EECS 395 - WEIR
Tweet Geo-Extraction

Kai Hayashi, Cary Lee, & Daniel Myers

Introduction

Twitter contains a vast recorded collection of human interaction. With its easy data scraping and open nature, it represents a nice collection of information for a researcher. Recently, Twitter added a feature allowing users to keep track of their location via their phone's GPS coordinates. We are interested in the following question: "Can someone determine a tweeter's location based off of the content of the tweet alone?"

Depending on what people use twitter for this task seems reasonable. If users tweet about their surroundings, then correlating the text to their location should be straightforward given enough tweets. Regional linguistic patterns may also play a role in determining location [4].

This question has been a hot topic in the news as of late, with reported robberies and many concerns over the public data available in tweets [3], [2]. The GPS information combined with Twitter's large user base, may provide enough data to build a distribution of user locations given the tweet text.

In this report we present our method of gathering Twitter data and provide some results of our basic attempt at solving this problem. We utilize WEKA and the Twitter API to collect and process tweets. We represent the tweets as a word vector and run a Naive Bayes classifier on this representation. We also present results from two other slices of this data: one tracking a user's location history, and one examining only the tweets emanating from New York and Los Angeles.

Data Collection

We collected the tweet data by scraping twitter once an hour for a week. We wrote a short scraping script that collects tweets from the public timeline, collects the friends of those users, and then grabs 200 most recent tweets from each user in this collection. In this report we only examine tweets that were produced with a geo-tagged location

provided by a GPS device. The geo-tags contain bounding-box information and a location string. We used this string to produce city, state/country pairs.

After running this scraper for a week we obtained just over 22,000 geo-tagged tweets from 626 different users representing 1474 different places.

Due to the restrictions of the twitter API (there is a limit of 350 queries per hour) we were prohibited from obtaining much more data. As will be discussed later there were also significant scaling issues with running the classifier on even a dataset with 22,000 tweets in it. In addition, our scraping algorithm may bias our data by not allowing for a diverse number of unique users.

Data Slices

In addition to examining our full dataset, we created two subsets from this data out of curiosity and to help improve our accuracy. The first dataset is a collapsed version of all of the data. The second examines the tweets from two common locations in our dataset, New York City and Los Angeles.

To build the first data slice we ran through a user's tweets and determined which city they most frequently tweeted from. We then concatenated all tweet text from that user into a single corpus identified with this location. This dataset contains corpora from 626 different users.

The second data slice was constructed by selecting all tweets from New York and all tweets from Los Angeles. Unlike the last dataset, we kept each tweet text separate. This dataset contained 433 different tweets from 21 different users. 314 of these tweets were from New York, while 119 were from Los Angeles.

Results

The results we present here were obtained using a similar method to one used by Brent Hecht, et al [1]. We combined our corpora of tweet text and corresponding locations as a list of pairs. We then converted the corpora into a word vector and built a model using the Naive Bayes classification algorithm. The results of 10-fold cross-validation are summarized in Table 1.

From our first data slice, we did not have great success yielding only 2.6 percent precision and 4.2 percent recall. From our NY vs LA slice we had roughly 74 percent precision and 76 percent recall.

The low accuracy of the User Location data can be explained by the lower number of unique users in our dataset. Hand inspection of our data reveals that most users tweet most often from a unique city in our dataset. The model will have a difficult time

Dataset	Precision (%)	Recall (%)
Full Dataset	30.9*	35.6*
User Location	2.6	4.2
NY vs LA	73.9	76.0

Table 1: Results from the three data slices. These are the values from the preliminary report run on 1/3 of the full dataset. We have been having some trouble getting the full dataset to run.

City	Words
NY	RT, MoMA, Thanks, exhibition, lol, shit, Ai
LA	LA, time, love, day, im, RT, bamboozle

Table 2: Indicative words for NY and LA.

discerning between cities as only one corpora represents that city. Thus, by concatenating all tweets, regardless of whether they emanated from the most popular location, together introduces significant noise into the corpora. For example, one user tweeted from both New York and Los Angeles. If he had talked about the Statue of Liberty when in NY, then our model would correlate the Statue of Liberty, incorrectly, with LA.

The excellent results from the NY vs LA data are not very indicative of the failures of this algorithm. The small user set makes it easy for the model to classify a tweet from LA just based off of a users particular tweeting style. Thus for this data to be meaningful, beyond being interesting for determining the words involved, we would need significantly more tweets from both of these cities.

Conclusions

There were a number of discoveries after analyzing the results of our experiment however the most apparent was that we needed more data. Despite scraping for nearly 2 million tweets, only slightly over 1 percent of those were geo-tagged. Furthermore, after creating subsets of the data, the number of tuples being examined was reduced into the hundreds yielding results that may or may not hold if the datasets were significantly larger. Lastly, the way in which we collected our tweets may have affected our results by getting users who were friends of other users, restricting our geographic diversity.

That being said, there were some observations suggesting that this method can be used and could easily be more successful with some minor refinements. For instance, Table 2 shows some of the indicative words for both New York and Los Angeles. While some words certainly do not seem like they are closely related to location (e.g. “lol”, “shit”, “Thanks”), others such as “MoMA” (The Museum of Modern Art in New York) are highly correlated to a specific place. Similarly, “LA” and “bamboozle” (a music festival in LA) were found to be indicative of Los Angeles.

The high precision and recall seen in the NY vs LA certainly is encouraging however the small size of this particular subset could call into question the validity of these results. Having so few tweets means that there was not a lot of noise present and given that these tweets all came from a mere 21 tweeters the text in our corpus is not likely to be overly indicative of the New York or Los Angeles vernacular.

Next Steps

The most import next step would be to rebuild the dataset. Rather than grabbing tweets from friends and rather than grabbing a larger number of tweets from a given user, it would be better from a diversity standpoint to just grab random tweets off of the twitter timeline. This would eliminate having any location clustering or biasing as well as having more user coverage.

To run and build a naive bayes classifier, WEKA required 15 GB of RAM. If this project were to continue we would need a lot more data and a way to run experiments on it. There are probably better approaches to performing the machine learning that should be considered.

Another interesting next step would be to try and use the bounding box associated with the geotagged tweets rather than the location names. After that, it would be possible to build a probability distribution across different regional granularities for a given tweet. This would make more sense for some tweets as a given tweet might contain fine granularity information like an address or instead might contain courser information like "y'all".

References

[1] Bongwon Suh Ed H. Chi Brent Hecht, Lichan Hong. Tweets from justin bieber's heart: The dynamics of the "location" field in user profiles. *CHI 2011*, May 7-12 2011.

[2] Danna Harman. How twitter is upending british privacy laws. *Chrisitan Science Monitor*, June 2011.

[3] WAGT Staff. Could twitter robbers get to you? *NBC*, 2009.

[4] Kyumin Lee Zhiyuan Cheng, James Caverlee. You are where you tweet: A content-based approach to geo-locating twitter users. *CIKM 2010*, October 26-30 2010.