GEOLOCATION THROUGH LANGUAGE RECOGNITION

Will Potter

Adviser: Professor David Kauchak

A Thesis

Presented to the Faculty of the Computer Science Department

of Middlebury College

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Arts

May 2014

ABSTRACT

With an increasing amount of text being shared on the web, through blogs, social media, websites pictures, it is becoming increasingly more difficult to translate the text in these mediums into geographic coordinates and physical locations. While GPS-enabled devices are becoming more popular, many people with mobile phones would prefer to not share their locations with applications and companies. Additionally, IP geolocation lacks the precision that GPS-enabled devices have. Yet, while Internet users don't explicitly share their GPS-location, they often will share information about their location in the form of textual status updates. Using geotagged tweets and other geotagged information, it should be possible to identify similarities between non-geotagged texts and classify someone's location by the words included in their tweet. With this information, more intelligence can be gathered about people tweeting, even if they haven't included their specific geographic coordinates with the tweet.

This thesis will focus specifically on classifying text to a variety of regions, including countries, states, counties and towns. It will use a variety of supervised learning classification techniques including SVM's and Naive Bayes. While the classifier is importance, the study will also focus on feature preprocessing, as that will most likely have a great impact on the results of the final product.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Classification of Social Media

The rapid growth of social media in the past 10 years has shown that people are living their lives increasingly more through the internet and are sharing more information with more people than ever before in history. At the end of 2013, Facebook had 1.23 billion monthly active users [2] and Twitter had 241 million monthly active users [1] – and those are just the largest two networks. People are sharing content online through Pinterest, Foursquare, Tumblr and many other sites. Yet while the number of social media users has expanded, much about their behavior is still difficult to determine.

As people share information, companies analyze user information to help data monetization efforts, such as advertising. For example, if 95% of a companys users were under the age of 30, advertising for retirement communities wouldn't be an efficient use of advertising space. They also review user behavior to improve their product and determine where efforts should be expanded. If only 3% of Facebook users were to use Facebook chat, spending effort elsewhere would be prudent.

While this principle seems similar, analyzing this data becomes increasingly more complex as the size of the data increases – Twitter, alone, manages 500 million tweets per day. Deriving useful conclusions from such a large dataset requires an intricate knowledge of the data before effective analysis can be applied.

1.2 Smartphones and Geotagging

In addition to the growth of social media, the world has seen the adoption of the Internet connected smartphone. Business Insider found that 22% of the global population owned

a smartphone at the end of 2013. Quite simply, people are sharing more and doing so wherever they go. As someone uses their phone, they also have the ability to share their phone's GPS location and attach it to their social media updates. This premise gave way to the creation of Foursquare, a geographic social network that allows users to share their location with other users and local businesses for perks. As Foursquare refined their idea, Facebook and Twitter soon followed with location-sharing services inside their mobile apps. Even though these features are now a part of most social media apps, adoption has not be excellent. Estimates show that only 1 in 5 tweets are geotagged, which still beats Facebooks poor geographic integration (Need Better Citation).

Geotagging is particularly useful for advertising, as well as business intelligence. Twitter, while an incomplete sample of the population, can serve as a tracking agent for businesses. While the requirement that people "opt-in" to sharing their location may be a bias in the data, companies still may use geotagged tweets as a random sample of all social media users when running analytics to generalize conclusions across an entire user base. For example, McDonald's might like to show an advertisement for someone who works immediately next to one, but has never tweeted or checked-in at a their restaurant. A user's proximity to a certain business makes them a more appealing ad target, which in turn provides a premium experience for advertisers to use. A McDonald's or Target would greatly prefer to target people who live nearby or pass through town rather than simply targeting based on a user's interests, tweets or follows.

Given that Twitter only records a definitive location for 20% of their tweets, they lose out on location information for the rest of the almost 80% of their users. If a social network were to mandate location sharing, the company would be perceived as grossly violating individual privacy rights and would almost certainly lose many users, thus decreasing the underlying value of their service. Even if the locations were automatically collected and kept private, smartphone manufacturers, like Apple, could disable

GPS-use on their phones to protect users. If a social network had some way at guessing the location of a user based purely based on the content of their update (i.e. a tweet's body or a Facebook status' text), they would be able to provide a superior analytic experience without storing the exact geographic coordinates of a user. Additionally, the network would be using information that the user implicitly consents to public sharing (by nature of using a social network in the first place).

This thesis looks to examine the possibility of determining a user's location at the point of sharing (or close to the point of sharing) based on text classification and clustering around the content of their update. Specifically, it will use a set of geotagged tweets acquired from Twitter's public API as training data for running a series of machine learning algorithms on the text of tweets. By training on already geo-tagged tweets, which appear to have virtually the same content as non-geotagged tweets, anyone could later predict locations for the remaining 80% of tweets. This thesis will focus on English language tweets, as tagged by the Twitter API, but the methods could hopefully carry over to other languages as well.

While initial efforts target exclusively the body text of tweets, incorporating user information into classification should improve results. Accounting for the location of other tweets by the same user and the user's stated location in their settings should help influence the classification result especially if classification efforts narrowed it down to a small number of places. Additionally, using the time as a feature should help, as most people operate in regular cycles (daytime at a place of work or school, and nighttime at home).

As many words bear certain significance to a particular geographic location, it can be expected that different types of words will have an effect on predicting the location of a tweet. At the simplest level, referring to place names, such as New York, Boston, or San Francisco, can be expected to have a correlation to the location of the tweeter. It may

relate to something in the past ("Just got back from New York City...great weekend."), present ("I'm at New York Hilton Midtown - @hiltonhotels (New York, NY)"), future ("3 hours to go until New York will be calling! #fashion #opportunity #career"). The user may not even be planning to go to the place, but rather just is referencing the place ("If only New York wasn't so far away").

Additionally, the use of neighboorhoods or other place names may indicate a geographic location. Tweets like "I SNOW #nyc @ Hell's Kitchen - NYC" refer to the Hell's Kitchen neighboorhood in NYC but tweets refer to other cultural features, like the TV show "Hell's Kitchen": "Literally can't get enough of @GordonRamsay 'a Hell's Kitchen' - absolute stormer of a show, can't wait for the new series".

Finally, the presence of particular words, like "frappe" and "milkshake" may indicate if someone is in a particular location. Tweets with "frappe" are less popular and appear in the Northeast mainly, while "milkshake" appears more frequently and across a greater geographic area.

Using a fair degree of knowledge about the use cases of Twitter, we can hope to see some success in analyzing a tweet's textual body to infer the location of a user.

CHAPTER 2

RELATED WORK

Founded in 2007, Twitter is a relatively new platform. Like most parts of the high tech sector, it has changed and evolved rapidly, starting as a text messaging based app that ultimately transitioned to mobile and web application based network. As Twitter has grown, it has seen increased interest from academics for its representation of realtime human dialogue and movement. Kwak et al. [10] found that Twitter had transformed into a hybrid social network and news medium. Given the low user reciprocity around sharing as well as the high average retweet count for an average tweet, it resembles a crowd-sourced news network almost as much as a person-to-person network.

Building on the crowd sourced news network description, researchers created systems that record and plot geo-tagged tweets and photos on a map, to show a heatmap effect regarding Twitter activity at any given moment. [13] [20]

Kinsella et al. [9] started to consider the idea of geotagging tweets by comparing it with several other forms of geotagging. Using a baseline constructed from a tweet's GPS coordinates, they considered geotagging based on a user's self reported location as well as a tweets content. They then compared this to the majority label in a given dataset.

Twitter's ability to analyze realtime communication across the globe has allured researchers to consider various questions, especially where users are when they tweet. Despite Twitter's included GPS coordinates, others have tried to extract location information from a tweets content. Using a highly specific dataset consisiting of Twitter users with over 1000 tweets and a listed profile location in the continental United States, Cheng et al. [5] were able to achieve an accuracy rate of 51% when classifiying user's locations. Initially, they identified local words, such as "tortilla", "redsox" or "canyon". By combining the geographic centers of each of these words, they probabalistically de-

termined a new centroid for a likely area of the tweet in question.

Roller et al. [16] compared the text from tweets to definitive, fact-heavy text from Wikipedia downloaded. Their experiments could correctly geolocate the tweets with 161km of its true location 90% of the time. By using a k-dimensional tree, they recursively broke up the geographic grid to best identify which node fit the testing data. Training data included the wikipedia data containing many place names and toponyms. They then compared a tweet to the wikipedia text, knowing that placenames included in tweets would correlate to a particular document from wikipedia often.

Additionally, the tweet corpus has been analyzed to search for spatiotemporal anomalies, such as natural disasters or societal events, like riots or protests through clustering tweets across time. Their study produced relevant visualizations that would accurately overlay the type of disturbance on a map, based on the trending terms used in tweets. [18]

CHAPTER 3

DATA AND PREPROCESSING

3.1 Data

Data was collected from the Twitter API by querying their streaming API for all tweets with an attached pair of geographic coordinates or an attached geofenced region. Tweets with a region attached were assigned the midpoint of the region. A program, running on a personal computer would run for a period of time downloading new tweets during that period and storing them in a database.

The dataset is comprised of 1,656,146 tweets with 612,728 unique users tweeting within a period of February 13, 2014 and March 6th, 2014. This sample represents all geotagged, English-language tweets collected over random time intervals during the above period. Collection from the Twitter API occasionally times out, leaving gaps where tweets were not collected. As Twitter publishes approximately 500 million tweets per day, this dataset of 1.6 million represents a vast minority of tweets between the dates of collection, yet it still provides a significant number for running experiments.

In addition to Twitter data, Wikipedia articles were used to provide another training base. Given the large number of place specific terms

3.2 Preprocessing

Tweets tend to be informal and are filled with insignificant, popular words (“the”, “I”, “a”) as well as a range of misspellings. In addition, tweets contain links and usernames that don’t necessarily have a strong correlation to one’s location. For this reason, simply counting each word as a feature is a nave approach that introduces an unwanted amount of noise to classification experiments. Additionally, examining bigrams, or particular

phrases, as opposed to simply looking at 1-gram words is expected to increase the accuracy of classification. The bigram, “in Boston”, is more likely to imply the user is actually in Boston compared to “from Boston”.

3.3 Labeling

Tweets are labeled with geographic coordinates or regions pertaining to particular cities or areas. The geographic locations come from mobile phone GPS units or Twitter’s own location detection attempts that incorporate IP addresses as well as some textual features.

CHAPTER 4

EXPERIMENTAL METHODS

4.1 Posing the question as a classification problem

Taking the dataset of geotagged tweets, this study structures its experiments as a classification problem. The tweets in question are already geotagged, thus providing a definitive definition of where they were written. With this in mind, the goal is to predict the geographic coordinates of tweets without included geodata.

4.2 Scope

While the dataset only comprises English language tweets, tweets are saved from around the world. Given the lack of English speakers in many regions of the world, it could be problematic to label the entire world, as a region in Africa will have substantially fewer training examples than a region in Fairfield County, CT would have. Experiments will examine the continental US separately in addition to attempting worldwide classification.

4.3 Labels

This study examines several different methodologies for labeling tweets.

4.3.1 Geographic Grid Regions

In running experiments, labels are derived from the geographic coordinates of a tweet. As the labels run across as wide range of real numbers in 2 dimensions, effectively labeling the training features is a non-trivial process. The easiest way to label tweets is

to break up the geographic into regions of equal latitude and longitude intervals. This, however, ignores cultural boundaries and dense areas. A 1-degree x 1-degree "bucket" in Montana has a smaller and more homogenous population with fewer places. However, the "bucket" including New York City would have many more people, places and material and determining if a tweet originated there would be difficult. Given the inequality of area as latitudes change, this can be expected to behave differently for regions on the equator when compared with polar regions. While most of the world's population and tweets do not originate from the extreme poles, northern/southern regions will be broken into smaller buckets than central areas.

4.3.2 Political Boundaries

Using political boundaries or census designated places, as labels would be a slightly more intelligent way to break up the corpus of tweets, but still doesn't account for a large difference in population density. Breaking up a heavily populated state like Massachusetts or New Jersey into more buckets than Montana or Wyoming would give more authentic labels, due to the diversity and increased activity on social media. Additionally, latitude and longitude are often not the most effective ways of differentiating a population as cultural boundaries are often not straight lines.

To simplify experiments, tweets outside of the 50 United States were discarded and each tweet was applied

4.3.3 N-sized Example Buckets

Using a k-dimensional tree to break up tweets into n-sized buckets by their latitude and longitude, experiments will break the dataset into labels that more accurately represent population distribution across the planet. This method is similar to the work of Roller et

al. [16], except it will train the model on tweets rather than Wikipedia data.

4.4 Data Training and Testing Protocols

4.4.1 Randomness and Cross Validation

Due to performance implications, most experiments ran with 10000 examples. From the 1.6 million-tweet corpus, tweets are randomly selected for each experiment. All models then run on the same random, 10000-example sample to avoid variation that might stem from pulling drastically different sets from the complete corpus.

Additionally, to account for variation that might occur within 1 experiment, 10-fold cross validation is used rather than simply running testing the whole sample in one go. Within each fold, the data is split into 80% training and 20% testing.

4.5 Supervised Learning Models

4.5.1 K-Nearest Neighbor

K-Nearest Neighbor classification examines a set of training examples and using their features and internally stores them on an n-dimensional space. In our case, features are words or n-grams included in the body text of tweets. Then with a test example, the model identifies the K-nearest examples and chooses the majority label.

In order to mitigate the risk of insignificant terms weighting the distance metrics for each example, a term frequency-inverse document frequency (TF-IDF) vectorizer can be used instead of a normal count vector. TF-IDF vectors weight words that appear the most in a given document but negatively weight words that appear in many documents.

4.5.2 Naive Bayes

Using the Naive Bayes model for text classification introduces another method to improve accuracy. While an effective model, our data should be preprocessed to ignore many of the random noise words included in most tweets. With each ineffective term in the training set, the overall effectiveness of Naive Bayes decreases. When training on a toponym heavy dataset like Wikipedia, Naive Bayes can be quite effective.

CHAPTER 5

EXAMPLES/RESULTS

Still in progress

5.1 Examples

TODO: Find some interesting examples from testing

5.2 Results

5.2.1 KNN/Grid Labeling Results

As shown in the table below, K-Nearest Neighbor increases accuracy with an increase in clusters. With both labeling schemes, they don't appear to improve beyond K values of 128. K values of 128 appear to slightly dominate choosing exclusively the majority label.

K-Nearest Neighbor Test Data Results		
Labeling Grid:	2° x2°	10°x10°
Unique Labels:	623	139
Majority	0.0632124352332	0.168604651163
2 NN	0.0103626943005	0.094476744186
4 NN	0.0129533678756	**0.180717054264**
8 NN	0.0150259067358	0.101259689922
16 NN	0.0290155440415	0.140988372093
32 NN	0.0461139896373	**0.176356589147**
64 NN	0.059067357513	0.132751937984
128 NN	**0.0652849740933**	**0.171511627907**
256 NN	**0.0678756476684**	**0.184108527132**
512 NN	**0.0652849740933**	**0.179748062016**

5.2.2 State Boundary Results

5.2.3 K-dimensional Tree Results

5.2.4 Naive Bayes Results

5.2.5 Discussion on Preprocessing

CHAPTER 6

CONCLUSION

BIBLIOGRAPHY

[1] About twitter, inc. — about. https://about.twitter.com/company. (Visited on 03/20/2014).

[2] Company info — facebook newsroom. https://newsroom.fb.com/company-info/. (Visited on 03/20/2014).

[3] Smartphone and tablet penetration - business insider. http://www.businessinsider.com/smartphone-and-tablet-penetration-2013-10. (Visited on 03/20/2014).

[4] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[5] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[6] Clayton Fink, Christine D Piatko, James Mayfield, Tim Finin, and Justin Martineau. Geolocating blogs from their textual content. In *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pages 25–26, 2009.

[7] Mark Graham, Scott A. Hale, and Devin Gaffney. Where in the world are you? geolocation and language identification in twitter. *CoRR*, abs/1308.0683, 2013.

[8] Shinya Hiruta, Takuro Yonezawa, Marko Jurmu, and Hideyuki Tokuda. Detection, classification and visualization of place-triggered geotagged tweets. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 956–963. ACM, 2012.

[9] Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. I'm eating a sandwich in glasgow: modeling locations with tweets. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 61–68. ACM, 2011.

[10] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[11] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceedings of the 33rd international*

ACM SIGIR conference on Research and development in information retrieval, pages 435–442. ACM, 2010.

[12] Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), 2013.

[13] Yusuke Nakaji and Keiji Yanai. Visualization of real-world events with geotagged tweet photos. In *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*, pages 272–277. IEEE, 2012.

[14] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *FLAIRS Conference*, 2011.

[15] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[16] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510. Association for Computational Linguistics, 2012.

[17] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 18:1–18:10, New York, NY, USA, 2008. ACM.

[18] Dennis Thom, Harald Bosch, Steffen Koch, Michael Worner, and Thomas Ertl. Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 41–48. IEEE, 2012.

[19] Benjamin P Wing and Jason Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 955–964. Association for Computational Linguistics, 2011.

[20] Keiji Yanai. World seer: a realtime geo-tweet photo mapping system. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, page 65. ACM, 2012.