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

DATA AND PREPROCESSING

2.1 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.

2.2 Labels

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.

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.

2.3 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".

CHAPTER 3

METHODS OF CLASSIFICATION

CHAPTER 4

EXAMPLES/RESULTS

CHAPTER 5

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] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adap-

tive 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.

[12] 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.