Web Extraction & Information Retrieval
Project Update

Kai Hayashi, Cary Lee, & Daniel Myers

Introduction:

Twitter contains a vast recorded collection of human interaction. With its easy data scraping and free and open nature, it represents a treasure trove of information for the well-equipped scientist. 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. 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 preliminary results of a very 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 Naive Bayes with this representation.

Data:

We collected that 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.

Limitations: [!] = cut if we are over 2 pgs

In the last year Twitter has imposed several restrictions on how a user can scrape Twitter data, and on how a user can protect themselves from being scraped. Due to these restrictions we are limited to making 350 queries per hour. This has limited

the number of tweets that we can pull. In addition, approximately 1% of users protect themselves from being scraped. This limits how many tweets we are able to collect.

Preliminary Results:

Next Steps: