

Western Kentucky University
TopSCHOLAR®

Masters Theses & Specialist Projects

Graduate School

5-2014

Using Statistical Methods to Determine Geolocation Via Twitter

Christopher M. Wright

Western Kentucky University, wright715@gmail.com

Follow this and additional works at: <http://digitalcommons.wku.edu/theses>



Part of the [Communication Technology and New Media Commons](#), [OS and Networks Commons](#), [Software Engineering Commons](#), and the [Theory and Algorithms Commons](#)

Recommended Citation

Wright, Christopher M., "Using Statistical Methods to Determine Geolocation Via Twitter" (2014). *Masters Theses & Specialist Projects*. Paper 1372.

<http://digitalcommons.wku.edu/theses/1372>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact connie.foster@wku.edu.

USING STATISTICAL METHODS TO DETERMINE GEOLOCATION VIA
TWITTER

A Thesis
Presented to
The Faculty of the Department of Physics and Astronomy
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
of the Requirement for the Degree
Master of Science

By
Christopher M. Wright

May 2014

USING STATISTICAL METHODS TO DETERMINE GEOLOCATION VIA
TWITTER

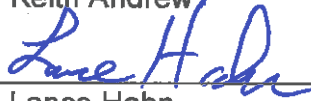
Date Recommended 5/1/2014



Dr. Phillip Womble - Director of Thesis



Dr. Keith Andrew



Dr. Lance Hahn



Dean, Office of Graduate Studies and Research

6-4-14

Date

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Phillip Womble, for his mentorship, guidance, patience, time, and for a workspace throughout my graduate studies at Western Kentucky University. I would also like to thank Dr. Keith Andrew for his kind advice and willingness to always lend time to help.

Also, I would like to thank EWA Government Systems Inc. for their help in data collection and for lending me a place to work on this thesis. I would like to thank Dr. Lance Hahn for his help with getting me started on research and for his help in writing the code for the data collection.

Thank you to a few of my undergraduate professors at Lindsey Wilson College, Dr. Scott Dillery and Dr. Mark McKinnon. You all made every class throughout college fun and entertaining. For being there for advice about things in an out of the classroom and for pointing me in the right direction in life.

Finally I would like to thank my lovely girlfriend, Whitley Gribbins, who stuck with me through the long nights, moments of doubt and stress and whose love and devotion has been a constant rock through my time at WKU. Her patience throughout my wavering moods is proof of her unwavering love. I could not have gotten through this without you. Thank you to my parents, Mary Newcomb and Gary Wright, who have always been a major inspiration and source of encouragement for me. Putting up with all of my late night calls and reinforcing your belief in me kept my mind on track. And to all of my friends and family, you know who you are. Thank you all!

CONTENTS

| | |
|-------------------|----|
| Introduction..... | 1 |
| Related Work..... | 11 |
| Results | 21 |
| Conclusion..... | 25 |

LIST OF FIGURES

| | |
|---|----|
| Figure 1. Map of US with 50 different cities highlighted..... | 17 |
| Figure 2. Text File to N-gram and Frequency Flow Chart..... | 18 |
| Figure 3. Flowchart of Excel Macro: Finding Unique Words among 50 Cities | 20 |
| Figure 4. Map of Cities with Unique entries (green): At least 50 occurrences | 22 |
| Figure 5. Map of Cities with Unique entries (green): At least 20 occurrences | 23 |

LIST OF TABLES

| | |
|---|----|
| Table 1. Types of OSN | 3 |
| Table 2. List of Twitter Search Operators | 15 |
| Table 3. Results for Words with at least 50 occurrences | 21 |
| Table 4. Results for Words with at least 20 occurrences | 23 |

USING STATISTICAL METHODS TO DETERMINE GEOLOCATION VIA TWITTER

Chris Wright

May 2014

32 Pages

Directed by: Dr. Phillip Womble, Dr. Keith Andrew, and Dr. Lance Hahn

Department of Physics and Astronomy

Western Kentucky University

With the ever expanding usage of social media websites such as Twitter, it is possible to use statistical inquires to form a geographic location of a person using solely the content of their tweets. According to a study done in 2010, Zhiyuan Cheng, was able to detect a location of a Twitter user within 100 miles of their actual location 51% of the time. While this may seem like an already significant find, this study was done while Twitter was still finding its ground to stand on. In 2010, Twitter had 75 million unique users registered, as of March 2013, Twitter has around 500 million unique users.

In this thesis, my own dataset was collected and using Excel macros, a comparison of my results to that of Cheng's will see if the results have changed over the three years since his study. If found to be that Cheng's 51% can be shown more efficiently using a simpler methodology, this could have a significant impact on Homeland Security and cyber security measures.

CHAPTER I

INTRODUCTION

Human Intelligence

Human Intelligence (HUMINT) is defined as any information that can be gathered from human sources. HUMINT is collected for the purpose of acquiring information from individuals who access vital foreign intelligence on the full range of our national security issues (CIA, 2010). It plays a critical role in developing and implementing U.S. foreign and national security policy and in protecting U.S. interests. (CIA, 2010)

Open Source Intelligence

Open Source Intelligence (OSINT) is information that can be gathered from publically available sources as well. Examples may include websites one may browse, broadcasts watched, or articles we read. (CIA, 2010) This type of intelligence plays a crucial role in giving insight and context to the national security community for a relatively low cost.

OSINT can be obtained from publicly available material, including:

- The Internet
- Traditional mass media (e.g. TV, radio, newspapers, articles)
- Specialized journals, conferences, public studies
- Photos
- Geospatial information (e.g. maps and commercial imagery)

OSINT has been an important part of all-source analysis, but society's progression in information technology has given a voice to a larger group of people and made it possible to answer new intelligence questions.

“For example, open sources can tell us how various groups overseas react to a speech by the president, we don't have to settle for the 'official' view but can assess various groups' perceptions as well as track trends over time.” – OSC Director Douglas Naquin (CIA, 2010)

SOCIAL NETWORKING

What is a social network?

A social network in its most simple definition is a set of people (or organizations or other social entities) connected by a set of social relationships (friendship, co-workers, or simply exchanging information) (Wilson, 2008). Social networks have been around since societies have been able to interact. Whether it be Neanderthal tribes communicating with other tribes about when the next boar roast social will be, or in the current generation with the 'Kevin Bacon Game', where we connect seemingly insignificant actors to one another based upon their movie roles and the costars of their costars, etc.

Traditional social networks usually consist of an actual meeting with a person or group of persons about a specific topic or interest. For example, a book club, a University alumni gathering, or even a sporting event could all be considered traditional social networks.

Online Social Networks

The sudden emergence of online social media websites, henceforth referred to as OSNs, began to explode in the mid-1990s when people began to realize that the Internet was an amazing resource to spread the word about their new business, or their views on certain topics that may be of interest to a select group of people.

Table 1: Types of OSN

| | | |
|-------------------------------------|---|--|
| Business Networking & Professionals | Business networking communities are for like-minded professionals to connect. | <ul style="list-style-type: none"> • LinkedIn.com |
| Family | Family-based social networks are places for families to stay connected. These have the possibility to be private so that only family members can view the site. | <ul style="list-style-type: none"> • CafeMom.com • Ancestry.com |
| Friends | The largest group of OSN, used to stay connected with friends. | <ul style="list-style-type: none"> • Facebook.com • Friendster.com • MySpace.com |
| Hobbies & Interests | Many online communities are centered on hobbies and interests. These OSNs are for people interested in arts, politics, movies, books, etc. | <ul style="list-style-type: none"> • Flixter.com • Coastr.com • vSocial.com • Shelfari.com |

| | | |
|------------------------------|--|---|
| Languages | Language communities are developed to help learn a foreign language by using others within the community to practice their language skills. | <ul style="list-style-type: none"> • Chalu • FriendsAbroad.com |
| Photo, Video & Audio Sharing | These media communities allow people to post different types of media to share with others to view, rate, share, and comment on. | <ul style="list-style-type: none"> • YouTube.com • Flickr.com • Picasa.com • Last.FM • Pandora.com |
| Shopping | Shopping OSNs allow members to share sales and discounts at local retailers and to give reviews on products. | <ul style="list-style-type: none"> • MyStore.com |
| Social Bookmarking | These websites allow members to store their favorite links from the Internet and shared with other members. | <ul style="list-style-type: none"> • Reddit.com • StumbleUpon.com |
| Students | These OSNs are primarily for students. Facebook.com in its earliest days was an OSN for only students to interact with one another, but was opened up to everyone after popularity rose. | <ul style="list-style-type: none"> • AlumWire.com • RateMyProfessor.com |
| Travel | Travel-themed OSNs are a way to learn about travel destinations by reading other members experiences. These are also a good way to meet people locally. | <ul style="list-style-type: none"> • TravBuddy.com • Zoodango.com |

(Delgado, 2013)

Online Social Networking versus Traditional Social Networking

There are differences between these OSNs and more traditional style of social networks. OSNs are more convenient in this fast paced world that we currently live in. OSNs also allow for easier research methods than a traditional social network by allowing the creation of algorithms that can allow for social networking analysis to be computed more easily. This keeps a researcher from having to conduct an in-person survey, for example (Wilson, 2008). Traditional social networks have that personal connection with meeting with another person and communicating in a social setting and getting a feel of who the person really is. While in an OSN a user does not know who he/she might be really dealing with behind the keys of their keyboard.

OSN websites, like Facebook and Twitter, have grown exponentially over the past few years. A staggering one in four people across the world will use some form of social networking in 2013, nearly 1.73 billion people and it is projected to grow even larger in the coming years (eMarketer, 2013). Twitter alone as of March 2013 has reached over 500 million unique users in its seven years of existence (Smith, 2013). With such a large population size on Twitter, this creates a large diversified data pool that can be used to accurately display the views of the population as a whole.

This ever expanding usage of Twitter allows us the possibility to use statistical inquiries to form a geographic location of a person using solely the contents of their posts, or tweets.

Language as a Geographical Indicator

A person's language can be very telling of where they were raised, or where they live. Popular culture has many examples of this, most famously being the character of Professor Henry Higgins in George Bernard Shaw's play "Pygmalion". In the play, Higgins is able to deduce the origins of a speaker "within six miles" by phonetics, a branch of linguistics (Shaw, 1912). Another example, a person from Georgia or other southern states may use the word 'pie' and it always means some form of baked fruit pie, while for someone in northern cities, like New York City and Chicago, a 'pie' may refer to a pizza pie. Subtle language cues can tell a lot about a person. With a group of data from different cities, it is possible to analyze each city and consider a person and how their language can help identify where they are located. Another example, looking at the Chevrolet Corvette, a sports car that is sold all across the United States; The word Corvette, as a single entry would not be an identifier, but it is possible that there is more talk about Corvettes around the Bowling Green, Kentucky area, because this is where the Corvette Plant and the National Corvette Museum are located. So a multitude of identifier words could center our location and help narrow down the results.

Authentication of Other Digital Identities

The Internet phenomenon "Catfishing", where an Internet user creates a fake personal profile on social media sites, using someone else's pictures, biographical information, etc. These "catfish" pretend to be a more 'appealing' identity than their true self to potentially meet a person or persons to fall in love

with in online dating. The term “catfish” came from an American documentary film, *Catfish*, about a young photographer named Nev Schulman, who befriends an eight year old artistic prodigy, Abby, online who painted a picture of one of his photographs. Nev’s friendship with Abby also led to a friendship with the entire family from Michigan and soon begins an online relationship with Abby’s 19-year-old half-sister Megan. When Nev began to do a bit of an online background check on Megan and Abby he found bits of information to be either partially or entirely fabricated, and they were not who they were portraying. The girl’s mother, Angela had created fake profile accounts to create a fake identity of her daughter and pursued a fake online relationship with Nev, who believed it all to be real (Henry Joost, 2010). In the film Angela’s husband Vince explains an anecdote about how when codfish were shipped from Alaska to China they would put catfish in with the cod to keep them agile, so that their flesh would not get mushy. He went on to say that “there are those people who are catfish in real life...those that keep you guessing, they keep you thinking, they keep you fresh.” (Zimmer, 2013)

Another example of ‘Catfish’ is that of Notre Dame Football star and linebacker, Manti Te’o, who was in a serious online relationship with, Lennay Kekua, a girl he met online. A few months into their relationship, Kekua supposedly developed leukemia and died. Distraught, Te’o used his pain from losing the girl he loved as inspiration to put on an amazing performance in the latter part of the college football season. But the whole relationship was a hoax, created by Ronaiah Tuiasosopo, a high school classmate of the girl whose

identity he stole. Tuiasosopo created Kekua's identity and fake profiles from the ground up and pursued an online relationship with Te'o and faked Kekua's death to cover up his lies and deceit. When the story of the hoax broke, because of all of the publicity and inspiration that had arisen from the story, Te'o's credibility as an elite football player had been questioned and suffered as a result.

Catfishing is one example of what this Twitter geolocation can help solve. If an online user has created a fake profile of a person from a completely different area it would be possible to determine if he/she is being truthful about their location.

PREVIOUS RESEARCH

In Dr. Lance W. Hahn's paper *The End is Near: the Statistical Regularities of Sentence Endings*, (Hahn, 2011) a dataset was used that was collected by Google in January 2006 that, sampled over 95 billion sentences on public English-language websites. This dataset was not used during this thesis due to the lack of location information within the data set. The dataset was organized by n-gram length from 1 to 5 tokens (e.g. words) with each n-gram length having a separate dataset. Using this dataset, Hahn was able to identify short-range statistical regularities in English word usage that occur at the end of a sentence. If there are statistical regularities telling us when a sentence is about to end, there could be some form of regularity tying a set of words to a specific geographic region.

In 2010, Dr. Zhiuan Cheng of Texas A&M University wrote a paper on geo-locating Twitter users. In his paper, *You Are Where You Tweet: A Content-*

Based Approach to Geo-locating Twitter Users, Cheng states that a large majority of Twitter users' location information is absent from their profiles and to overcome the lack of location information they proposed and researched a framework for estimating a Twitter user's city-level location based on the content of the user's tweets. In their research they found that they could place 51% of Twitter users within 100 miles of their actual location (Cheng, Caverlee, & Lee, 2010).

Present Work

The purpose of this thesis is to, using a simple search and collection of data, determine a geographic location of an individual Twitter user using only the content of their tweets.

TWITTER

Social networking is an online platform to build relationships that may share similar ideas, hobbies, backgrounds, etc. A social network service allows users to interact with these people of similar interests through e-mail or instant messaging. One website that uses a unique way of interaction is Twitter. Developed in 2006, by Jack Dorsey and Evan Williams, Twitter is a social media website that allows users to express their feelings, ideas, opinions, news, daily activities, etc. through a 140 character or less short message, also known as a 'tweet'. (MacArthur, 2013) Why only 140 characters? The reason for the limitation is because Twitter was originally designed as an SMS (short message service) mobile phone-based platform. 140 characters is the limit that mobile carriers

enforced on SMS messages. When Twitter expanded to a web-based platform, the constraint remained. These messages can either be viewed by the whole of the Twitter population or to only a select few of users, known as 'followers', that have subscribed to private Twitter users.

Efficacy of Using Twitter

While Facebook is a much larger network, currently standing at 1.31 billion users, (Statistic Brain, 2014) Twitter is more user friendly than other social networking sites when it comes to allowing access to users that are not part of your core friends. While it is still possible to set a user's profile to private on Twitter, not as many users choose to do so. Because the tweets on Twitter seem to have much less personal information included in them users feel more inclined to keep their profiles open to the public, while sites like Facebook delve much deeper into our personal lives. It is for the above reasons that we chose to analyze Twitter over Facebook.

Thesis Overview

In this work, we will examine how Twitter can be used to perform the geolocation. As part of this research, we have developed an alternative method to collecting and analyzing tweets which we will describe in Chapter 2. In Chapter 3, we will discuss our result and compare them to those of Cheng's (Cheng, Caverlee, & Lee, 2010). In Chapter 4, we will summarize our work and suggest future work based on this research.

CHAPTER II

RELATED WORK

The rise in popularity of Twitter over the past 5 years has attracted the attention of multiple researchers. In 2010, Dr. Zhiyuan Cheng, conducted research on the topic at hand (Cheng, Caverlee, & Lee, 2010). Cheng concentrated on three key features: i) to rely solely on tweet content omitting IP information, log-in information or external knowledge bases; ii) a classification component to automatically identify words with a strong local geo-scope within tweets; iii) a lattice-based smoothing model to refine a user's location estimate.

During Cheng's extensive data collection two complementary crawling strategies were implemented to help avoid bias: to crawl through Twitter's public timeline API to collect users and tweets, and using a breadth-first search to crawl through the social edges of each user's friends (following) and followers.

Using the open-source library twitter4j (Twitter4j open-source library, 2013), Cheng collected an unfiltered dataset of 1,074,375 Twitter users consisting of 29,479,600 status updates. Filtering out the users without a user submitted specified location, Cheng finds that 72.05% have a non-empty location. Since a user can manually input a location into their profile, there are an abundance of nonsensical location entries (e.g. Wonderland or the Moon). Taking these profiles into account, Cheng further filtered out all profiles that did not include a city name and found 223,418 user profiles (21%) list a city name and only 61,355 (5%) list a location as a latitude/longitude coordinate. In the final filtering process, Cheng narrowed the dataset down to profiles that only included

a valid city label (e.g. “cityName”, “cityName, stateName” and “cityName, stateAbbreviation”), accounting for 130,689 users and 4,124,960 tweets (12% of all sampled users). Cheng’s final sample of Twitter users will represent the population of the United States.

To assist with the building of their models and reduce bias, Cheng created a test dataset separate from their initial dataset. They collected a set of active users with over 1,000 tweets who had their exact location listed as a latitude/longitude coordinate and arrived at 5,190 test users and more than 5 million tweets.

Using this test data they formulated an algorithm to parse the tweets and found 25,987 per-city word distributions from a base set of 481,209 distinct words that occurred more than 50 times within the database. Using this baseline method, they were only able to identify 10.12% of users within the test set within 100 miles of their actual location, with an average error distance of 1,773 miles.

To improve upon this result they began to look at two problems: i) that most words are distributed evenly throughout the population across different cities and ii) most cities have a sparse set of words within their tweets creating an inaccurate word distribution for the cities and creating large estimation errors.

To manage the first challenge, Cheng analyzed all of the “local” words that were generated by the algorithm. By examining all of these words from the estimator they found that the performance was lacking because of the presence of noise words that would not necessarily portray a sense of location. To filter

those words out Cheng used what is called Local Filtering. Using a model of spatial variation for analyzing the geographic distribution of terms in search engine query logs created by Backstrom et al. (Backstrom, Kleinberg, Kumar, & Novak, 2008), Cheng was able to filter out the insignificant noise words. The main focus of the filtering is that each word would have a center geographic focus on the map and that as the distance from the center increases the frequency will decrease. A method was then used to divide the map into lattices and for the center of each lattice they searched for the optimized central frequency for the word. They then zoomed into the lattice and performed the search again. After repeating the process a number of times they were able to find the optimal center point and the most likely geographic center for the given word.

To overcome the deficiency of words across their sampled Twitter dataset, Cheng et al tried a number of smoothing methods for the probability distributions. They found the most effective method to be a lattice-based neighborhood smoothing method, with the map of the continental United States first divided into lattices of 1 x 1 square degrees. The probability per-lattice can be calculated with the neighbor lattices being taken into account.

Using a combination of both processes in the original algorithm, Cheng was able to improve upon the original baseline findings of a 10.12% success rate to a 51% success rate of finding a user's estimated location within 100 miles of their actual location. They also improved upon the average error distance from 1,773 miles to 539 miles.

DATA COLLECTION

Twitter's Search Engine

Twitter's massive collection of tweets since its creation is a goldmine of potentially useful information. Using Twitter's search engine, you can sort through that minefield and filter down to a specific topic, person, or idea. For example, to find people that have tweeted about "President Obama", insert the keyword into the search bar, "President Obama" or "Obama" and the result is a list of recently updated tweets from users about the Commander in Chief.

Specific topics can also be queried within Twitter's search engine. Twitter uses a method of categorizing tweets by topic by having users include a hashtag character (#) in front of the topic. For example, if we wanted to search for anyone that has tweeted anything about the 2014 Winter Olympics, in the search bar we would insert #WinterOlympics and any tweet that has included that particular hashtag will be shown.

More detailed operators are available that can be used within the search system to further filter out undesired tweets. Below is a table showing a few of the possible operators that are available within the Twitter search engine.

Table 2. List of Twitter Search Operators

| Operator | Finds tweets... |
|-----------------------------------|---|
| twitter search | containing both "twitter" and "search". This is the default operator. |
| "happy hour" | containing the exact phrase "happy hour". |
| love OR hate | containing either "love" or "hate" (or both). |
| beer -root | containing "beer" but not "root". |
| #haiku | containing the hashtag "haiku". |
| from:alexiskold | sent from person "alexiskold". |
| to:techcrunch | sent to person "techcrunch". |
| @mashable | referencing person "mashable". |
| "happy hour" near:"san francisco" | containing the exact phrase "happy hour" and sent near "san francisco". |
| near:NYC within:15mi | sent within 15 miles of "NYC". |
| superhero since:2010-12-27 | containing "superhero" and sent since date "2010-12-27" (year-month-day). |
| ftw until:2010-12-27 | containing "ftw" and sent up to date "2010-12-27". |
| movie -scary :) | containing "movie", but not "scary", and with a positive attitude. |
| flight :(| containing "flight" and with a negative attitude. |
| traffic ? | containing "traffic" and asking a question. |
| hilarious filter:links | containing "hilarious" and linking to URLs. |
| news source:twitterfeed | containing "news" and entered via TwitterFeed |

Before collection was started, the Institutional Review Board of Western Kentucky University approved the collection that all collection would maintain anonymity of Twitter users and that all usernames were scrubbed from the datasets during the collection by only collecting the message portion of the tweet. A local cyber security firm, EWA Government Systems, Inc., assisted in the collection of data from the Twitter search engine. We gathered the data by forming an advanced Twitter search for 'near:*city* within:*distance*' where *city* is the city, state format and *distance* is the distance in miles away from the geographic center of the *city*, which was 50 miles in this study, for 50 different cities across the United States as illustrated in Figure 1 below. A gathering script was written and is set to identify the more tweets button at the end of the search page, and loops through 200 pages of results for each city. The parsing script then takes the tweet message portion only and saves those to another file for analysis. Usernames have been omitted to preserve anonymity. Performing the search in the manner allows for the filtering out of users locations. Twitter users are allowed to enter a location of their choice when creating a profile. Users can enter their city and state (e.g., Bowling Green, KY), just their state (e.g., Kentucky), or attach a nonsensical location to their profile (e.g., the Moon, Gotham City, etc.) or leave it blank. By collecting the data using Twitter's location search query we can filter out the users despite what information they put in the location bar of their profile. Twitter does have an option to use a geolocation within the tweets themselves which would give a latitude/longitude location of

where the tweet was sent out, however as of June 2012 only 0.77% of users take advantage of this feature. (Lunden, 2012)



Figure 1: Map of US with 50 different cities highlighted

Data Set

After the collection of data through the week of September 1st – 9th, 2013 we had amassed a database of tweets from 50 different cities. Accounting for around 3,700 tweets per city and 185,000 tweets overall.

Sorting Data Set into Word Occurrences

With the help of Dr. Hahn, a script was written that would allow the raw data set to be sorted through and a frequency count would be performed on each word. Written in Perl, the script would run through each city's dataset and line by

line it would count the tokens (words) separated by a space character. With this script it was possible to also collect any number of *n-gram* word sets (combinations of *n* number of words). When the script reaches the last line of the file, a new Excel file would be made that would list the individual word and its frequency count. In Figure 2 below is a flowchart that explains the steps taken by the script to arrive at the results.

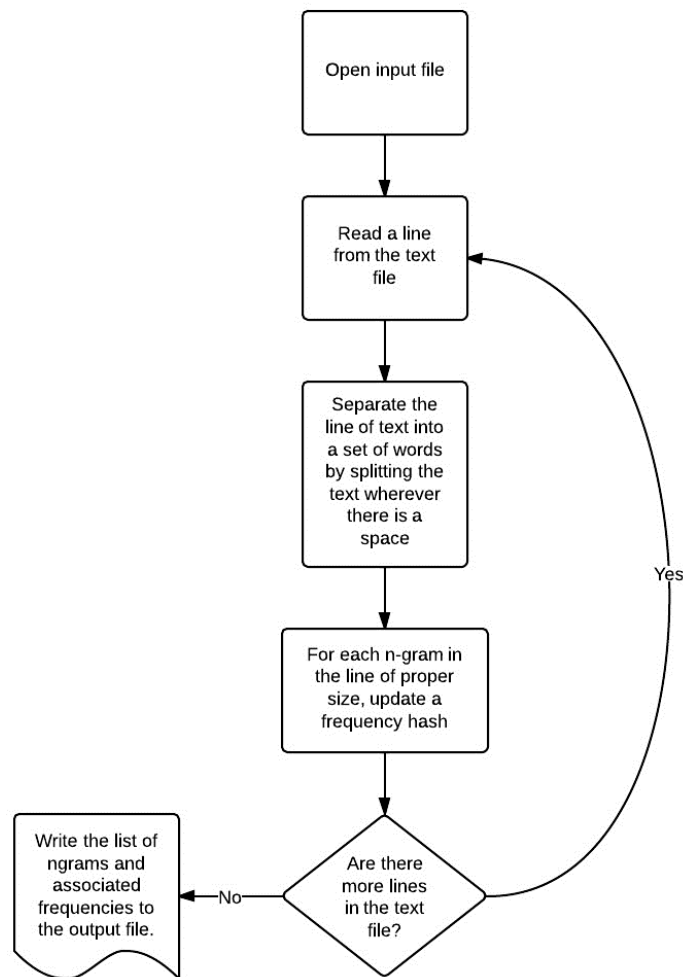


Figure 2: Text File to N-gram and Frequency Flow Chart

After the initial collection of the individual word frequencies with a simple change in the Perl code we were able to collect two more data sets of 2-gram and 3-gram word sets.

Filtering the Unique Words

A method was then needed to compare the different frequency files for each city we had obtained to try and discover any unique identifier words that would give us evidences to the location of the tweet illustrated by flowchart in Figure 3 below. A majority of the tokens that were counted had very low frequencies within the cities database. For example 91.8% of the tokens (words) within Atlanta's database occurred less than three times. To avoid the inconsistency we performed two tests using only the words with more than 50 occurrences and on words with more than 20 occurrences within each city's databases. With the help of Dr. Womble, we created an Excel Macro that, when executed, would loop through said file to identify any unique words within the different cities.

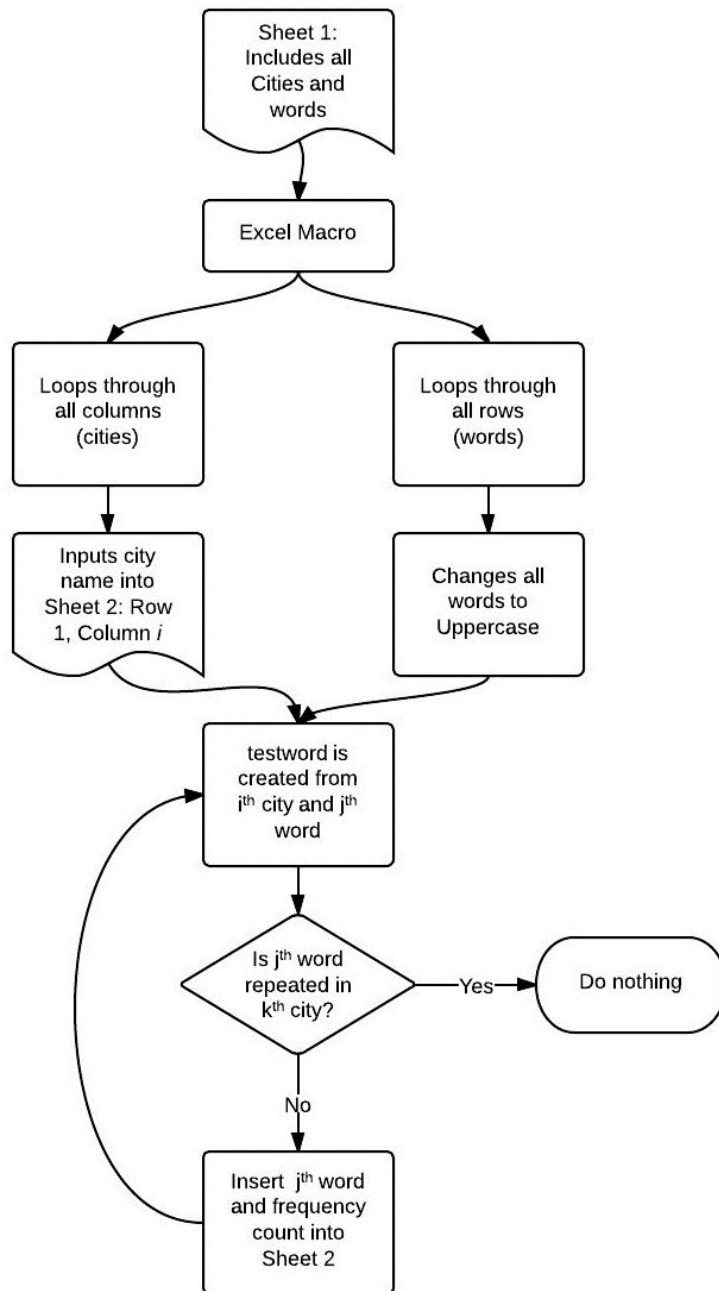


Figure 3: Flowchart of Excel Macro: Finding Unique Words among 50 Cities

CHAPTER III

RESULTS

Using the initial unique word filtering method described in Chapter II, inconclusive results were determined. Out of the 50 cities that were collected, only 11 had at least one unique word that occurs at least 50 times within the entire collected dataset and only 0.39% of the words tested were unique. In Table 3 below are the words, in relation to Homeland security that stood out the most when reviewing the analysis. Figure 4 below shows the cities that had unique entries in green and those that had no unique entries are in red.

Table 3: Results for Words with at least 50 occurrences

| CITY | UNIQUE WORD (OCCURANCES) |
|--------------------|---------------------------------|
| CHARLESTON | SYRIA (51) |
| LOS ANGELES | GROUP (51) |
| MIAMI | FINGERPRINT (74) |
| OMAHA | LIVE (78) EVENT (56) |
| ORLANDO | CUSTODY (66) GUN (50) |



Figure 4: Map of Cities with Unique entries (green): At least 50 occurrences

There happened to be even fewer unique words when we expanded the analysis and queried the words with at least 20 occurrences within the entire data set. Only 5 cities had at least one unique word within this analysis and only 0.07% of the words were unique.

Table 4: Results for Words with at least 20 occurrences

| CITY | UNIQUE WORD (OCCURANCES) |
|-------------|--------------------------|
| LOS ANGELES | GROUP (51) |
| MIAMI | FINGERPRINT (74) |
| MONTGOMERY | RATE (48) |
| | JOBS (47) |
| ORLANDO | CUSTODY (66) |
| | GUN (50) |



Figure 5: Map of Cities with Unique entries (green): At least 20 occurrences

The last set of data that was analyzed was words that occurred at least twice within the entire data set. This set was much larger than the previous two by a significant margin. We found that 2.66% of the 197,453 words that occurred

more than twice all 50 cities were unique words. There were a total of 5,264 unique occurrences at an average of 107.43 unique words per city.

When these different analyses are compared it is determined that this method did not provide significant results. There are a multitude of reasons why the method did not work.

- (i) The possibility for statistical error is too high.
- (ii) There was not enough data collected to provide a significant result.
- (iii) Data collection did not occur over a long enough period of time.
- (iv) There were too many nonsensical entries that skewed the data and interfered with the results.

Looking at Poisson distribution, the standard error is the square root of the number of counts. $Error = \sqrt{N}$ So, for example, our results in the analysis of the data set with at least 50 occurrences, we had 16 occurrences of unique words. This leaves us with an error of 4. Because our results were so minimal, we calculated a very high error. In order to obtain significant results, our error needs to be considerably lower.

In review, comparing the total amount of data collected and analyzed by this study to Cheng (Cheng, Caverlee, & Lee, 2010), the amount of analyzed data is quite different. Over the course of this thesis research, a data set of around 3,600 tweets per city and 180,000 tweets in total had been collected. During Cheng's research over 1,000,000 user profiles with over 29,400,000 tweets were collected. This large difference in the number of tweets tested

created a massive impact on our individual results. With such a small amount of unique words per city with a large enough hit count finding any words that are unique to a specific city would prove difficult using this method. To keep within Institutional Review Board standards we preserved anonymity during our data collection, but this process actually hindered our results. While Cheng was able to collect, on average, 30 tweets per user profile, our method of collection only amassed a large collection of individual tweets. We were unable to determine how many profiles these tweets came from because of the scrubbing of the user names during the collection process. Because Cheng was able to collect a set of data from particular users, he was able to determine how a user tweets on a regular basis.

Another potential problem with the results is that the collection of data did not occur over a very long period of time. Collection of the data set in this research took place over a ten day span at the beginning of September 2013. In this case, data was only collected for a short period of time at one point in the year. Trends and Twitter topics change constantly as fads come and go throughout the year. What is talked about most in March of 2014 may not be discussed 3 months later. Results could very well be improved by collecting more sets of data at staggered points in the year or by collecting steadily throughout the year.

When the data was analyzed there were a lot of words that came up as nonsensical or 'gibberish'. The data set with two or more occurrences had many more nonsensical unique words than the other data sets. For example, the top

unique word among all 50 cities was 'caw' which occurred only 33 times; this is because one user tweeted the word repeatedly 33 times, making the entry a false positive. Other examples include a lot of URLs that linked to YouTube videos or image sharing websites.

MULTIPLE WORD ANALYSIS

After failing to come to any conclusions with the initial data set of single unique word occurrences, an analysis of 2-gram and 3-gram word sets was conducted.

In the 2-gram analysis, there were a total of 158,305 word combinations and 15,402 (9.73%) of those combinations were unique to a city. But, once again a majority of these entries were nonsensical unless one was able to know the context in which they were used. The 3-gram analysis, had a total of 1.95 million word combinations and 20,048 (1.02%) of the combinations were unique.

Chapter IV

CONCLUSION

In this thesis it is shown that a streamlined approach to geolocating a Twitter user is not the most effective manner of obtaining significant results. The original data collection of about 3,700 tweets per city. That data was then sorted into word occurrences and n-gram occurrences using a Perl script. The data was then furthermore deduced to unique words per city using Excel macros. After obtaining the unique words in each city I manually looked over each set of data and was unable to deduce any words with locational significance as they were skewed by nonsensical data.

Cheng was able to deduce a user's location to within 100 miles of a city with 51% accuracy (Cheng, Caverlee, & Lee, 2010); numbers that were unobtainable during this research due to the lack of a sufficient dataset. The data collected using this method was only 0.6% of the total number of tweets that Cheng collected. In order to duplicate or improve upon these results, the data set must be equal in size to that of Cheng's research.

Future research should include a more extensive collection of tweets, however a problem that occurred during research was using the Microsoft Excel macros to perform our analysis. Although large data sets were being worked with, the macros required a large period of time to complete each analysis and were very fragile in the process while performing on my own personal computer. On multiple occasions either a wrong button was pushed or something to that

effect caused the macro to halt its process demanding a restart. In the future, I believe finding a more effective and sturdy approach to analyzing the data would be beneficial to the next researcher. One possible solution would be to use high performance computing (HPC) or a supercomputer. This would decrease the amount of time needed to analyze these datasets and any other larger data sets that are collected in the future. HPC would also help with the fragileness of the analysis method as there would be no other processes that would hinder it.

To improve upon the 51% accuracy within 100 miles that Cheng established using his own method, there will need to be a more widespread collection of tweets should extend through all parts of the year to cover changing trends and seasonal consistencies. Additionally, to improve results and improve the analysis, more frequency data should be analyzed. Time constraints limited only three sets of data to be analyzed at the end; words with more than one occurrence, more than 20 occurrences, and more than 50 occurrences. In order to see any trends and find an optimal occurrence analysis, more occurrence data sets should be analyzed. Looking at the average unique words per city, the results showed that 0.33 words per city occurred in the 50+ occurrences, 0.167 words per city in the 20+ occurrences, and 107.43 words per city with more than one occurrence. Since the results did not occur in a linear formation there will have to be a peak occurrence point. Exhausting all of the analysis techniques used in this thesis would not be ideal because of time constraints and the tediousness of the methods.

One area within the government in which a method such as this would greatly improve would be in Law Enforcement. New social media surveillance software can provide real-time monitoring of on-line communities. A company out of Virginia, ECM Universe, has developed a new software, Rapid Content Analytics (RCA) for Law Enforcement. RCA for Law Enforcement can monitor for credible threats to a community, fraud, bullying, employee behavior and suicidal tendencies. However, in ECMU's research about 20 to 30 percent of social media users enable geolocation and they are unable to dispatch officers to real-time street locations where threats may be originating from (Kanable, 2014). With a refined method, this could be more plausible and could benefit those that are using a software such as this.

Another interesting idea is that a process, like the one used in this thesis, could be used for more than just geolocation purposes. Communities tend to grasp onto news stories or current issues to fuel the topics for their tweets. A data set of this size could possibly reveal some trends within certain communities on how a particular story spreads. For example, a data set collected during election time could indicate the political views of a particular community and what might direction they are intending to vote.

Using other Online Social Networks in conjunction with Twitter could be possible in order to expand the data throughout the social network. However, with the increase in social media privacy throughout all OSNs this could prove to be difficult. Facebook has implemented new privacy policies which make it simple for a user to block any other user that is not included in their friends list

and given permission to view their posts. Other OSNs could also prove to be more difficult to collect from due to their ever changing formats. The reason Twitter was chosen for this particular thesis is because of the simplicity of the format and the ease of collecting a mass amount of data.

Could this method of Twitter analysis be beneficial toward the Department of Homeland Security? Only future research can answer this. If the correct research and analysis is performed it could very well help with the detection and pinpointing the location of Twitter users to within a specified range.

REFERENCES

- Backstrom, L., Kleinberg, J., Kumar, R., & Novak, J. (2008). Spatial Variation in Search Engine Queries.
- Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. *19th ACM International Conference on Information and Knowledge Management*, (p. 10). Toronto.
- CIA. (2010, October 21). *INTelligence: Human Intelligence*. Retrieved February 7, 2014, from Central Intelligence Agency: <https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/intelligence-human-intelligence.html>
- CIA. (2010, July 23). *INTelligence: Open Source Intelligence*. Retrieved February 7, 2014, from Cental Intelligence Agency: <https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/open-source-intelligence.html>
- Delgado, M. (2013). *Types of Online Social Networks*. Retrieved from Online Brand Manager: <http://onlinebrandmanager.org/social-media/social-network-types/>
- eMarketer. (2013, June 18). *Social Networking Reaches Nearly One in Four Around the World*. Retrieved from eMarketer: The data is from doing a twitter search for 'near:ciy within:50mi' where city is the city, state format from your list. The gathering script is set to identify the more tweets button at the end of the page, and loop through 100 pages of results for a cit
- Hahn, D. L. (2011). The End is Near: the Statistical Regularities of Sentence Endings.
- Henry Joost, A. S. (Director). (2010). *Catfish* [Motion Picture].
- Kanable, R. (n.d.). *New Social Media Surveillance Tools*. Retrieved from Blue Sheepdog: <http://www.bluesheepdog.com/social-media-surveillance-tools/>
- Lunden, I. (2012, July 30). *Analyst: Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City*. Retrieved from TechCrunch: <http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/>
- MacArthur, A. (2013). *The Real History of Twitter, In Brief*. Retrieved from About.com: <http://twitter.about.com/od/Twitter-Basics/a/The-Real-History-Of-Twitter-In-Brief.htm>

- Shaw, G. B. (1912). *The Project Gutenberg EBook of Pygmalion, by George Bernard Shaw*. Retrieved February 2014, from Project Gutenberg:
<http://www.gutenberg.org/files/3825/3825-h/3825-h.htm>
- Smith, C. (2013, May 26). *How Many People Use the Top Social Media, Apps & Services*. Retrieved from Digital Market Ramblings:
<http://expandedramblings.com/index.php/social-media-user-stat-infographic/>
- Statistic Brain. (2014, 1 1). *Facebook Statistics*. Retrieved from Statistic Brain:
<http://www.statisticbrain.com/facebook-statistics/>
- Twitter4j open-source library*. (2013, February). Retrieved from
<http://twitter4j.org/en/index.html>
- Wilson, M. (2008). *The comparison of online social networks in terms of structure and evolution*. Ann Arbor, United States: UMI Dissertations Publishing. Retrieved from <http://books.google.com/books?id=QBex-J7yvuQC&pg=PA13&lpg=PA13&dq=online+social+networks+traditional+social+networks+media&source=bl&ots=fKZA5meWDC&sig=BmeofOKqonFdLF-W62f9PuyrdX4&hl=en&sa=X&ei=50doUt2PM5fk4AOP0YCQBg&ved=0CF0Q6AEwBA#v=onepage&q=online%20>
- Zimmer, B. (2013, January 27). *Catfish: How Manti Te'o's imaginary romance got its name*. Retrieved from Boston Globe:
<http://www.bostonglobe.com/ideas/2013/01/27/catfish-how-manti-imaginary-romance-got-its-name/inqu9zV8RQ7j19BRGQkH7H/story.html>