EECS 395 - WEIR: Final Report
Twitter 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 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?"

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 [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. We used this location to produce city, state/country pairs.

After running this scrapper for a week we obtained just over 22,000 geo-tagged tweets from 621 different users. ****INsert number of cites here (maybe a table that summarizes this information?)***

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 to help improve our accuracy and out of general curiosity. 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 this first data slice we run through a user's tweets and determine which city they most frequently tweet from. We then concatenate all tweets from that user into a single corpus and run the naive bayes algorithm

Results

The results we present here were obtained using a similar method to one used by Brent Hecht, et al [1].

Conclusions

Next Steps

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