name of the location
my $dirname = "F:\\INOSCLogs\\${location}\\2_DailySearchLogs";
open FILE_NumOfIPs, ">" . $location . "__TotalIPs.txt" or die $!;
open FILE_AllIPs, ">" . $location . "__AllIPs.txt" or die $!;

my %UniqueIPs = ();
my $line;

#*********************************************
# Open directory with files to be read/written
#*********************************************
opendir DIR,$dirname or die "open failed : $!\n";
for(readdir DIR) {
  if (/$location/i) {
    if (/Searches/i) {
      open FILE, $_ or die $!;

      #*********************************************
      # loop to analyze every line in the log
      #*********************************************
      while ($line = <FILE>){
        my $w = "(.+?)";
        $line =~ m/^$w $w $w $w $w $w $w $w "$w" $w $w $w $w $w $w $w $w
                   $w $w $w $w $w $w $w $w "$w"/;

        #*********************************************
        # log fields
        #*********************************************
        my $date = $1;
        my $time = $2;
        my $timetaken = $3;
        my $c_ip = $4;
        my $cs_username = $5;
        my $cs_authgroup = $6;
        my $x_exceptionID = $7;
        my $sc_filter_result = $8;
        my $cs_categories = $9;
        my $cs_referer = $10;
        my $sc_status = $11;
        my $s_action = $12;
        my $cs_method = $13;
        my $rs_content_type = $14;
        my $cs_uri_scheme = $15;
```

```
        my $cs_host = $16;
        my $cs_uri_port = $17;
        my $cs_uri_path = $18;
        my $cs_uri_query = $19;
        my $cs_uri_extension = $20;
        my $s_ip = $21;
        my $sc_bytes = $22;
        my $cs_bytes = $23;
        my $x_virus_id = $24;
        my $r_ip = $25;
        my $cs_user_agent = $26;

        my $IPAddress = $c_ip;
        my $referer = $cs_referer;
        my $ContentType = $rs_content_type;
        my $path = $cs_uri_path;
        my $query = $cs_uri_query;

        #*********************************************
        # line-by-line processing
        #*********************************************
        if ($ContentType =~ m/HTML/i && $path =~ m/search/i){
          if ( $referer =~ m/google/i ) {
            if ( $query =~ m/((&q=).*?&)/ ) {$UniqueIPs{$IPAddress}++;}}
          elsif ( $referer =~ m/bing/i ) {
            if ( $query =~ m/((\?q=).*?&)/ ) {
              if ( $referer =~ m/handlers/ ) {}
              else {$UniqueIPs{$IPAddress}++;}}}
          elsif ( $referer =~ m/ask/i ) {
            if ( $query =~ m/((\?q=).*?&)/ ) {$UniqueIPs{$IPAddress}++;}}
          elsif ( $referer =~ m/yahoo/i ) {
            if ( $query =~ m/((\?p=).*?&)/ ) {
              if ( $referer =~ m/handlers/) {}
              else {$UniqueIPs{$IPAddress}++;}}}
          elsif ( $referer =~ m/lycos/i ) {
            if ($query =~ m/((&query=).*?&)/){$UniqueIPs{$IPAddress}++;}}
          elsif (($referer =~ m/youtube/i) && ($referer =~ m/query/i)){
            if ( $query =~ m/(=.*?&)/ ) {$UniqueIPs{$IPAddress}++;}}
          elsif ( $referer =~ m/(forecast.weather.gov)/i ) {
            if ( $query =~ m/(=.*?&)/ ) {$UniqueIPs{$IPAddress}++;}}}}}
  close(FILE);
  print "file complete\n";}}

my $NumUniqueIPs = keys %UniqueIPs;
print FILE_NumOfIPs "${location} had ${NumUniqueIPs} Unique IPs from 7
                     Nov 11 through 7 Dec 11.";
close(FILE_NumOfIPs);

foreach $NumUniqueIPs (keys %UniqueIPs){print FILE_AllIPs
"$NumUniqueIPs\n";}
exit 0;
```

B.4 Pre-processing - IP Filter

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

#***********************************************
# The program goes the the month-long SearchLog, creates a new file for
# each IP address, and saves the search queries originating from that
# IP to the file, in chronological order
#***********************************************

#***********************************************
# Declarations to make before running program
#***********************************************
my $base = "Charleston";
my $dirname = "F:\\INOSCLogs\\$base\\2_DailySearchLogs";
open FILE, $base."_SearchLog.txt" or die $!;
my $line;
my $SearchInput;
my $cIP;
my $LineCount = 0;

my @AllIPs = ();
open FILE_AllIPs, $base."__AllIPs.txt" or die $!;
  while ($line = <FILE_AllIPs>) {
    chomp $line;
    push(@AllIPs,$line);}
 close(FILE_AllIPs);

#***********************************************
# loop to analyze every line in file, printing out those categorized as
search engines/portals
#***********************************************
while ($line = <FILE>){
  $LineCount++;
  print "$LineCount\n";

  my $w = "(.+?)";
  $line =~ m/^$w $w $w $w $w $w $w $w "$w" $w $w $w $w $w $w $w $w $w
              $w $w $w $w $w $w $w "$w"/;

  #***********************************************
  # log fields
  #***********************************************
  my $date = $1;
  my $time = $2;
  my $timetaken = $3;
  my $c_ip = $4;
  my $cs_username = $5;
  my $cs_authgroup = $6;
  my $x_exceptionID = $7;
  my $sc_filter_result = $8;
  my $cs_categories = $9;
```

```
  my $cs_referer = $10;
  my $sc_status = $11;
  my $s_action = $12;
  my $cs_method = $13;
  my $rs_content_type = $14;
  my $cs_uri_scheme = $15;
  my $cs_host = $16;
  my $cs_uri_port = $17;
  my $cs_uri_path = $18;
  my $cs_uri_query = $19;
  my $cs_uri_extension = $20;
  my $s_ip = $21;
  my $sc_bytes = $22;
  my $cs_bytes = $23;
  my $x_virus_id = $24;
  my $r_ip = $25;
  my $cs_user_agent = $26;

  my $query = $cs_uri_query;
  my $referer = $cs_referer;
  $cIP = $c_ip;

  if ( $referer =~ m/google/i ) {
    if ( $query =~ m/((&q=).*?&)/ ) {
       $SearchInput = substr($1,3);
       &PrintSearchQuery;}}
    elsif ( $referer =~ m/bing/i ) {
       if ( $query =~ m/((\?q=).*?&)/ ) {
           if ( $referer =~ m/handlers/ ) {}
           else {
               $SearchInput = substr($1,3);
               &PrintSearchQuery;}}}
    elsif ( $referer =~ m/ask/i ) {
       if ( $query =~ m/((\?q=).*?&)/ ) {
           $SearchInput = substr($1,3);
           &PrintSearchQuery;}}
    elsif ( $referer =~ m/yahoo/i ) {
       if ( $query =~ m/((\?p=).*?&)/ ) {
           if ( $referer =~ m/handlers/) {}
           else {
               $SearchInput = substr($1,3);
               &PrintSearchQuery;}}}
    elsif ( $referer =~ m/lycos/i ) {
       if ( $query =~ m/((&query=).*?&)/ ) {
           $SearchInput = substr($1,3);
           &PrintSearchQuery;}}
    elsif ( ($referer =~ m/youtube/i) &&  ($referer =~ m/query/i) ) {
       if ( $query =~ m/(=.*?&)/ ) {
           $SearchInput = substr($1,1);
           &PrintSearchQuery;}}
    elsif ( $referer =~ m/(forecast.weather.gov)/i ) {
       if ( $query =~ m/(=.*?&)/ ) {
           $SearchInput = substr($1,1);
           &PrintSearchQuery;}}}
close(FILE);
```

```
exit 0;

sub PrintSearchQuery {
    #$searchcount++;
    $SearchInput = substr($SearchInput,0,-1);  # remove the last char
    $SearchInput = lc($SearchInput);           # convert to lowercase
    $SearchInput =~ tr/+/ /;                   # convert + to whitespace
    $SearchInput =~ tr/-/ /;                   # convert - to whitespace
    $SearchInput =~ s/%([a-f0-9][a-f0-9])/chr(hex($1))/ieg; # convert
                                               #hex-ascii to whitespace
    $SearchInput =~ s/[[:punct:]]//g;          # remove any punctuation
    $SearchInput =~ s/^\s+|\s+$//g;            # remove any extra spaces

    if ($SearchInput =~ m/€/i) {}
    else {
        open my $cIP, ">>C_" . $cIP . "_IPFilter.txt" or die $!;
        print $cIP "$SearchInput\n";
        close($cIP);}}
```

B.5 Disorder-Related Search Histories

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

#*********************************************
# Declarations to make before running program
#*********************************************
my $line;
my $TotalAnxietyHits = 0;
my $TotalPTSDHits = 0;
my $TotalSuicideHits = 0;
my $TotalInterestingHits = 0;
my $FileCount = 0;
my $dirname = "C:\\Users\\cmiller\\Desktop\\IPFileScrub";
open FILE_BucketSort, ">BucketSort_Hits.txt" or die $!;
open FILE_AnxietyIPs, ">AnxietyIPs.txt" or die $!;
open FILE_PTSDIPs, ">PTSDIPs.txt" or die $!;
open FILE_SuicideIPs, ">SuicideIPs.txt" or die $!;
open FILE_AllIPs, ">AllIPs.txt" or die $!;

#*********************************************
# move words from dictionaries to array for comparison
#*********************************************
open FILE_AnxietyWords, "AnxietyWords.txt" or die $!;
my @AnxietyWords = ();
while ($line = <FILE_AnxietyWords>) {
  chomp $line;
  push(@AnxietyWords,$line); }
close(FILE_AnxietyWords);

open FILE_PTSDWords, "PTSDWords.txt" or die $!;
my @PTSDWords = ();
while ($line = <FILE_PTSDWords>) {
  chomp $line;
  push(@PTSDWords,$line); }
close(FILE_PTSDWords);

open FILE_SuicideWords, "SuicideWords.txt" or die $!;
my @SuicideWords = ();
while ($line = <FILE_SuicideWords>) {
  chomp $line;
  push(@SuicideWords,$line); }
close(FILE_SuicideWords);

opendir DIR,$dirname or die "open failed : $!\n";
for(readdir DIR) {
  if (/IPFilter/i) {
    open FILE, $_ or die $!;
    my $FileName = $_;
    my $InterestingHits = 0;
    my $AllClusterIPs = 0;
    my $Word;
```

```
    my @Matches = ();
    my @Dictionary = ();

    print FILE_AllIPs "$FileName\n";

    my %InterestingWordTotals = (Anxiety=>0, PTSD=>0, Suicide=>0);
    my $sum = 0;

    while ($line = <FILE>){
      chomp $line;
      foreach $Word (@AnxietyWords){if ($line =~ m/^$Word/i)
           {push(@Matches,$line); push(@Dictionary,$Word);
           $InterestingWordTotals{'Anxiety'}++; $InterestingHits++;}}
      foreach $Word (@PTSDWords){if ($line =~ m/^$Word/i)
           {push(@Matches,$line); push(@Dictionary,$Word);
           $InterestingWordTotals{'PTSD'}++; $InterestingHits++;}}
      foreach $Word (@SuicideWords){if ($line =~ m/^$Word/i)
           {push(@Matches,$line); push(@Dictionary,$Word);
           $InterestingWordTotals{'Suicide'}++; $InterestingHits++;}}}

      if ( $InterestingHits != 0) {
       my $highest_val = (sort { $InterestingWordTotals{$b} <=>
           $InterestingWordTotals{$a} } keys %InterestingWordTotals)[0];
       print FILE_BucketSort "$FileName,$highest_val";
       foreach(@Matches){
         my $Match = pop(@Dictionary);
         print FILE_BucketSort ",($Match)$_";}
         print FILE_BucketSort "\n";
         foreach my $key (keys %InterestingWordTotals) {$sum +=
                                          $InterestingWordTotals{$key};}

       if ($highest_val eq "Anxiety") {print FILE_AnxietyIPs
                              "$FileName,$sum\n"; $TotalAnxietyHits++;}
       elsif ($highest_val eq "PTSD") {print FILE_PTSDIPs
                                 "$FileName,$sum\n"; $TotalPTSDHits++;}
       elsif ($highest_val eq "Suicide") {print FILE_SuicideIPs
                             "$FileName,$sum\n"; $TotalSuicideHits++;}}

    $FileCount++;
    print "$FileCount\n";
    close(FILE);}}

close(FILE_AnxietyIPs);
close(FILE_PTSDIPs);
close(FILE_SuicideIPs);
close(FILE_AllIPs);

print FILE_BucketSort "Anxiety: $TotalAnxietyHits, PTSD:
$TotalPTSDHits, Suicide: $TotalSuicideHits";
close(FILE_BucketSort);
exit 0;

sub largest_value (\%) {
  my $hash = shift;
  keys %$hash;       # reset the each iterator
```

```
my ($large_key, $large_val) = each %$hash;
while (my ($key, $val) = each %$hash) {
  if ($val > $large_val) {
    $large_val = $val;
    $large_key = $key;}}
$large_key;}
```

B.6 Cluster Statistics

```
#!/usr/bin/perl
use 5.010;
use strict;
use warnings;

my $dirname = "C:\\Users\\cmiller\\Desktop\\IPFileScrub";
open FILE_ClusterStats, ">ClusterStats.txt" or die $!;
print FILE_ClusterStats "ClusterID,# of Docs,# of
                                  Searches,SC,MD,ND,AR,FL,KS,NJ,IL,CA\n";

#*********************************************
# Declarations to make before running program
#*********************************************
my $line;
my $Base;
my $ClusterID;
my %ClusterDocs = ();
my %Searches = ();

opendir DIR,$dirname or die "open failed : $!\n";
for(readdir DIR) {
  if (/ClusterID/i) {
    open FILE_ClusterID, $_ or die $!;
    $ClusterID = $_;
    my %BaseDocCounter = ();
    while ($line = <FILE_ClusterID>){
      %ClusterDocs{$ClusterID}++;
      open FILE_IP, $line or die $!;
        $Base = substr($_,0,2);
        $BaseDocCounter{$Base}++;
        while ($line = <FILE_IP>){$Searches{$ClusterID}++;}
        close(FILE_IP);
        print FILE_ClusterStats
            "$ClusterID,$ClusterDocs{$ClusterID},$Searches{$ClusterID},
            $BaseDocCounter{'C_'},$BaseDocCounter{'D_'},
            $BaseDocCounter{'G_'},$BaseDocCounter{'L_'},
            $BaseDocCounter{'MD'},$BaseDocCounter{'MC'},
            $BaseDocCounter{'MG'},$BaseDocCounter{'S_'},
            $BaseDocCounter{'T__'}\n";}}
  close(FILE_ClusterID);}

close(FILE_ClusterStats);
exit 0;
```

Appendix C. LDASOM Topics

Topic 0: 45 western airmen math animals route bin groups definition run
Topic 1: did death wikipedia cartoon guy marathon results say watches characters
Topic 2: people data magazine joseph vista solar 06 purple iii fine
Topic 3: date cover release gmc earth save earthquake function julian sierra
Topic 4: force air adls enlisted learn learning advanced distributed training cbt
Topic 5: park child divorce support laws coloring masks suicide knives comic
Topic 6: dog model gold dogs bikes shops reason border bruce breeders
Topic 7: afi 91 31 33 afosh chapter 501 201 usc volume
Topic 8: use application single download remove create waste 64 poem bit
Topic 9: dental medal robert awards study ribbon ribbons decorations challenge dentist
Topic 10: fire kindle ups amazon 35 spring cities kate costume smart
Topic 11: civilian age square budget value cuts flow suit head right
Topic 12: running nike diagram shoe directions ham mid spanish honey 97
Topic 13: template lowes need photography mr cleaning patrick carpet cakes wave
Topic 14: hand tsp oregon pharmacy columbus overseas hiring specialist freeze eyes
Topic 15: king bed plans wife 2001 wright bath burger sister leg
Topic 16: war market watch inspection electronic memorial stadium austin meat abbreviations
Topic 17: current difference acronym contact stone thesaurus reports finder catholic consumer
Topic 18: red hair msn salon tracker styles natural cup instructions simmons
Topic 19: city oklahoma wrestling atlantic baseball ocean castle reverse 89 miller
Topic 20: ice smith bases williams bob hockey ok ticketmaster england amy
Topic 21: credit federal union andrews cooper pentagon chef freedom pampered earned
Topic 22: calendar santa holidays rose candy julian boot nuclear claus dec
Topic 23: pc bell asus farm journal evo impact vet helicopter interview
Topic 24: tax rate numbers chinese plane yellow lower calculate taxes clock
Topic 25: kit 2008 social lift kits riddle volkswagen embry drop cc
Topic 26: university college community phoenix ashford colleges athletics kaplan transcripts campus
Topic 27: phone number 20 cell point 00 phones wi wisconsin slim
Topic 28: fish points colonel communications shift separation recovery channel gallon aquarium
Topic 29: sale 2004 1000 pistol boats optima ducati cva 308 dealers
Topic 30: 11 101 24 23 13 boat afman vol 110 21
Topic 31: travel government citi camp writing apa az 5k essay kentucky
Topic 32: va john contract disability load rating deere ratings compensation benefits
Topic 33: music building tool songs brothers convert hit pine fm murder
Topic 34: training pressure emergency signs faa heat symptoms brain bite injury
Topic 35: list gun gas promotion annual sets guns tea survey craigs
Topic 36: benefits martin pregnancy pregnant early write bird allen ebis weeks
Topic 37: pictures excel picture dates research sheets soup employment cells add
Topic 38: nj burlington hamilton wrightstown browns mills pemberton lumberton bordentown medford
Topic 39: form dd steve 1351 hudson blank 2875 214 worksheet fillable
Topic 40: review reviews pad adams leap leappad emblem roland lg optima
Topic 41: medical hospital childrens recruiter regional peter division religion choice 130j
Topic 42: access company furniture run far warrior foundation franklin pounds stores
Topic 43: hill secret cherry 2011 living ma victoria shower oakland massage
Topic 44: recipe recipes cake pie homemade chocolate pumpkin chili cookie salad
Topic 45: shoes team non great large lottery shows metro crab close
Topic 46: grand forks nd mn cola minnesota coca dakota hockey und
Topic 47: dodge ram trucks 1500 charger rims chevy cab mirror truck
Topic 48: best buy iphone 4s cases otterbox older jailbreak 3gs shrine
Topic 49: af form imt 988 afto 910 931 geo prizm 171
Topic 50: classic german info signature royal puppies month collection wildlife british
Topic 51: new jersey egypt mikes arab rep biodiesel metlife brent transit

Topic 52: 4 f left 150 49ers standard issue corvette clean locker
Topic 53: 8 station valley jack 0 el jordan inside lodge puppy
Topic 54: control process orm analysis step risk hazard making event authority
Topic 55: mount mt human holly pleasant hunter laurel drum rice rights
Topic 56: charleston sc summerville dorchester ladson rivers sceg 843 tanger lowcountry
Topic 57: mcguire afb ny nj dix nyc lakehurst trenton philly jbmdl
Topic 58: japan symbol certificate install diesel account print f250 birth 250
Topic 59: dress girl surgery blues regulations lane concert kohls drill christina
Topic 60: two dictionary clep da pass crossfit planet sgt dining contracting
Topic 61: safety number photos hd topics routing briefing extended representative adults
Topic 62: quotes business happy quote language drake nsips small smoke album
Topic 63: code zip codes coupon area promo gulfport zipcode promotional cpt
Topic 64: best airforce bad career famous lee temperature way daniel check
Topic 65: mypay 3 modern warfare american wii ucmj african wire 86
Topic 66: hp driver golden printer password connect reset ip corral monitor
Topic 67: lyrics amazon mac michelle want jesus highest hate congress cast
Topic 68: 2009 block southern manager columbia cap fit gray pipe hr
Topic 69: online programs liberty masters courses anderson degrees science graduate banking
Topic 70: address bar way angel lookup italian 26 stories mailing 47
Topic 71: west epr sales bullets key bullet ashley yard volunteer wwe
Topic 72: james report mil 130 military sandusky jerry member orders spouse
Topic 73: wedding year country dresses kim kardashian kay jewelers bridal engagement
Topic 74: game general face ms major bear bass christopher teddy empire
Topic 75: health care clinic wa med alarm healthcare theatre pediatric swank
Topic 76: south carolina creek sc charleston fishing goose clemson palmetto bb
Topic 77: vs brown ufc cdc dark fight night wayne pacquiao mayweather
Topic 78: nissan tablet android app image apps saw spray airmans vinyl
Topic 79: family group washington academy dc readiness boxes esd leader francis
Topic 80: il scott illinois belleville ofallon mascoutah shiloh clair 375 62225
Topic 81: women baby short girls woman mma hairstyles poems cars haircuts
Topic 82: low location 500 tools race mississippi chest multi shock vacuum
Topic 83: 2003 maintenance start directory room pack upgrade fusion hardware hidden
Topic 84: 2007 word email outlook sharepoint text files copy powerpoint message
Topic 85: jeep class 50 napa cherokee edition dot woods suspension tj
Topic 86: real estate technical professional giants trident cmsgt past stocks java
Topic 87: pro replacement file birthday jacket 3d kia jackets indoor macbook
Topic 88: big womens dawn breaking moore twilight wrangler guam forever mods
Topic 89: login super ii cbs generator fantasy vacation control pistol units
Topic 90: toyota rates 2012 bah runner diem tacoma military corolla crime
Topic 91: hunting paul deer land w george tn ron 29 jim
Topic 92: license k development drivers story ct marriage richard conference testing
Topic 93: operations operation requirements mission commander planning response approved commanders freedom
Topic 94: officer 2000 problems 2002 rank mass terminal issues mileage shot
Topic 95: dr life ako christian webmail bus faith greyhound dre cycle
Topic 96: county cadillac solano hillsborough berkeley jail sheriff arrest cts clerk
Topic 97: guard unit award job william description air outstanding meritorious matt
Topic 98: master chief combat air chase sergeant h badge force crew
Topic 99: table ball silverado jet z plastic dragon laser frame tables
Topic 100: facebook commercial mart wal walmart aolcom aol musicians trading ships
Topic 101: car rental cars rc lincoln ugly town dealerships sweaters enterprise
Topic 102: weather basketball band cold channel forecast dream bands extreme underground
Topic 103: stock harley alaska 16 davidson dealers motorcycles c5 peterson xmas
Topic 104: coupons camera digital outlet wine practice bottle plug tests elements

Topic 105: house times express rentals arts seating trip polar martial seven
Topic 106: pa marine dvd area player corps tennessee pennsylvania newspaper models
Topic 107: course door glass leadership sweet kitchen five pot craftsman doors
Topic 108: help boston exercise statistics mother arrested warner military causes robins
Topic 109: 2011 paint dallas december ps3 cowboys bread fall players bodybuilding
Topic 110: club golf boots dodgers sams alcohol duck raiders clubs kings
Topic 111: news fox daily msnbc abc democrat bnd bbc megan cnn
Topic 112: 2 ipad action supplement 57 1993 figures atrix takes sales
Topic 113: food level gsa gym advantage protection 400 nose lexus multicam
Topic 114: american mustang cisco 2013 gt arizona vw native gti tallahassee
Topic 115: squadron wing nc airlift sex fighter air expeditionary kid refueling
Topic 116: line la crossword answer pocket queen employees packages holder sunglasses
Topic 117: leave 30 ed codes ultimate fleece andrew ready burning mg
Topic 118: turkey box party shadow cook deep fryer fried monkey incirlik
Topic 119: bank america j td weathercom walmart locations beneficial wives citizens
Topic 120: security forces icao place weapons armed homeland authorized raven afcent
Topic 121: 17 river pilot toms taylor armor 41 ultra lease 42
Topic 122: men images photo clothing shirts canon soul beer msncom drinking
Topic 123: vehicle electric airfield chain edge registration walk 1995 brian rules
Topic 124: delaware repair pics usb device knee wilmington soft winner hdmi
Topic 125: football b gear landing 58 bags bridge coach championship 52
Topic 126: beach disney things resort ski houston resorts stay recreation myrtle
Topic 127: amu housing dmhrsi heart plate dhmrsi website indiana bush luke
Topic 128: sign colorado springs shooting chris van lisa rescue auction cape
Topic 129: uniform 32 window ave cda battle dencom syndrome tablets debt
Topic 130: program equipment supply record ray logistics standards aerospace hq activity
Topic 131: radio afsc lake street tahoe listing occupy turtle systems shack
Topic 132: ford island f150 steel ranger explorer raptor focus parker coil
Topic 133: book airlines southwest delta regulation justin gay spirit bieber continental
Topic 134: afb lodging checklist ridge lackland sheppard maxwell langley billeting eglin
Topic 135: 2010 exchange small nursing crash leather rick shore grinch ross
Topic 136: following best hazardous oxygen procedures required associated materials responsible authorized
Topic 137: day navy veterans yahoocom parade nko veteran discounts serco advancement
Topic 138: afi 36 coffee 2903 lakes maker keurig 3003 makers regulations
Topic 139: guide amc tire tires 18 firestone johns projector papa mats
Topic 140: wheels wheel camaro racing pontiac steering 95 trans rear 98
Topic 141: free art clip printable clipart perry animated katy gluten sounds
Topic 142: type types following material person individual soldier apply burn snake
Topic 143: tv board website inch jewelry lcd korea ats osan c130
Topic 144: e pubs 14 forms publishing lsu meter elf hobby reserves
Topic 145: center network education mask dish convention massachusetts conference 68 m50
Topic 146: direct forum aol mexico tom core tiger puerto forums rico
Topic 147: chevy homes tricare pain floor ss prime frank saic impala
Topic 148: tampa macdill bay orlando pete tough mudder brighthouse petersburg teco
Topic 149: fuel pump sensor valve head instructions 1996 infiniti snowboard knock
Topic 150: mean movies workout filter turn disease posters coming demotivational future
Topic 151: usaa retirement dmv locations stars hotmailcom stripes chase contacts dancing
Topic 152: microsoft windows remote multiple code adobe vulnerability execution vulnerabilities update
Topic 153: black friday ads deals ad sales shopping hardaway mos capsule
Topic 154: money battery chicken calories roster translate meals soup starter quick
Topic 155: 6 target support sears switch wars serial panasonic speakers washer
Topic 156: nfl fitness week years chicago scores picks bears draft eve
Topic 157: rent 60 houses los agency activities intelligence angeles buy townhomes

Topic 158: air force bonus reenlistment enlistment bonuses employer arcnet bastion mahi
Topic 159: white 2005 max therapy audi pages vw porsche sarah beetle
Topic 160: 2011 november pdg miap increments yu calvin e7 e5 oct
Topic 161: military central eye basic working covers institute allowance eagles doctor
Topic 162: 7 power windows read xp win cord smsgt iso warriors
Topic 163: 1 m kelly rear disa sights underwood carbine 1c m1
Topic 164: card star citibank hotmail memory payment mattress nm cac login
Topic 165: near hotels 100 places eat hunt mercury eastern cedar douglas
Topic 166: defense nsn heights 01 hall fairview act sleep osha soldiers
Topic 167: service civil customer members usps teaching positions postal kyle taxi
Topic 168: samsung come reader charter father solutions zombie ky highland airplane
Topic 169: base air vegas las aviano nellis nv kandahar thumrait champs
Topic 170: new 12 york cnn uniforms nicole subway uso manhattan nyc
Topic 171: home kids depot funeral dfas know fix desk instructor obituaries
Topic 172: pay federal scale gs 2012 cfr period naf employees opm
Topic 173: thanksgiving panel dinner section year utility bathroom rule re hyatt
Topic 174: discount diamond aviation denver hood better plaza industrial broncos exam
Topic 175: time zone converter right abbreviation outdoor jimmy zulu storm guys
Topic 176: command georgia mobility books enterprise telephone howard gateway electrical birds
Topic 177: bowl mountain tour bike rankings god rings lines hour bcs
Topic 178: open rifle remington safe ammo cabelas rifles shotgun arms guns
Topic 179: 3 manual mens series construction 05 battlefield fb owners historical
Topic 180: color lock instruction nook pin russian changes gloves magic clear
Topic 181: air force cissp garner rocky someones hershey hutchinson 433rd 1104
Topic 182: youtube water heater sweater row third saints smoking tobacco neck
Topic 183: joint public mobile 22 aafes affairs flying youtubecom virgin jeans
Topic 184: games funny sports paper notams fun monster dins jokes adds
Topic 185: schedule dod cancer league medicine breast milk 365 solid roller
Topic 186: letter air force policy sample traffic letters counseling failure recommendation
Topic 187: price internet ruger flat springfield glock carry 1911 sig taurus
Topic 188: green cheap flights packers hat sage bean second manning aaron
Topic 189: david thomas col lt cheese jb female mdl factor male
Topic 190: craigslist walmart tundra batman craiglist minute gander peoples reed boards
Topic 191: ideas gift 19 worth got plant gmail std basket anniversary
Topic 192: world cut end mapquest sea biggest canada largest dave height
Topic 193: court msds ground portable 99 lawn tmobile salt mower trial
Topic 194: 2012 chart pay hyundai payroll sonata 703 elantra rmd e6
Topic 195: order change build self 25 controller wiring patterns really concrete
Topic 196: trailer certification nba lab trade inches bumper trailers rumors ceremony
Topic 197: skyrim server error sql oracle user database file configuration banner
Topic 198: calculator webflis loan mortgage score student waps cream pants ernie
Topic 199: storage mini led tank ring comcast lighting supplies italy 55
Topic 200: ca vacaville fairfield sacramento vallejo solano suisun kaiser walnut roseville
Topic 201: driving screen speed tips snow vehicles virtual chiefs lead operator
Topic 202: google maps truck saint leo blackboard suzuki decals decal bing
Topic 203: records personnel title request vision 38 management handbook foia reddit
Topic 204: hot al port virginia naval mary stewart township salem aerial
Topic 205: texas eagle biography clark bow lewis indian prayer auburn coins
Topic 206: specs mazda turbo brake soccer mile carbon 600 fiber talking
Topic 207: 2006 cat fargo alabama wells conversion transportation shipping dealer arctic
Topic 208: work lights wood stand display pattern wooden rod hope natalie
Topic 209: first epubs mypay dual tag aid 1st replace buying trust
Topic 210: cost average half total bull price month costs year material
Topic 211: church pdf property afpc personal format secure mls husband spain

Topic 212: long custom pink abu gauge charts sms sleeve problem summary
Topic 213: weight fat loss effects diet standard nutrition lose nfpa protein
Topic 214: menu restaurant engineering restaurants jones wings seat wild buffalo japanese
Topic 215: old motorcycle song mw3 dance fly republic anti film singer
Topic 216: man d dead heavy walking counter away iron pepper ghost
Topic 217: services sprint tracking local boys cool fedex package mitsubishi bedroom
Topic 218: movie holiday inn theater suites hilton attack marriott grove hampton
Topic 219: dts germany shirt clearance foreign ramstein iraq pittsburgh nascar currency
Topic 220: video boy questions urban china girlfriend fireplace anthony oneal korean
Topic 221: live special ga view atlanta ear avenue savannah madison stove
Topic 222: r michael toys l finance prices arows wesley sd babies
Topic 223: fl tampa brandon florida dale riverview clearwater mabry usf petersburg
Topic 224: 9 shop p 40 universal cast bars 55 210 creed
Topic 225: florida schools jackson miami patch 44 dies lightning ultrasound marlins
Topic 226: state ohio penn joe coach sport paterno rodeo scandal oem
Topic 227: days events just rack moving meaning adapter break grant roof
Topic 228: area motor metal dsn distance seal del administration recall 93
Topic 229: aircraft light feet parking 37 tower runway cargo hold instrument
Topic 230: san wall francisco antonio sf diego knowledge street knife tx
Topic 231: usaf tattoo design tattoos sony speech ssgt designs tsgt msgt
Topic 232: apartments tx look village technology apartment dan position manas oaks
Topic 233: north dakota jr harry 1997 carrier dame notre eddie penny
Topic 234: information sheet systems act privacy purchase critical item opsec process
Topic 235: air force law log falls memorandum niagara ti enforcement cabin
Topic 236: love page bible pet mall horse cd pull strategy flooring
Topic 237: space statement true statements project three identify permit confined entry
Topic 238: test blood press sugar unemployment drug cain herman vietnam french
Topic 239: definition web link site go81 htc youth g081 flis shark
Topic 240: financial patient treatment infection effect grow radiation following feeling incident
Topic 241: map grill quest arnold items arena wingman ymca household toddlers
Topic 242: wallpaper winter davis desktop garage background backgrounds tile noise registry
Topic 243: drive att v hard blackberry 1998 external trane price share
Topic 244: philadelphia plan cruise facts flowers savings flower tony delivery ranks
Topic 245: case guitar nov style oak mexican judge studio les weekend
Topic 246: christmas tree gifts trees wolf reindeer stocking decorating stockings ornament
Topic 247: tube reserve cyber monday deals ipod touch generation laptops nano
Topic 248: fort set ft michigan 200 sam sauce dix grace piece
Topic 249: body gen nikon bio scope works wide maj gingrich acer
Topic 250: insurance pt mark names companies technician charlie bowling cafe 90
Topic 251: u chevrolet 2011 g assistant muscle trader quality autotrader lion
Topic 252: oil salary transmission best subaru wrx fluid change sti miles
Topic 253: rock little ar arkansas pulaski littlerock 19th baptist 72076 miley
Topic 254: af portal epubs afportal myafmil mentor csc tcu porta uopx
Topic 255: blue bmw field energy gto shield bluetooth source headset gallery
Topic 256: cards ab 135 belt charles flash kc track tail iowa
Topic 257: logo define symptoms agent tongue quill chemical skin culture agents
Topic 258: verizon wireless apple association laptop deal router fios refurbished 82
Topic 259: california national staff silver northern chuck norris detroit lions usmc
Topic 260: united states co employee assistance fisher foxnews pacific tuition communities
Topic 261: air force reflective belts sketch graduates osteoporosis resiliency 2108 psm
Topic 262: good pizza season play o foot hut evaluation flyer episode
Topic 263: ar arkansas jacksonville cabot sherwood asu lrafb beebe conway branson
Topic 264: office afghanistan fss library forest patriots y seattle fresh married
Topic 265: road duty active deployment pre elite marines stainless utah voice

Topic 266: espn report twitter sports drudge nation jason recruiting tomahawk 247
Topic 267: cable harbor status pearl passport toy rv pcs jk cowboy
Topic 268: 10 words kc hawaii usajobs friends hickam cheat headphones gomez
Topic 269: parts auto performance exhaust advance 350 gm autozone aftermarket salvage
Topic 270: global audio legal today double sun amp mom installation 350z
Topic 271: dover de delaware mortuary camden wilmington milford smyrna wboc 19901
Topic 272: wichita ks kansas mcconnell derby cox attnet andover butler intrust
Topic 273: flight dell galaxy johnson droid motorola ticket dollar elizabeth 4g
Topic 274: afb travis keesler bx bookoo beatty balfour commissary bldg beale
Topic 275: tech spa tim wind sound wear jennifer chrysler sons ang
Topic 276: yahoo mail tickets airline pool century cheap inbox tspgov horoscope
Topic 277: air force history east coast trail goods sporting heritage assignment
Topic 278: roll search barrel english sedgwick 03 crystal translator rodney arthur
Topic 279: store stores machine rd garden facebookcom lauren talk circle mill
Topic 280: department police acura night justice uss criminal poster mike harris
Topic 281: army examples example resume 21 templates reading nco salvation mla
Topic 282: airport international md flag maryland baltimore paintball flags ravens realtor
Topic 283: train main steam accident engineer lackawanna erie locomotive 51 branch
Topic 284: military loans pioneer discounts spouses consolidation feel lending payday wilsons
Topic 285: st louis mo missouri stl cardinals rams vuitton stlouis metrolink
Topic 286: software capital article balance cash ways coleman feedback zero statement
Topic 287: xbox 360 animal cycle shelter plus update skyrim abuse gamestop
Topic 288: management degree society eating disorder disorders guidelines treatment nurse assessment
Topic 289: 5 bag range check formula percent tint available extension randolph
Topic 290: 15 hours tactical toshiba ccaf powder upper satellite complete sight
Topic 291: jobs usa opm 71 1999 job cop careers employment sf
Topic 292: school high district middle elementary teacher jrotc ballot leslie dorchester
Topic 293: post gi transfer 911 lady courier jessica 07 kevin montgomery
Topic 294: rate fed term dollars money demand graph economy point percentage
Topic 295: wiki answers obama president bomb hip resources moon hop clinton
Topic 296: cross training 300 70 physical classes lost mc baptist weekly
Topic 297: hotel size airman senior smart dui going grip macys rules
Topic 298: children young barnes products ge piano cause exercises noble adult
Topic 299: honda engine civic yamaha accord steelers engines cam heads jetta

Works Cited

[1] J. Battelle, The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture, New York: Penguin, 2005.

[2] J. Ginsberg, M. Mahebbi, R. Patel, L. Brammer, M. Smolinski and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature,* vol. 457, pp. 1012-1014, 2009.

[3] G. Koehler-Derrick and J. Goldstein, "Using google insights to assess Egypt's Jasmine Revolution," *CTC Sentinel,* pp. 4-8, March 2011.

[4] Department of Defense Task Force on the Prevention of Suicide by Members of the Armed Forces, "The challenge and the promise: Strengthening the force, preventing suicide, and saving lives," Department of Defense Suicide Event Report, 2010.

[5] LtGen Darrell D. Jones, Deputy Chief of Staff Manpower, Personnell and Services, United States Air Force, *Hearing to examine current status of suicide prevention programs in the Air Force,* Washington D.C.: Department of the Air Force presentation to the Subcommittee on Military Personnel, Committee on Armed Service, United States House of Representatives, 2011.

[6] Air Force Medical Service, "Suicidal Behaviors," Air Force Suicide Prevention Program, 2012. [Online]. Available: http://airforcemedicine.afms.mil/idc/groups/public/documents/webcontent/knowledgejunction.hcst?functionalarea=LeadersGuideDistress&doctype=subpage&docname=CTB_204436. [Accessed 7 May 2012].

[7] Need to find, "Spike in suicides," *Air Force Times,* pp. XX-XX, 7 May 2012.

[8] N. Carr, The Shallows: What the Internet is doing to our brains, New York: W.W. Norton & Company, Inc., 2011.

[9] D. Cox, M. Ghahramanlou-Holloway, F. Greene, J. Bakalar, C. Schendel and M. Nademin, "Suicide in the United States Air Force: Risk factors communicated,"

Journal of Affective Disorders, vol. 133, pp. 398-405, 2011.

[10] R. Bucklin and C. Sismeiro, "Click here for internet insight: Advances in clickstream data analysis in marketing," *Journal of Interactive Marketing,* vol. 23, pp. 35-48, 2009.

[11] B. Mobasher, "Web Usage Mining," in *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*, New York, Singer, 2007, pp. 449-483.

[12] I. Witten, E. Frank and M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Burlington: Morgan Kaufmann, 2011.

[13] B. Liu, Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, New York: Springer, 2007.

[14] A. Yang, S. Tsai, N. Huang and C. Peng, "Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004-2009," *Journal of Affective Disorders*, vol. 132, pp. 179-184, 2011.

[15] I. Ajzen, Attitudes, personality, and behavior (2nd. Edition), Milton-Keynes, England: McGraw-Hill, 2005.

[16] L. Bartholomew, G. Parcel, G. Kok, N. Gottlieb and M. Fernandez, Planning health promotion programs: an intervention mapping approach, San Francisco: Jossey-Bass, 2011.

[17] M. Stead, S. Tagg, A. MacKintosh and D. Eadie, "Development and evaluation of a mass media Theory of Planned Behaviour intervention to reduce speeding," *Health Education Research,* vol. 20, no. 1, pp. 36-50, 2005.

[18] S. Taylor and P. Todd, "Decomposition and crossover effects in the theory of planned behavior: a study of consumer adoption intentions," *International Journal of Research in Marketingq,* vol. 12, no. 2, pp. 137-155, 1995.

[19] D. Robinson, "Cyber-based behavioral modeling," PhD Thesis, Dartmouth College (Thayer School of Engineering), July 2010.

[20] A. Lipsman, "Why is Cyber Monday Becoming More Important to Holiday Season E-Commerce?," comScore, Inc., 22 November 2011. [Online]. Available:

http://blog.comscore.com/2011/11/cyber_monday_work_computers.html. [Accessed 20 01 2012].

[21] R. Feldman and J. Sanger, The text mining handbook: andvanced approaches in analyzing unstructured data, Cambridge: Cambridge University Press, 2007.

[22] Q. Mei, X. Shen and C. Zhai, "Automated labeling of multinomial topic models," in *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2007.

[23] P. Treeratpituk and J. Callan, "Automatically labeling hierarchal clusters," in *dg.o '06: Proceedings of the 2006 International Conference on Digital Government Research*, New York, 2006.

[24] V. Vapnik, Statistical learning theory, New York: John Wiley & Sons, 1998.

[25] K. Nigam, A. McCallum, S. Thrun and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103-134, 2000.

[26] N. Andrews and E. Fox, "Recent developments in document clustering," Department of Computer Science, Virginia Tech, Blacksburg, 2007.

[27] A. McCallum, A. Corrada-Emmanuel and X. Wang, "Topic and role discovery in social networks," Computer Science Department Faculty Publication Series, 2005.

[28] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, Arlington, 2004.

[29] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proc Natl Acad Sci U S A*, pp. 5228-5235, 6 April 2004.

[30] S. Ananiadou, D. Sullivan and W. Black, "Named Entity Recognition for Bacterial Type IV Secretion Systems," 29 March 2011. [Online]. Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0014780. [Accessed 20 January 2012].

[31] R. Batista-Navarro and S. Ananiadou, "Discovering Potential Drugs by Extracting Biological Activities of Natural Products," in *Proceedings of the Workshop on Mining the Pharmacogenomics Literature, Pacific Symopsium on Biocomputing*, Stanford, 2011.

[32] B. Kolluru, L. Hawizy and P. Murray-Rust, "Using Workflows to Explore and Optimise Named Entity Recognition for Chemistry," PLoS ONE, 2011. [Online]. Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020181. [Accessed 12 April 2012].

[33] M. Shaw, C. Subramaniam, G. Tan and M. Welge, "Knowledge Management and Data Mining for Marketing," *Decision Support Systems*, vol. 31, no. 1, pp. 127-137, 2001.

[34] M. Bush, "Text Mining Provides Marketers With the 'Why' Behind Demand," Ad Age Digital, 27 June 2009. [Online]. Available: http://adage.com/article/digital/marketing-text-mining-demand/138110/. [Accessed 12 April 2012].

[35] J. Giles, "2020 vision: The crystal ball internet," NewScientist, 21 May 2011. [Online]. Available: www.newscientist.com/article/mg21028121.900-2020-vision-the-crystal-ball-internet.html. [Accessed 12 April 2012].

[36] C. Metz, "Software: Text Mining, Uncovering telltale patterns," PCMag.com, 1 July 2003. [Online]. Available: http://www.pcmag.com/article2/0,2817,1130911,00.asp. [Accessed 13 April 2012].

[37] D. Sallach, "Data Theory, Discourse Mining and Thresholds," in *AAAI Fall Symposium*, Arlington, 2009.

[38] B. Thuraisingham, "Data Mining for Counter Terrorism," *AAAI Press*, vol. 1, pp. 191-218, 2002.

[39] S. Lee, J. Song and Y. Kim, "An Empirical Comparison of Four Text Mining Methods," *Journal of Computer Information Systems*, vol. 3, pp. 1-10, 2010.

[40] A. Wilson and P. Chew, "Term Weighting Schemes for Latent Dirichlet Allocation," in *COLING-ACL 2006: 21st International Conference on Computational*

Linguistics and 44th annual meeting of the Association for Computational Linguistics, Stroudsburg, 2006.

[41] D. Blei and J. Lafferty, "Topic Models," *Text mining: classification, clustering, and applications* , pp. 71-83, 2009.

[42] T. Miller, "Acquisition program problem detection using text mining methods," Air Force Institute of Technology [Thesis], 2012.

[43] T. Kohonen, Self-Organizing Maps (3rd Edition), New York: Springer, 2001.

[44] J. Millar, G. Peterson and M. Mendenhall, "Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps," in *Proceedings of the 22nd International FLAIRS Conference*, 2009.

[45] H. Bauer and K. Pawelzik, "Quantifying the neighborhood preservation of self-organizing feature maps," *IEEE Transactions on Neural Networks*, vol. 3, no. 4, 1992.

[46] B. Liu, W. Hsu and Y. Ma, "Mining Association Rules with Multiple Minimum Supports," School of Computing National University of Singapore, Singapore, 1999.

[47] W. Lin, S. Alvarez and C. Ruiz, "Collaborative recommendation via adaptive association rule mining," Department of Computer Science Worcester Polytechnic Institute, Worcester, 2000.

[48] M. Forte, C. Hummel, N. Morris, E. Pratsch, R. Shi, J. Bao and P. Beling, "Learning human behavioral profiles in a cyber environment," in *Systems and Information Engineering Design Symposium*, Charlottesville, 2010.

[49] AFI33-129, *Transmission of Information via the Internet*, Secretary of the Air Force, 2001.

[50] X. Wei and W. Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[51] M. Maltais, "Facebook offers suicide-prevention lifeline for military families," LA Times, 9 May 2012. [Online]. Available: http://www.latimes.com/business/technology/la-fi-tn-facebook-military-suicide-20120508,0,1357844.story. [Accessed 9 May 2012].

[52] Anxiety and Depression Association of America, "Understanding anxiety," 2012. [Online]. Available: http://www.adaa.org/understanding-anxiety. [Accessed 20 April 2012].

[53] A. Baum and D. Polsusnzy, "Health psychology: mapping biobehavioral contributions to health and illness," *Annual Review of Psychology,* vol. 50, pp. 137-163, 1999.

[54] N. Anderson and P. Anderson, Emotional Longevity: what really determines how long you live, New York: Viking, 2003.

[55] R. Sinha, "Chronic Stress, Drug Use, and Vulnerability to Addiction," *Annals of the New York Academy of Sciences,* vol. 1141, pp. 105-130, 2008.

[56] K. Alvord, K. Davidson, J. Kelly and K. McGuiness, "Understanding chronic stress," American Psychological Association, 2012. [Online]. Available: http://www.apa.org/news/press/releases/stress/2011/chronic-stress.aspx. [Accessed 20 April 2012].

[57] Law Offices of LaVan & Neidenberg, "Post-Traumatic Stress Disorder," 2012. [Online]. Available: http://www.disabilitylawclaims.com/library/signs-and-symptoms-of-post-traumatic-stress-disorder-ptsd.cfm. [Accessed 20 April 2012].

[58] Law Offices of LaVan & Neidenberg, "Depression," 2012. [Online]. Available: http://www.disabilitylawclaims.com/library/signs-symptoms-definition-of-depression.cfm. [Accessed 20 April 2012].

[59] National Institute of Mental Health, "Depression," U.S. Department of Health & Human Services, NIH Publication No. 11-3561, 2011.

[60] Center for Disease Control and Prevention, "Web-based Injury Statistics Query and Reporting System (WISQARS)," 2011. [Online]. Available: http://www.cdc.gov/injury/wisqars/index.html. [Accessed 16 May 2012].

[61] J. Milton and J. Arnold, Introduction to Porbability and Statistics, New York: McGraw-Hill, 2003.

[62] Applied Statistics Handbook, "Coefficients for Measuring Association," AcaStat, 2012. [Online]. Available: http://www.acastat.com/Statbook/chisqassoc.htm. [Accessed 1 May 2012].

[63] M. Ricks, "Spike in Suicides," *Air Force Times*, 7 May 2012.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
14-06-2012	Master's Thesis	August 2010 – June 2012

TITLE AND SUBTITLE

Cyberspace and real-world behavioral relationships: Towards the application of Internet search queries to identify individuals at-risk for suicide

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Miller, Casey, C., Captain, USAF

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

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

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GCE/ENG/12-08

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

Center for Cyberspace Research
Attn: Maj Jonathan Butts
2950 Hobson Way
WPAFB OH 45433-7765
(937) 255-3636, ext 4332 Jonathan.Butts@afit.edu

10. SPONSOR/MONITOR'S ACRONYM(S)
AFIT/CCR

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Internet has become an integral and pervasive aspect of society. Not surprisingly, the growth of ecommerce has led to focused research on identifying relationships between user behavior in cyberspace and the real world – retailers are tracking items customers are viewing and purchasing in order to recommend additional products and to better direct advertising. As the relationship between online search patterns and real-world behavior becomes more understood, the practice is likely to expand to other applications. Indeed, Google Flu Trends has implemented an algorithm that accurately charts the relationship between the number of people searching for flu-related topics on the Internet, and the number of people who actually have flu symptoms in that region. Because the results are real-time, studies show Google Flu Trends estimates are typically two weeks ahead of the Center for Disease Control.

The Air Force has devoted considerable resources to suicide awareness and prevention. Despite these efforts, suicide rates have remained largely unaffected. The Air Force Suicide Prevention Program assists family, friends, and co-workers of airmen in recognizing and discussing behavioral changes with at-risk individuals. Based on other successes in correlating behaviors in cyberspace and the real world, is it possible to leverage online activities to help identify individuals that exhibit suicidal or depression-related symptoms?

This research explores the notion of using Internet search queries to classify individuals with common search patterns. Text mining was performed on user search histories for a one-month period from nine Air Force installations. The search histories were clustered based on search term probabilities, providing the ability to identify relationships between individuals searching for common terms. Analysis was then performed to identify relationships between individuals searching for key terms associated with suicide, anxiety, and post-traumatic stress disorder. Findings based on the calculated χ^2-test statistic demonstrate a strong correlation between the individuals who searched for the key terms. The results demonstrate the utility of clustering individuals who exhibit similar search patterns and provide the foundation for future efforts to bridge the gap between cyberspace and real-world situational awareness for identifying at-risk individuals.

15. SUBJECT TERMS

cyber behavior, internet search queries, suicide, PTSD, text mining, LDA, SOM, clustering

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Maj Jonathan Butts (ENG)
U	U	U	UU	103	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636, ext 4332 (jonathan.butts@afit.edu)

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