Jacob Petersen

Josue Martinez-Montero

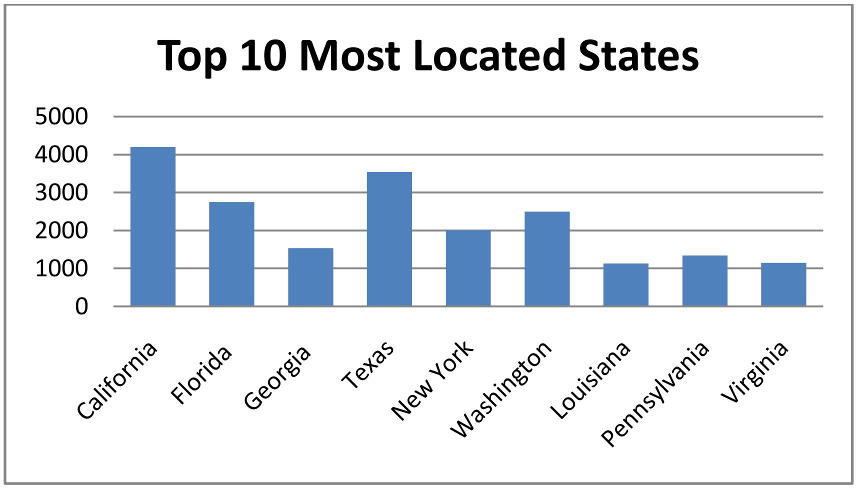
Wesley Weaver

Checkpoint 1

The data that we collected in the first phase of our project consisted of tweets pulled using Twitter’s streaming API. We streamed relevant tweets by filtering all of the tweets by the keywords “Obama”, “Barack Obama”, “Barack”, “President Obama”, and “President Barack Obama”. We also filtered the tweets by location so that the tweets we streamed were from users that were located within the continental United States. The original format of the data we received consisted of tweets that have many data fields that we are not interested in. In order to condense this data we used a variant of the python script, clean.py that Jeff provided, to remove all of the data fields that were not relevant. After manipulating the raw data and trimming it down the format of our tweets looks similar to this:

```
{"user":{"locat":false,"profile_image_url_https":"https:\/\/si0.twimg.com\/profile_images\/1674686129\/Profile_pic_normal.j
pg","id":25872911,"followers_count":1661,"friends_count":1727,"location":"","name":"laynie","lang":"en","favourites_count":
3,"screen_name":"laynier"},” text":". @edshow President Obama wins popular vote by close to 3 million
votes.","id":266352207509741568,"retweet_count":0}
```

We also plan on manipulating the data so that each user’s tweets are concatenated each day so that any spam or high-traffic users do not have a negative effect on our data analysis. The following graph shows the number of tweets that have a relevant and recognizable location for the top 10 states.


Top 10 Most Located States
California
Florida
Georgia
Texas
New York
Washington
Louisiana
Pennsylvania
Virginia