2019

Applying Text Analytics to Derive Value from Blog Posts

Ivan Tang
Pace University

Applying Text Analytics to Derive Value from Blog Posts

Ivan Tang
Computer Science
Faculty Advisor: Andreea Cotoranu
Seidenberg School Of Computer Science and Information Systems

May 8, 2019 Poster Presentation - May 2019 Graduation

Placeholder for Advisor Approval Page

Abstract

Text analytics consists of examining the unstructured data contained in natural language using various methods, techniques, and tools. This form of analytics has been growing in popularity due in part to social media and its ability to record public opinion on topics of interest. Blogs are a powerful form of social media in that they are easy to create and maintain and enable any Internet user to become a content creator. Deriving meaning from such open source content can be applied to many domains including marketing, product development, and intelligence gathering.

We apply text analytic tools to derive value from a cybersecurity blog authored by a prominent figure in the cybersecurity field. The blog is popular among cybersecurity aficionados who have an interest in the author's expert opinion on cybersecurity topics such as vulnerabilities, patches, and privacy. The author has been actively blogging on this platform since 2004.

We implement website scraping tools and generate a text data set by harvesting text from the cybersecurity blog. We process and analyze this text using the Natural Language Toolkit in Python. We process the removing stop-words and punctuation, converting all letters to lowercase, and removing inflectional word endings. We apply word frequency measures to analyze the processed text and reveal the most frequent topics discussed by the blog's author. In addition, we map these frequent topics to a timeline to uncover meaningful patterns in topics of interest across the blog and how discussions of cybersecurity topics have evolved over 15 years in this author's blog.

Table of Contents

Introduction

Social media on the Internet is an integral way to effectively communicate and conduct business. There are currently about 3.886 billion active social media users, which equates to 51.8% of the world's population [13]. Furthermore, these Internet users have an average of 7.6 social media accounts [9]. Every minute about 3.3 million Facebook posts and 448,800 Tweets are constructed. Blogs are also a powerful form of social media as they enable any Internet user to become a content creator and solely express their thoughts and opinions. WordPress is one of the most popular blogging platforms. On WordPress alone, 74.7 million blog posts are published every month [15]. Social media sites archives vast amounts of unstructured data including images, video, and text. It is estimated that only 20% of the data today is structured (e.g. names, addresses, dates, social security numbers); the rest is unstructured (e.g. images, audio files, video files, and text) [9]. Presently, less than 0.5% of publicly available data is being processed [13]. This provides an excess of opportunity for content analysis to be utilized.

Processing and analysis of such large amounts of data is possible with automated web harvesting and text analytics. Deriving value from text with the use of automated tools can be valuable in many domains including but not limited to marketing and intelligence. For example, textual analysis can be used commercially. Businesses are able to gather data from blogs and social media and analyze them to make their decisions. The most common use of social media analytics is to mine customer sentiment to support marketing and customer service activities [13]. In addition, social media provides platforms for customer feedback, which businesses are then able to collect and respond, whether it be positive or negative, to build customer support and confidence.

Blogs, much like other social media platforms such as Reddit, Facebook, or Twitter, archive public opinion and personal information. Through the process of data harvesting, data preprocessing, and applying various analytic tools, large data sets can be processed systematically. Data analysis often unveils patterns in behavior or preferences that are of value in advertising campaigns, product development, and intelligence gathering. With the rise of machine learning, the usage of content analysis has grown exponentially. Machines are able to efficiently categorize sentences or words into specific categories using classifiers and test sets, allowing for a plethora of text to be analyzed.

This research aims to apply text analytic tools to derive value from blog posts. We provide the methods and tools necessary to reproduce and apply onto other text. Instead of collecting raw data from surveys or other user participation, textual analysis relies on open source intelligence and intelligence gathering methods, which is why social media is a big target. In this research, we exemplify the power of these tools through collecting and processing text from a blog authored by a renowned cybersecurity professional. The author has been blogging since 2004, and the blog has been read by over 250,000 people [12]. From this process, we can gain insight into the key topics the author has been writing about and demonstrate the power of text analytics and their applications in other fields.

Literature Review

Access to numerous social media platforms and search engines make open source intelligence gathering easily accessible. The practice of gathering intelligence through open sources, such as blogs, is common across industries today. Scripts can assist with gathering and

storing data from such open sources automatically. While open source intelligence gathering practices have been in existence, the techniques and tools have evolved and are accessible. Text analytics tools have become increasingly important for the processing of large data sets collected from the web. The following aspects of text collection and analysis and their implications are discussed in this literature review: data sources, data collection and storage, and data privacy.

Data Sources

A search for the implementation of text analytics on social media revealed that Facebook and Twitter were popular data sources for scraping. This could be due to the fact that these platforms have application programming interfaces (API)s which facilitate data scraping, or to the fact that these platforms collect and store vast amounts of personal data. This thesis differs from existing work in that data collection and analytics are applied to blogs, which have not been chosen for analysis prior. This thesis shows how scraping tools can be applied to collect data from the web. Furthermore, this shows that text analytic tools can assist with deriving meaning from blogs. We apply such techniques to a blog by a cybersecurity expert.

Facebook.

Soubhik Barari analyzed Facebook confession pages of college campuses to uncover what they would review to other peers or even people on the pages under complete anonymity.[1] Barari found "University students have different stresses and concerns at different points during the academic year. Final exams are likely to be the prime academic stressor for students at the end of an academic term while adjusting to a new social and academic climate is the more imminent concern at the start of a term." [1] He then tried to tie current events to some of the anxiety they were feeling, "global news events, as well as holidays, are

also likely to mobilize students to social media (e.g. major sports events, terrorist attacks)". [1] To achieve his answers for his research questions, he found the universities with confession pages in each state with the most posts; Tufts University (27,605), Purdue University (10,629), University of Toledo (8,686), Biola University (8,288), Cornell University (7,845) and Montana State University (6,845). He did not force any private information and only found what is on the public API.

The table he created categorizes all the posts he found and dives deeper into dividing into each campus with its national rank, as well as the frequency of each type of topic discussed. Barari found that "students at top-ranking elite universities post more about socioeconomics and mental/physical health." [11] He also concludes that they each have a different way about expressing it, however most importantly, they receive support on matters some may consider taboo, such as mental health and socioeconomic status.

Erin Willis and Patrick Ferrucci categorized 122 posts on 30 deceased users' Facebook walls. [11] What they had found was that these posts were motivated by entertainment, followed by integration and social interaction. Here the family members and friends were able to post memories, condolences, and interact with each other. "Prior to Facebook, newspaper obituaries and telephone calls were the most frequently used methods for notification of a person's death (Carroll & Landry, 2010); now, it may be Facebook that alerts us to a friend's passing." [11] Willis and Ferrucci bring up a good point in how news of death now travels a lot quicker and more efficiently. Posting on these walls also came as an emotional release for many users, who use Facebook as "a conduit for connecting with family, friends, and the larger social network

long after the deceased is gone." This especially holds important and influential, as mourning and grief can generally last longer than brief wakes or funerals that might be held. [11]

Twitter.

Kazutoshi Sasahara had taken on the task of finding the food relationship and personal attributes. "Food left-wing" were the vegetarians, while "food right-wing" was the fast food lovers. Basically, attempting to find healthy versus unhealthy eating habits is for which type of person in politics. To complete this, she used the Twitter API to create a scraper, collecting two data sets. 18 food-related keywords were used to gather tweets in Japanese. "Results show that food identity extends beyond the domain of food: the food left-wing has a strong interest in environmental issues, while the food right-wing has a higher interest in large-scale shopping malls and politically conservative issues." [8] This would then suggest that food can also be used as a gateway to finding personal attributes and offer commercial insights.

Kwon et al. also chose the route of using Twitter's public API to study and analyze the general public opinion of users during the Korean saber rattling in 2013. [5] Their study sought out to prove how society would react to certain current events. They gathered tweets using Twitter's API and then split each tweet using keywords."The results show that, while non-rumor narratives focus on policy-level responses to the threat situation in a similar manner to institutionalized opinion polling, rumors are less concerned with official responses, instead of reflective of hegemonic tensions between anti-leftwing political sentiments and the counteractive accounts." [5] Finding each division wasn't hard with keywords, "The OR generally reflected public desire to retrieve a sense of security and safety, while the WDR disclosed the unconscious source of intergroup hostility that could have exacerbated domestic instability under the crisis."

What they also found, that came to no surprise, is that the public uses humor, guesswork, and wishes to cope with fear.

Researchers have used text analysis to understand hate speech and gang violence. Vidgen and Yasseri were able to detect the differences between non-Islamophobia, weak, and strong Islamophobia on Twitter using machine learning. [10] Vidgen and Yasseri used Twitter's Rest API to gain 109,488 tweets produced by far-right accounts during 2017. First, a training set of 4,000 tweets were created. Vidgen and Yasseri then created models and tested them. 100 of the 109,488 Tweets were then sampled for each of the three categories (none, weak, and strong) and created a new dataset of 300 tweets. It was then applied to the total amount of tweets collected and they found, "most tweets are not Islamophobic (57,630 tweets), weak Islamophobia is considerably more prevalent (36,963 tweets) than strong Islamophobia (14,895 tweets)." The textual analysis provides great insight into online Islamophobia and how some users think. Their information was portrayed clear and efficient and when I create charts for my findings, I would hope it could come out as concise.

Using textual analysis and intelligence gathering, Chang et al. easily recognized gang activity on Twitter and sought to investigate. [3] Gang member's "experiences and intents" have been increasingly expressed further on these sites. "In some situations, when they experience the loss of a loved one, their online expression of emotion may evolve into aggression towards rival gangs and ultimately into real-world violence." [3] Chang et al. wanted to detect aggression and loss in social media related to gangs. First, they utilized a pre-existing dataset of tweets from a prominent female gang member, Gakirah Barnes. They then scraped more tweets from 279 users in the same network as Barnes. Again, with the use of data mining, tweets were categorized with

training sets created. This research has the potential to help avoid gang violence by automatically scraping tweets from these gang members which would be too difficult to achieve manually. Monitoring aggression and loss to identify the process, would allow proper intervention and could possibly end it in its beginning stages.

Data Collection and Storage

Batrinca and Treleaven's "Social media analytics: a survey of techniques, tools, and platforms," brings to light the wealth of social media software available to scrape, store, clean, and analyze text. This paper covers the terminology used in this line of research. Scraping refers to "collecting online data from social media and other Web sites in the form of unstructured text." The term is also known as site scraping, web harvesting or web data extraction. [2] Scraping is key in intelligence gathering, as where you scrape from is pivotal to the type of information and analysis performed. Some social networking sites do not provide an application programming interface (API) or access for scraping data. Examples include Bing, LinkedIn, and Skype. "While more and more social networks are shifting to publicly available content, many leading networks are restricting free access, even to academics." [2]

The scraped data can be stored in databases such as SQL databases, relational databases, or in plain text format. Compiling data into these databases makes parsing and organization easier.

Current analysis techniques include computational statistics, machine learning, or complexity science; which are executed in data mining or simulation mining. While there are companies that specialize in text analytics, open-source analytic tools are available as well. The data can then be visualized through charts and graphs.

Textual sources extend beyond social media. Gorskya & MacLeod demonstrate how textual analysis can assist in comparing the leadership norms and expectations in higher education based on published career advertisements from 2000-2004 as opposed to 2010-2014. The researchers scanned the advertisements to create an electronic composite document to assist in the process to be able to make keyword searches easier. Their first step was to create a list of keywords and phrases (common, unique, or not used at all in each timeframe). They then created a structured data collection form to help identify and compare themes. Finally, they placed each advertisement in the group they created to have been the identifying variable (collaboration, transparency, community-centeredness, accountability, and teams).

Their research shows that there was a major difference in the wording chosen for each of the analyzed variables, especially in accountability. While the authors noted the limitations of their study, they hoped to "stimulate further discussion about career pathways, leadership development, and management preparation."[4] This study shows that syntax and word choice have an impact on an advertisement. A similar methodology will be used to categorize blog posts and then perform frequency analysis.

Metoyer et al. applied automatic textual analysis to answer questions such as who, what, when, and where as well as gathering supporting evidence to assist in creating a "narrative visualization" of sports narratives on ESPN. "Support for linking visualizations of this information to the text is an underexplored area that could benefit each of these users." [7] Users are most likely to skip text and go straight to videos, interactive visualizations, and box score tables, which may "limit the reader's ability to interpret the writer's narrative in context." [7] Their integration may be a solution to meet both needs to keep things simple. To achieve this,

they first segmented the text into sentences and sought if each contained a reference to the W's they wanted to answer. They first created a list of NBA players, coaches, and referees. Answering the "what" they used a data mining method. Metoyer et al. created a training set of sentences that included both stats or not at all. "When and where" portion became similar to the stats where they created training sets of sentences. The next difficult part was to create an interactive design that is easily usable.

Metoyer et al. managed to make information easy to consume. Users are able to hover over a stat and an explanation is provided as highlighted text. While there is room for improvement, such as including videos, this completely highlights the use of text analysis as well as the potential it has in the future.

Data Privacy

While it may be easy to perform text analysis on content online, one may need to check if it is ethically correct and how to go about it. Data can be easily sold for quite a bit of money to companies that wish to utilize the data to turn a profit through targeted advertisements. However, this can be morally incorrect as user's privacy are not always kept in mind. This can easily turn abusive or has the potential to influence a user's life, for better or worse. Furthermore, it is even more dangerous given the youth have easier access to the Internet, in which they can be put in a position of putting more data about themselves when they don't fully understand the severity. Once data is put on the Internet, it makes it so much more difficult to remove. It is necessary to take extra precautions in this age of data privacy.

Light et al. took on the task of analyzing a public app that facilitates sex in public places. The researchers "developed a web scraper that carefully collected selected data from the app and

that data were then analyzed to help identify ethical issues". [6] They combatted legal issues by talking to legal scholars and found the best way to go about their research was, "collecting selected data for analytical purposes and disposing of the collection method and collected data once this process was complete was appropriate." [6] Then began the data collection. The app they chose did not include an application programming interface (API), so they created a Python script to collect the data. Light et al. chose to structure the collection in four steps. First, location ID and name, which returned "12,000 public sex locations across several countries and continents". [6] Second, user comments linked to each location, which returned an average of 61 comments per place. Third, additional data about the location. Lastly, profile information, which includes name and amount of comments. The average number of comments per user (120,000 total) was about 6. While not much other data or scripts were shared, they were still successful in their process and finding out the geo-location of each public sex spot. They even found that the most popular thing to ask was if anyone was available. "Using Python to look for the presence of popular words, we were able to establish the top 5 as anyone (7407), here (7032), around (3690), now (3681), and today (1581), with the average length of a comment being 39.2 characters." [6] Their findings could expose people who may live alternative lives, which is why they had to delete their scripts and data sets, which is why they chose to abide by ethics and delete all data sets and scripts they created. This article definitely helps with my thesis as it pertains to textual analysis and intelligence gathering. My thesis will include a data set that others can see and use for their own knowledge or advantage. It is a completely separate topic in dealing with Schneier's public blog and cyber topics, rather than seeking to find public sex locations and the

people who are doing so, which can take a dark turn of extortion, or even for commercial use to profit off these individuals. Handling sensitive information can have ethical implications.

Methodology

We take the following approach to collecting, processing and analyzing the data for this thesis.

1. Identify data source (blog)
2. Select web scraper tool (WebScraper.io)
3. Identify data fields for scraping

Dataset 1: Blog based

4. Scrape full blog and store in .csv file
5. Preprocess data
    a. Convert all letters to lowercase
    b. Remove punctuation, special characters, and digits
    c. Remove stop-words
    d. Tokenize words
6. Filter data by field
    a. Filter data by year
    b. Identify top tags for the full blog and for each year
    c. Filter data by the desired tag
7. Perform tag frequency count

Dataset 2: Tag-based

8. Scrape select posts, by tag, and store in .csv file
9. Preprocess data

   a. Convert all letters to lowercase
   b. Remove punctuation, special characters, and digits
   c. Remove stop-words
   d. Stem words
   e. Tokenize words
10. Perform word frequency count for all posts
11. Create charts and graphs for all the data
12. Draw conclusions based on results

Text Preprocessing

The raw data requires preprocessing before moving to data analysis. Preprocessing or cleaning the data is a process that involves: removing punctuation, special characters, digits, stop-words, stemming the words, making every word into lowercase, and tokenization. Stop-words are words that are commonly used and are deemed irrelevant to data analysis, i.e. and, is, the, etc. This same concept applies to punctuation digits, and special characters, neither of which should be counted and would make our output messier. Stemming is the process of normalizing words into their base or root form by removing morphological affixes, for example, the words “happening” and “happened” would reduce to the root form “happen” after a stemming. This process is automated with Python and its wide variety of libraries, including Natural Language ToolKit (NLTK).

The following stemming algorithms are widely used in text preprocessing: Porter, Snowball (Porter2), and Lancaster (Paice-Husk). We are applying the Snowball Stemmer because the roots of the stem words are more legible.

Tokenization is a major step in preprocessing and allows for strings or sentences to be broken up into individual words, otherwise known as tokens. This allows for understanding the importance of the words in the sentence. For our thesis, we use tokenization to break up sentences within the body of the blog posts to perform a frequency count on each individual token, resulting in which words appear the most.

Term Frequency

After text processing, the data can then be counted for the number of times a word will appear. Each token in the body is then able to be added into an array list to be counted and printed back into a separate .csv. This makes for the creation and analysis of the charts to be much easier and to easily see the separate frequencies that will appear. Stemming was an important step prior to frequency count as the suffix or prefix will count as a different number than the original word. Both of these key tools for content analysis has been provided in a script, as shown in figure 16. The script performs both text preprocessing and term frequency on a given column and will return it in the same file, or can produce a new one if wanted.

Research Questions

Through this study we are aiming to answer the following research questions:

1. What is the topic distribution across the author's blog from 2004-2019?
2. What are the most frequent (top 10) topics discussed in the author's blog from 2004-2019?
3. What are the most frequent topics discussed in the author's blog for each year?
4. What are the most frequent words (top 15) used to discuss each topic?

Experiment Design

Data Collection

A free online Google Chrome extension was used to harvest data from the author's blog, collecting multiple datasets. The Chrome extension, Web Scraper, https://www.webscraper.io, simply requires Chrome and is very simple to use. More documentation and tutorials are provided on their site. This scraper was chosen due to the functionality of the point and click interface and simplicity. It provided a very simple way to automate extraction, even with pagination, and export it in the form of a CSV file, which allows for easier data processing in Python.

Dataset 1.

The web scraper tool is used against the **entire blog** to create a data set to include the following fields: titles, links, dates, and tags for each post on the blog. The web scraper selector graph (Figure 2) highlights the selection path. Scraping starts with the main page (root) and continues with an individual blog post (link) followed by the title, tags, and date of the post. The data set includes 7114 posts posted from 10/1/2004 to 3/14/2019.

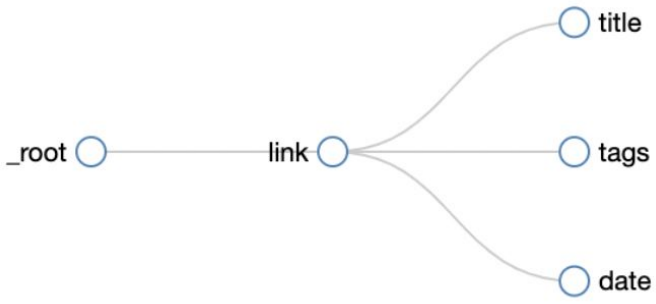


Figure 2 Web Scraper Selector Graph Dataset 1

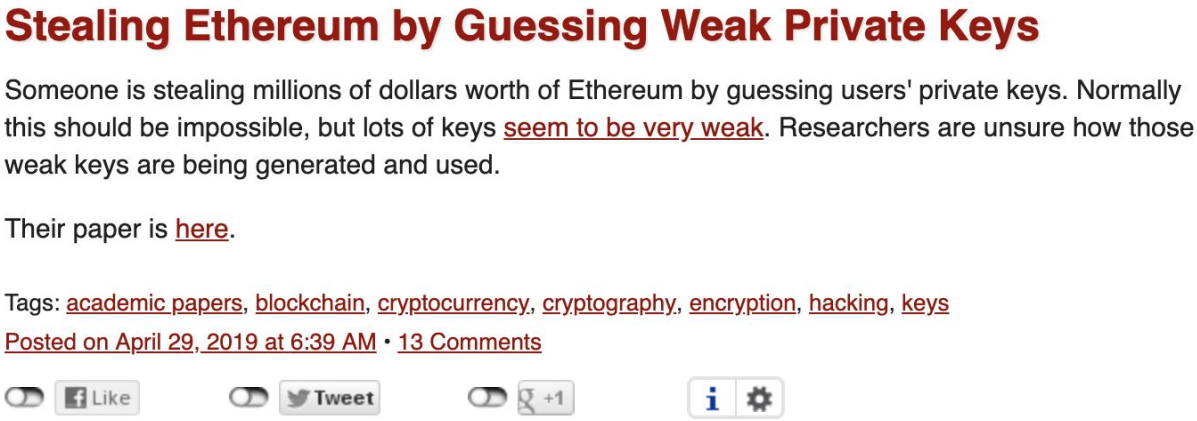


Figure 3. Example of text structure on blog

Dataset 2.

After receiving the topmost recurring topics within the entire blog, we create a second dataset to include only blog posts on the **top-10 topics**. Scraping starts with the main topic page (root) and continues with an individual blog post (link) followed by the entire body. What this allows for is another frequency count to see what words are most chosen by Schneier within the writing.

Figure 4. Web Scraper Selector Graph Dataset 2

Results & Discussion

Research Question 1: What is the topic distribution across the author's blog from 2004-2019?

From the beginning of the blog to when this scrape occurred, the author has written a total of 7,114 blog posts. The distribution of topics was found by the author's usage of tags to display what each article would focus on. The tags can then be quantified through the use of a word counter to unveil the number of times a topic appears.

Being a dedicated cybersecurity blog post, some topics will appear more often than others, resulting in the graph below displaying exponential decay. From a total of 693 topics, security is at the very top of the distribution because this is the main focus of the blog as well as multiple types of security that are possible: computer security, homeland security, economics of security, etc. The topics 1-110 are written about more than 100 times, which results in cell 111-693, falling closer to the x-axis. Cell 458-693 topics are written about less than 10 times, which is at the tail of the graph.

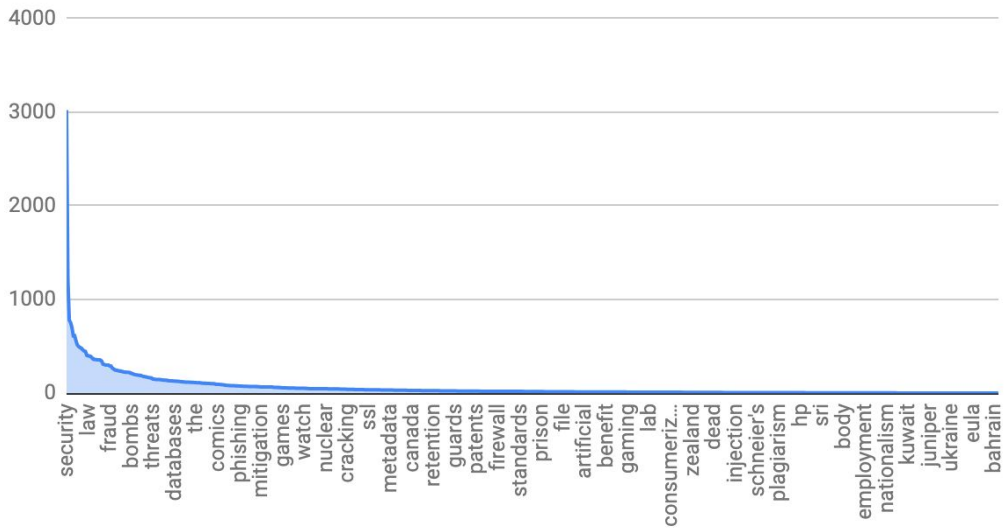


Figure 5. Distribution of 691 Topics by Frequency Across Blog

Research Question 2: What are the most frequent (top 10) topics discussed in the author's blog from 2004-2019?

The graph below follows a similar distribution as the figure 5 in that it will display decay. This will always be the case in how some topics will be discussed more than others. For this distribution, security should be ignored due to the focus of the blog as well as the number of types of security, which inflates the count. Our most frequent topics discussed in the blog from most to least are privacy (1215), terrorism (780), surveillance (754), cryptography (559), hacking (514), NSA (487), national security policy (481), encryption (451), air travel (446), and law enforcement (392).

Privacy in cyberspace has been one of the largest debates of the 21st century. How American legislators should handle privacy has recently surfaced. America has little to no privacy regulations as it is. Data privacy is an important factor and with how easily data is able to be obtained, it comes to no surprise to why the author would write about privacy often.

Sensitive information about oneself should confidential and only be accessible to those authorized to view.

Following privacy in hits is terrorism, with about 400 fewer articles. The duration of the blog occurs through a 15-year span, which also covers years after 9/11. While U.S. President Barack Obama announced on May 23, 2013, that the Global War on Terror is over, this may not be the case. Terrorism could have also taken a place in cyberspace with recruitment, cyberterrorism, airport security, and other lone attacks that the author discusses.

The author writes much about surveillance as a form of espionage. There are many surveillance devices and software being developed and it is important to stay up to date and be aware of what may be happening around us. Many devices may also be hacked into and be used for nefarious acts of surveillance.

Squids and academic papers are skipped and don't count towards popular topics. The author publishes squids on Friday's through a "Friday Squid Blogging" and is focused on the aquatic cephalopods. I am unsure as to why he posts these, however, it shows he has an interest towards them and it is probably a good way to distract himself from the constant cybersecurity topics. Academic papers are also skipped because it is difficult to grab text from them as links are simply provided, rather than the actual text, and the main factor that the author himself didn't publish it.

Cryptography another major topic of discussion. The author is an American cryptographer and has published books and has been involved in the creation of many algorithms, so it comes to no surprise as to why this topic has the ranking it does. Cryptography

is the field of techniques used to secure communication through encryption and decryption with a key.

Hacking is the tag the author uses when he shares vulnerabilities, tools, devices, and companies affected through unauthorized access. This tag may be a generalized idea of what cybersecurity would be like to proponents outside of the field. Cybersecurity, however, involves much more as the tags discussed may indicate. Hacking is another predictable topic to land in the most frequent tags. This is due to the constant vulnerabilities being developed and exploited to cause data breaches and damage to companies, resulting in a lack of trust with consumers. Examples can be found in Uber, Equifax, and Sony, which have had a wave of negative feedback in the way the situation has been handled.

The National Security Agency, otherwise known as NSA, is a government agency of the United States Department of Defense. The NSA holds the responsibility of foreign and domestic monitoring, collection, and processing of information and data for intelligence and counterintelligence purposes. The actions of the NSA have constantly been a matter of political controversy, with the defense of the nation compared to the unethical ways data is obtained.

National security policy is the tag used when the author discusses his perspective on regulatory issues and politics. This provides a forum for debate as well as the presentation of the many dangerous security aspects occurring on a national level. National security policy posts have definitely stepped up considering the lack of policy opposed to our need for it. Consumers should be protected and their privacy should be respected. Too much information can be readily available for anyone to use for selfish desires.

Encryption lies between air and travel, however, for our case is ranked above it due to the fact that air may have been included in other tags, travel shouldn't have been. Encryption falls under cryptography and is the process of encoding a message for only authorized parties to access. What lies in these tags are encryption backdoors, ciphers, and real-world application of encryption. What is a shock to me is that encryption has its own tag when it should technically fall under cryptography as well. I feel as if it is a little bit misleading, as readers would want to find cryptography instead.

Air travel is the 9th most frequent topic. This ranges for a variety of reasons. Based on the FAA, 2,600,000 passengers fly in and out of the U.S. every day [14]. Airports are responsible for these passengers and it is of vital importance that airlines are not hacked and people get to their destination safely. Airport security has to really be stepped up and the TSA has a difficult portion of that task. Within this tag, the author posts much about the TSA security, issues about airplane security, hacks that have occurred to planes and airports, and terrorism on air travel. Air travel was an unforeseen topic for me. When thinking about security, air travel does not usually cross my mind as much as the other topics would. However, it does make sense that it would appear, considering there are a lot of passengers who use air travel and would be negatively affected by something like a hack or TSA not properly doing their job.

Law enforcement was the tenth main topic to be scraped and analyzed. Law enforcement had a total of 392 articles and involved policing, drones, surveillance, and the FBI. Hacking and other cyber crimes are emerging worldwide, with anyone being a potential victim. The complexity and scope of the situation may be too difficult for most law enforcement agencies,

especially when they're the targets and house confidential information. It is crucial to stay up to date and raise awareness of the ongoing issue.

The author is able to post information he finds interesting, as well as events, and outbreaks. These all contribute to his growing daily readers, who find the topics he posts interesting. The topics that are most frequent are ongoing and hold much value to the field. It is necessary to know them and stay informed as they can apply to you at any given moment. Furthermore, this may act as an introduction to the field to someone who may be interested.

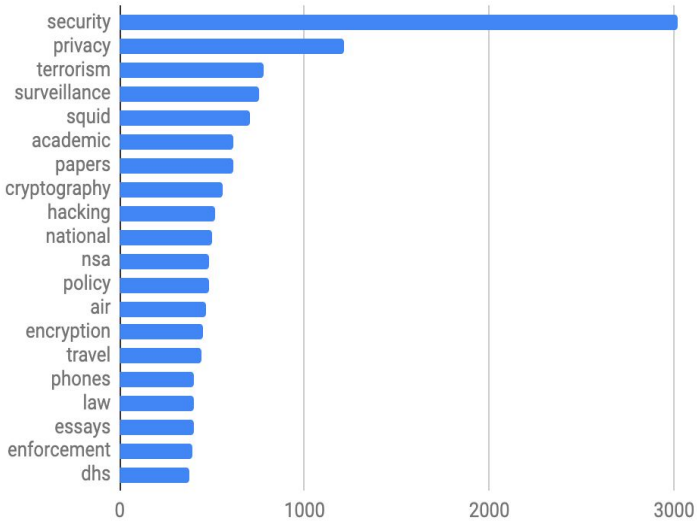


Figure 6. Distribution of Top 15 Topics

Research Question 3: What are the most frequent topics discussed in the author's blog for each year?

Figure 7 is a chart that depicts each year with their total amount of posts and the top 5 topics with their corresponding frequency. To find out a specific year's topics, it is meant to be read horizontally. The top 5 topics are ordered in most frequent to less frequent, topic 1 would be

the most posted for that year. For more information, the top 20 topics per year can be found in figures 17-24 in the appendix.

As time progressed, the author begins to post less throughout the years. His full-year began in 2005, where he posted 535 times. The next year, he posted 636 and 685 the following. Since then, he has been on the decline and has averaged about half the amount in 2017 (373) and 2018 (351). Still, from the duration of 15 years, he has posted 7114 times, which results in an average of 474 posts per year. This is an impressive feat and shows the dedication he has towards his readers.

Privacy and terrorism held the number 1 and 2 spots interchangeably from 2004-2012. Since 2013-present, terrorism hasn't been within the top 5 topics per year. Privacy still continued to be the hot number 1 topic from 2012-2015 and 2018, while remaining in the top 5 if it wasn't. This shows that user privacy is a top issue within cyberspace where public data is being bought and sold on a daily basis for commercial gain. Terrorism holding a spot in the most frequent topic show that this issue was ongoing for 8 years, and people were afraid of what was to come. The author is able to use this medium as a method of spreading information and informing the public of methods of recruitment, attacks, as well as prevention that may have occurred. Terrorism can also apply to cyberterrorism, where incidents may occur on the internet to cause disruption or chaos in order to achieve political or ideological gains.

From this data, we can also infer what world events may have peaked during the year. For example, for the year of 2013, we see the NSA's first introduction into the top 5 topics of the year at the number 3 spot after surveillance (119), following at number 5 is Edward Snowden (65). This sets up a narrative for what has occurred in 2013. Edward Snowden is an American

whistleblower who previously worked at the CIA. Snowden leaked highly classified information from the NSA and disclosed numerous of the global surveillance programs being run. Since then, his U.S. passport has been revoked and has charges against him, including violating the Espionage Act of 1917 and the theft of government property.

In 2016 and 2017, hacking was the most frequent topic, which can be a result of data breaches, a wide variety of vulnerabilities being found, or even the 2016 Presidential Election, where people suspected Russians had an influence. Since 2015, national security policy has mostly held the number 2 spot. This can be due to the aftermath of Snowden has put privacy into the minds of the public. Privacy regulation should be in place to protect the public and their data.

Year	Total Posts	Topic 1	Frequency	Topic 2	Frequency	Topic 3	Frequency	Topic 4	Frequency	Topic 5	Frequency
2004	58	Essays	17	Privacy	14	terrorism	12				
2005	535	privacy	125	terrorism	88	law	70	crime	61	economic	58
2006	636	privacy	127	terrorism	91	surveillance	71	air travel	62	crime	60
2007	685	terrorism	123	privacy	86	law enforcement	71	air travel	63	surveillance	55
2008	627	terrorism	117	privacy	77	air travel	74	homeland security	52	surveillance	51
2009	538	privacy	91	terrorism	78	cryptography	53	psychology	43	surveillance	40
2010	514	terrorism	87	privacy	77	air travel	51	cryptography	47	DHS	43
2011	447	privacy	62	terrorism	48	hacking	45	DHS	39	cryptography	37
2012	451	privacy	44	terrorism	42	cryptography	39	DHS	36	air travel	33
2013	494	privacy	140	surveillance	119	NSA	108	national security policy	66	Edward Snowden	65
2014	453	privacy	138	NSA	135	surveillance	121	phone	65	exploit	49
2015	453	privacy	93	surveillance	72	hacking	53	national security policy	52	encryption	50
2016	434	hacking	68	national security policy	63	privacy	54	encryption	43	surveillance	35
2017	373	hacking	64	national security policy	43	internet	42	privacy	41	NSA	38
2018	351	privacy	40	national security policy	38	encryption	34	vulnerabilities	33	hacking	27
2019	65	vulnerabilities	12	internet of things	8	cybersecurity	6	hacking	6	encryption	6

Figure 7. A chart depicting the years with top 5 topics and frequencies

Research Question 4: What are the most frequent words (top 15) used to discuss each topic?

Words in the dataset may be obfuscated due to stemming, which has occurred to allow for more accurate readings. To find the most frequent words used, each of the body of the blog posts was scraped and cleaned to find which will appear the most. Once again, overall, the term

“secur”, which may refer to secure or security will be found within the top more often than other terms. This is due to the subject of the blog, security, which needs to occur on all bases of our main topics.

To discuss privacy, data (2120), use (1669), and privacy (1508) were the most frequent. Data refers to what should be private. Use is what happens to the data in terms of privacy. Other important terms in this topic are surveil, NSA, and companies. The NSA has been notorious for surveilling the public, whether it is ethical or not. Companies are also known to violate Internet privacy and collect consumer data on users for commercial gain.

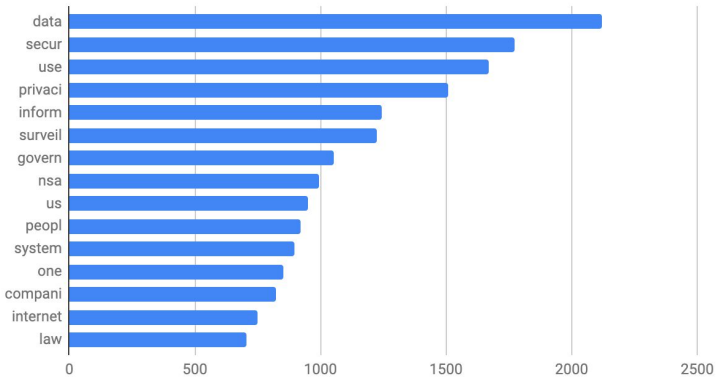


Figure 8. Top 15 Words Used in Privacy Posts

Within terrorism posts, the author uses terrorist (1673), terror (986), and people (837) in his writing. Terrorists are the people who are causing harm and terror, while the people or the public are the ones receiving this harm. The terrorists also attack (764) and make threats (558). Terrorism occurs on many levels, even in cyberspace, hoping to cause harm and spreading its ideologies to more followers.

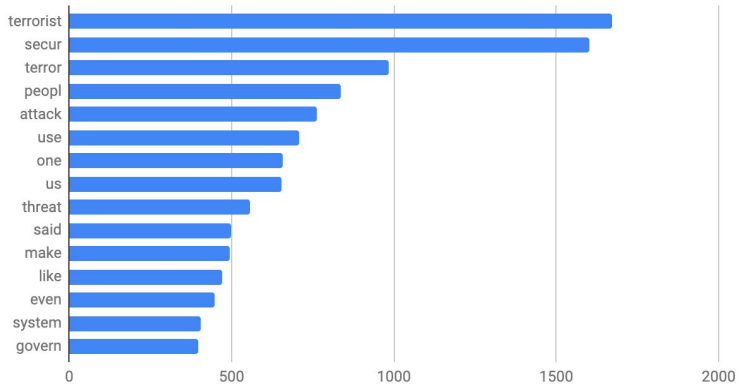


Figure 9. Top 15 Words Used in Terrorism Posts

Surveillance uses the words surveil (1231), data (1171), and use (1112). As previously mentioned, surveillance is a form of spying on the public and collecting data. It is notable that many topics coincide, for example, privacy and surveillance may have a connection in that it would violate privacy when spying occurs without proper authorization. Many devices, such as phones (535) are able to be hacked unknowingly and can be used to eavesdrop.

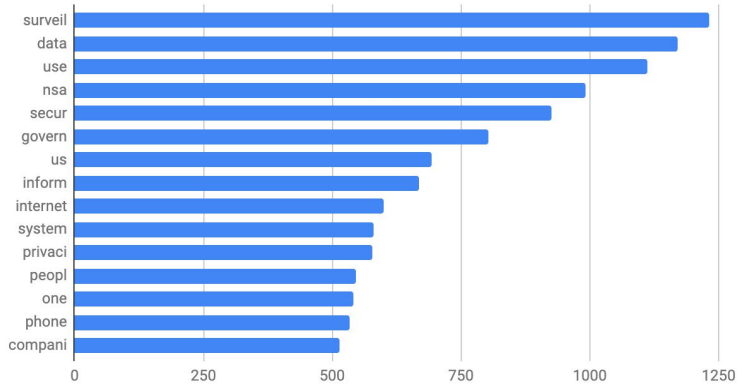


Figure 10. Top 15 Words Used in Surveillance Posts

Cryptography is our author's main field of study. Text is able to be encrypted (1141) and decrypted with a key (565). These algorithms (226) are still able to be crack through brute force with an attack (410). Hopefully, with the development of quantum computing (357), algorithms

are able to evolve to another level and become uncrackable. As previously mentioned, encryption would also fall under the same category as cryptography and has a very similar word usage.

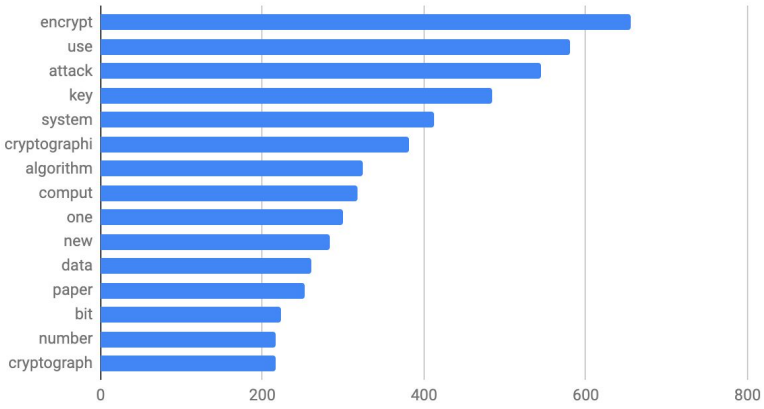


Figure 11. Top 15 Words Used in Cryptography Posts

The topic hacking has the words attack (986), hack (720), and use (623) at its most frequent. A hacker (470) would hack into a system (538) through an attack that exploits vulnerabilities (398) within the system. The hacker would then be able to obtain data (445) and use it for personal gain, or simply to cause destruction. Without proper security, many devices are able to be victims of attacks and infect networks, which could then spread to other devices. Examples can be seen in new IoT devices, where companies are rapidly pushing out products with weak onboard web interfaces and open ports, which are used to manage the device. These devices are then susceptible to attacks and result in a loss of human life where applicable.

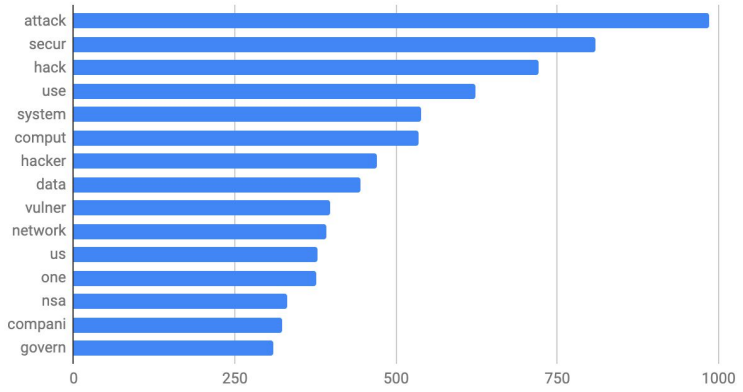


Figure 12. Top 15 Words Used in Hacking Posts

National security policy involves the government (832) and its lack of regulation within the U.S. (782). There is a very similar usage of the words data (584), NSA (583), and use (583), which could possibly indicate that these were always used in the same sentence. There are currently little to no privacy regulation, however, many critics say the U.S. should follow suit of the EU with the passing of the GDPR. The GDPR is OPT-IN legislation to protecting consumer data. The Consent Act has been proposed by Senator Richard Blumenthal and Ed Markey, which is actually similar to the GDPR in that it requires explicit opt-in from users before transactions regarding their data can be made.

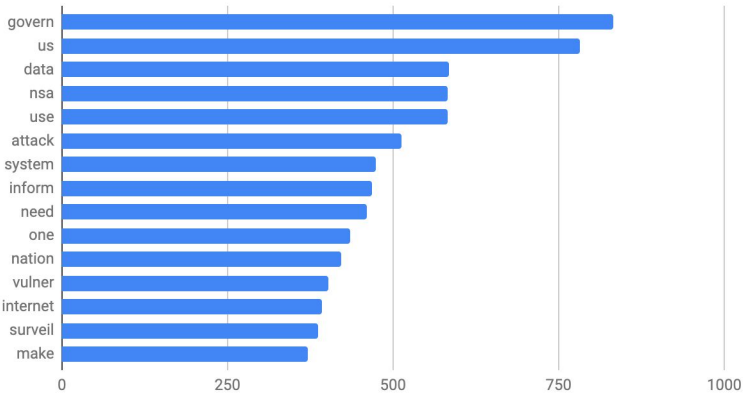


Figure 13. Top 15 Words Used in National Security Policy Posts

Air travel involves many airports (730) with TSA agents (649) to perform screenings (309) to ensure the safety of the public. Sadly, terrorist (517) is a common word used in air travel and invokes fear as passengers (432) are at risk when traveling. Air travel security is necessary as many people travel on a daily basis.

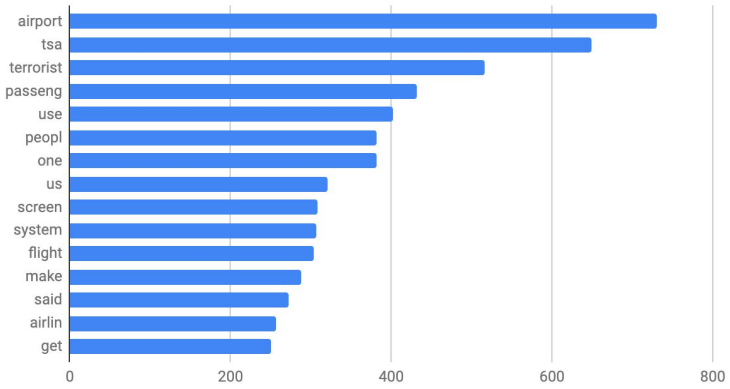


Figure 14. Top 15 Words Used in Air Travel Posts

Law enforcement is the last of the top 10 topics. The words utilized in the posts include police (642), law (362), data (250). Police are at the top as they are responsible for enforcing the laws, along with the FBI (171) in some cases, they are able to stop cybercriminals. With the introduction of new software and computing (231), there come new ways to crack cases, for example, digital forensics.

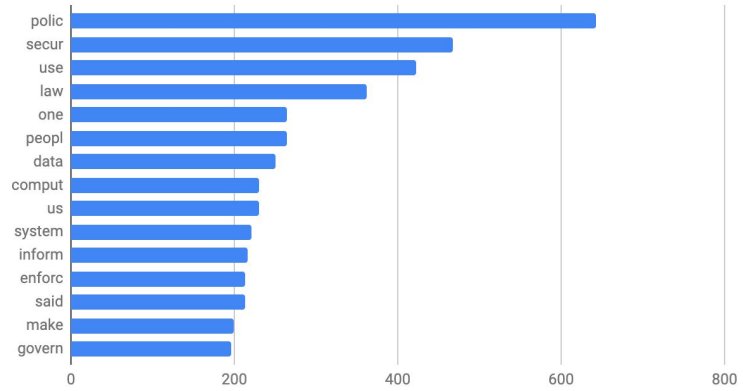


Figure 15. Top 15 Words Used in Law Enforcement Posts

Conclusion, Limitations, & Contributions

We harvested a total of 7,114 posts from a blog authored by a renowned cybersecurity professional using an open source web scraping tool and applied text analytic techniques to clean and normalize the text. Afterwards, word (unigram) frequency measures were applied to gain insight from the author's blog posts. Textual analytic methods and programs functioned as the tools in the automated scraping and processing of the 15 years of content, allowing for information to be portrayed accurately on a graph. The word distribution, as shown in figures 8-15, depicts an exponential decay; this means that the author is using some terms more frequently than others. However, the analysis shows that this distribution is similar for all topics and across the years, and therefore not necessarily a reflection of the author, but rather a reflection of how words are distributed in a given text.

The analysis revealed the topics the author writes about most frequently. The top 10 topics posted included: privacy, terrorism, surveillance, cryptography, hacking, NSA, national security policy, encryption, air travel, and law enforcement. Each of these topics holds importance in the field of cybersecurity today. This analysis can easily be applied to other blog posts or other forms of social media with the provided tools and script to uncover important insight and topics pertaining to scope provided.

With what has been demonstrated with the textual analysis included in this paper, it is easy to transpose these methods and tools on to other platforms for different purposes. While some APIs currently exist to assist with harvesting data, this is constantly changing with the arrival of privacy issues, as seen with Facebook allowing the denial of targeted advertisements. Twitter's API and many other social media sites are following suit allowing publicly available

content to be scraped. This can be bypassed with other open source tools found on the Internet. Scraping these posts from various platforms gives the ability to perform sentiment, advertisement, and marketing analysis to ultimately benefit consumers and producers. Researchers also have the opportunity to change lives as seen with the ability to monitor gang activity, gain insight into how users' opinions on certain social issues and digest information portrayed.

Differing from other work, provided are the tools and methods necessary for harvesting and applying textual analytics on other blogs or social media sites. Figure 16, in the appendix below, is the Python script created which houses most preprocessing steps and the unigram frequency counter to be applied on specific columns. The script is intuitive enough to be easily comprehended and modified to fit imperative needs. In this thesis, we simply show an example of how to apply web harvesting and text analytics to derive meaning from social media.

Works Cited

[1] S. Barari, "Anxiety, Alcohol, and Academics: A Textual Analysis of Student Facebook Confessions Pages," *Tufts Independent Data Journal*, Feb. 2015.

[2] B. Batrinca and P. C. Treleaven, "Social media analytics: a survey of techniques, tools and platforms," *Ai & Society*, vol. 30, no. 1, pp. 89–116, 2014.

[3] S. Chang, R. Zhong, E. Adams, F. Lee, S. Varia, D. Patton, W.R. Frey, C. Kedzie, and K. McKeown, "Detecting Gang-Involved Escalation on Social Media Using Context." *EMNLP*, 2018.

[4] D. Gorsky and A. Macleod, "Shifting norms and expectations for medical school leaders: a textual analysis of career advertisements 2000–2004 cf. 2010–2014," *Journal of Higher Education Policy and Management*, vol. 38, no. 1, pp. 5–18, 2015.

[5] K. H. Kwon, C. C. Bang, M. Egnoto, and H. R. Rao, "Social media rumors as improvised public opinion: semantic network analyses of twitter discourses during Korean saber rattling 2013," *Asian Journal of Communication*, vol. 26, no. 3, pp. 201–222, 2016.

[6] B. Light, P. Mitchell, and P. Wikström, "Big Data, Method and the Ethics of Location: A Case Study of a Hookup App for Men Who Have Sex with Men," *Social Media Society*, vol. 4, no. 2, p. 205630511876829, 2018.

[7] R. Metoyer, Q. Zhi, B. Janczuk, and W. Scheirer, "Coupling Story to Visualization," *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI 18*, 2018.

[8] K. Sasahara, "You are what you eat: A social media study of food identity." *CoRR* abs/1808.08428, 2018.

[9]K. Smith, "123 Amazing Social Media Statistics and Facts." *Brandwatch*, www.brandwatch.com/blog/amazing-social-media-statistics-and-facts/.

[10] B. Vidgen and T. Yasseri, "Detecting weak and strong Islamophobic hate speech on social media." *CoRR, abs/1812.10400, 2018*.

[11] E. Willis and P. Ferrucci, "Mourning and Grief on Facebook: An Examination of Motivations for Interacting With the Deceased," *OMEGA - Journal of Death and Dying*, vol. 76, no. 2, pp. 122–140, 2017.

[12] B. Schneier, "Schneier on Security," Blog. [Online]. Available: https://www.schneier.com/blog/about/.

[13] O. Wassén, "Big Data Facts - How Much Data Is out There?" *NodeGraph*, 1 Mar. 2019, www.nodegraph.se/big-data-facts/.

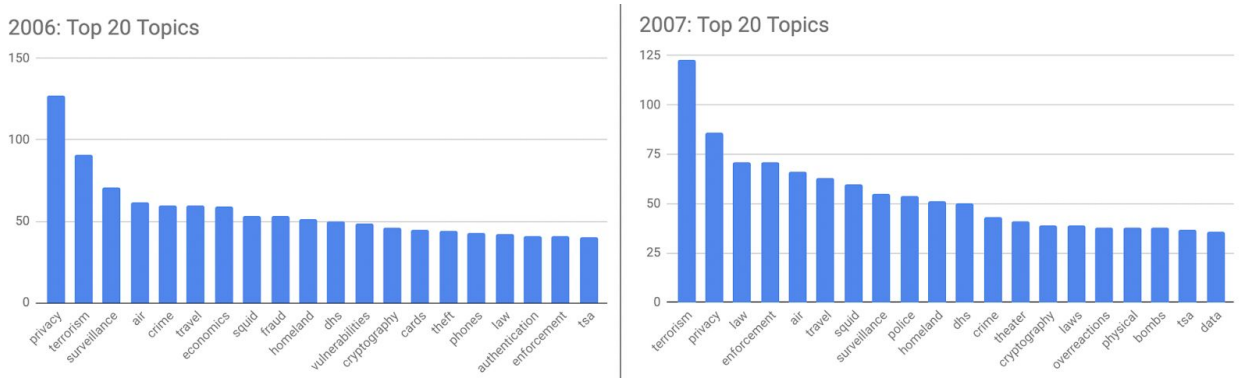
[14] https://www.faa.gov/air_traffic/by_the_numbers/

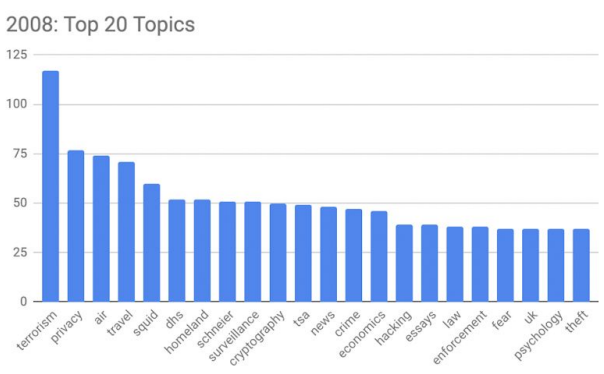
[15] https://wordpress.com/activity/

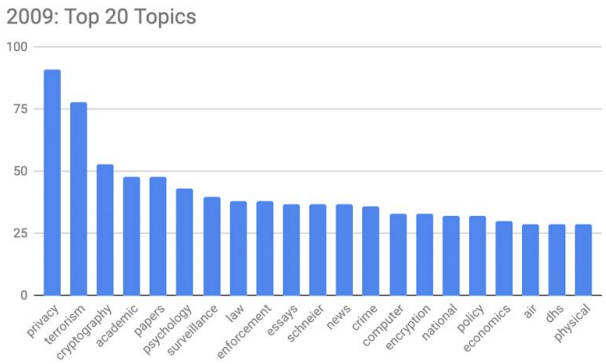
Appendix

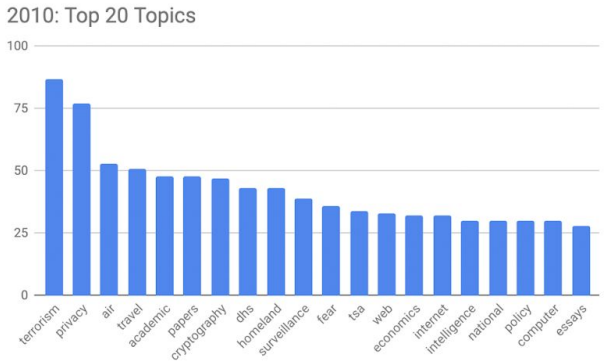
```
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
from nltk.stem.snowball import SnowballStemmer
import nltk
import pandas as pd
import csv
stop = stopwords.words('english')

set(stopwords.words('english'))

df = pd.read_csv("hacking.csv")
df['body'] = df['body'].str.replace(r'[^\w\s]+', '')
df['body'] = map(lambda x: x.lower(), df['body'])
df['body'] = df['body'].apply(lambda x: ' '.join([item for item in x.split() if item not in stop]))
df['body'] = df['body'].str.replace(r'\\b\\w\\b',' ').str.replace(r'\\s+', ' ').str.replace('\d+', ' '
df.to_csv("new_hacking.csv")

twt=df['body'].str.cat(sep=' ')
df = pd.read_csv("new_hacking.csv")

porter = SnowballStemmer('english') # Stemming code
twt=porter.stem(twt)
tojoin=[]
for word in twt.split(" "):
    tojoin.append(porter.stem(word))

twt=' '.join(tojoin) # Stemming code

texts=[twt] #Frequency Code
cv = CountVectorizer()

cv_fit=cv.fit_transform(texts).toarray().tolist()[0]
names=cv.get_feature_names()

auxList=[]
for i in range(len(cv_fit)):
    auxList.append([names[i],cv_fit[i]])
auxList.sort(key=lambda x: x[1],reverse=True)

print('word frequencies:')
for i in range(len(auxList)):
    print(auxList[i][0],' : ',auxList[i][1]) #frequency code

with open("stemmed_hacking.csv", "w") as f:
    writer = csv.writer(f)
    writer.writerows(auxList)
```

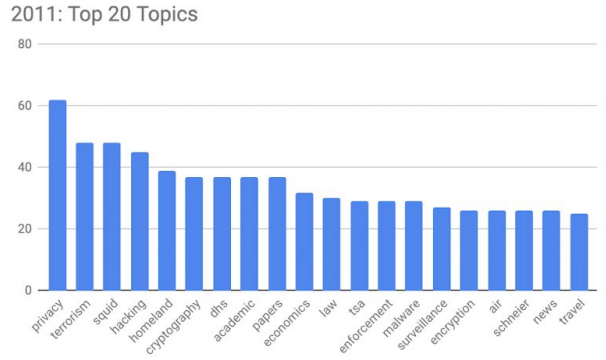
Figure 16. Python script used to preprocess data

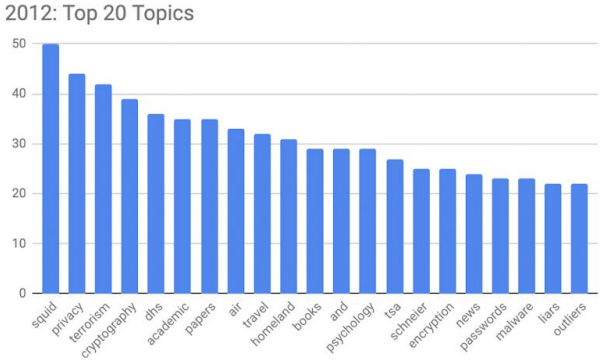


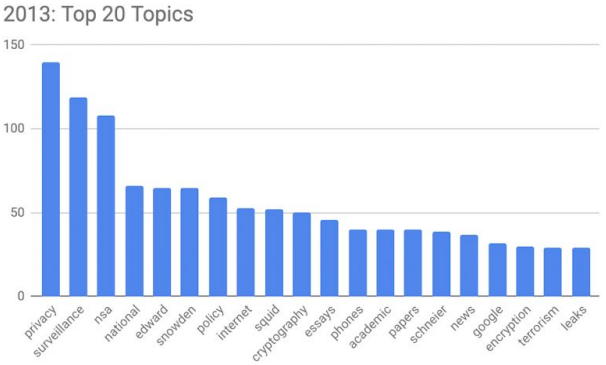

2008: Top 20 Topics
125
100
75
50
25
0
terrorism
privacy
air
travel
squid
dhs
homeland
schneier
surveillance
cryptography
tsa
news
crime
economics
hacking
essays
law
enforcement
fear
uk
psychology
theft

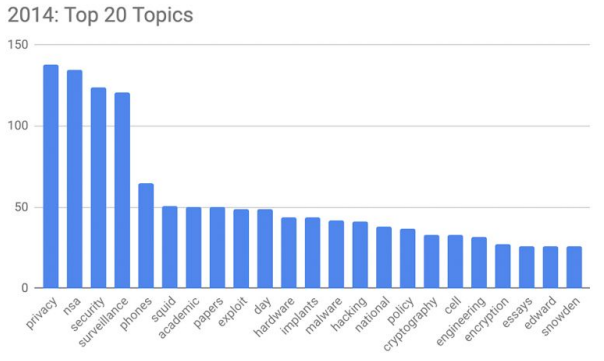

2009: Top 20 Topics
100
75
50
25
0
privacy
terrorism
cryptography
academic
papers
psychology
surveillance
law
enforcement
essays
schneier
news
crime
computer
encryption
national
policy
economics
air
dhs
physical

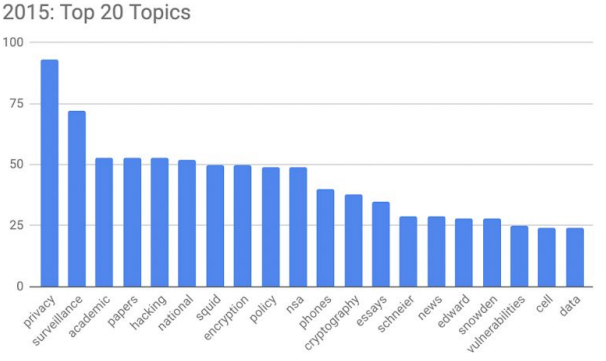

2010: Top 20 Topics
100
75
50
25
0
terrorism
privacy
air
travel
academic
papers
cryptography
dhs
homeland
surveillance
fear
tsa
web
economics
internet
intelligence
national
policy
computer
essays


2011: Top 20 Topics
80
60
40
20
0
privacy
terrorism
squid
hacking
homeland
cryptography
dhs
academic
papers
economics
law
tsa
enforcement
malware
surveillance
encryption
air
schneier
news
travel


2012: Top 20 Topics
50
40
30
20
10
0
squid
privacy
terrorism
cryptography
dhs
academic
papers
air
travel
homeland
books
and
psychology
tsa
schneier
encryption
news
passwords
malware
liars
outliers


2013: Top 20 Topics
150
100
50
0
privacy
surveillance
nsa
national
edward
snowden
policy
internet
squid
cryptography
essays
phones
academic
papers
schneier
news
google
encryption
terrorism
leaks


2014: Top 20 Topics
150
100
50
0
privacy
nsa
security
surveillance
phones
squid
academic
papers
exploit
day
hardware
implants
malware
hacking
national
policy
cryptography
cell
engineering
encryption
essays
edward
snowden


2015: Top 20 Topics
100
75
50
25
0
privacy
surveillance
academic
papers
hacking
national
squid
encryption
policy
nsa
phones
cryptography
essays
schneier
news
edward
snowden
vulnerabilities
cell
data

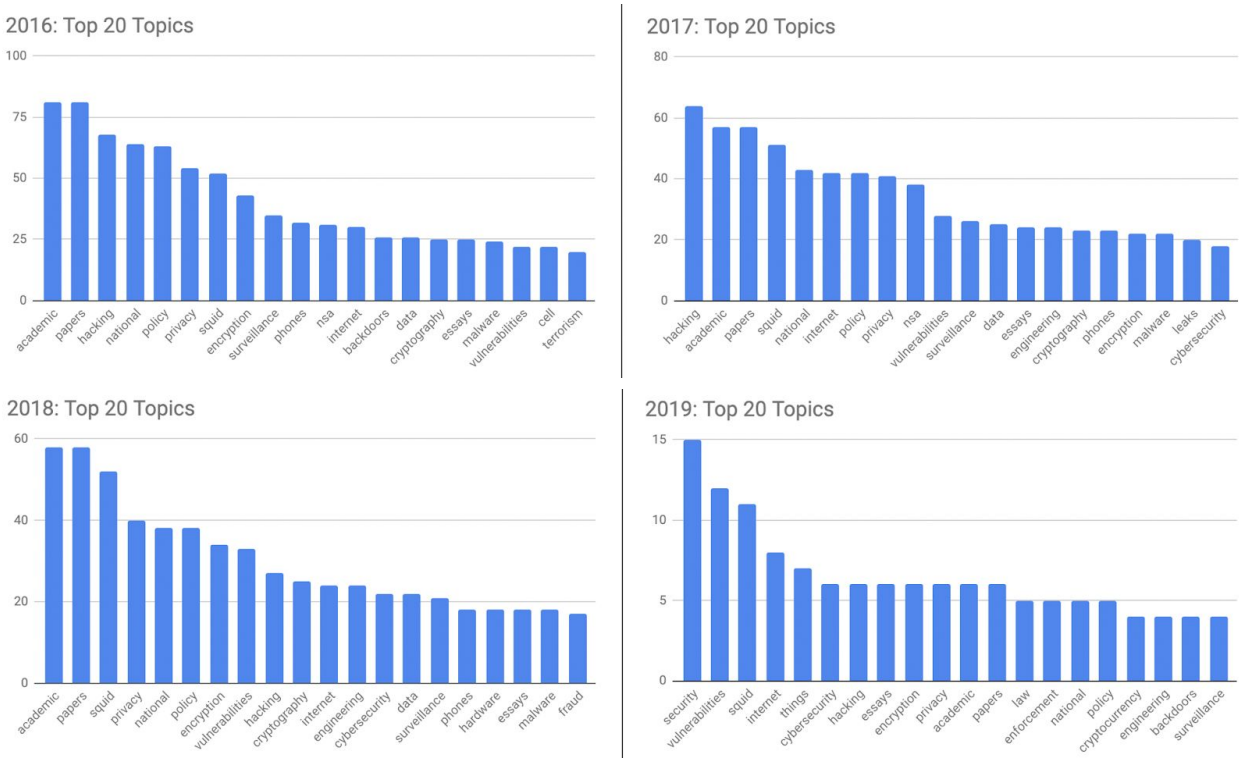


Figure 17-24. Distribution of top 20 topics over the years (2004-2019)