

Jonathan Candelaria
Jesse Gomez
Yuxuan (Leo) Li

SID: 861062229
SID: 861056174
SID: 861045931

NetID: jcand003
NetID: jgome026
NetID: yli066

CS179G: Bigdata Analysis Spring 2016

Project Phase 1 Report

Project Overview

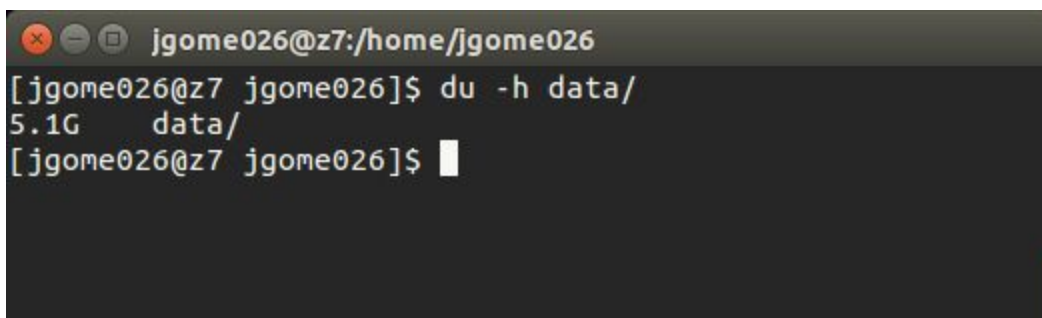
The goal of our project is to provide analytical information from Tweets from different areas of the United States. This information will include statistics related to various social interests such as entertainment, sports, politics, or other events. We will also be applying sentiment analysis to determine whether each Tweet is a positive or negative engagement with the topic and compare this data across the nation.

Data Collection

We are using the Twitter API and Tweepy library to collect the Tweets needed from each of the four regions of the continental United States. We divided it into these four regions to allow for better results of filtered Tweets from Twitter's streaming API, which uses bounding boxes based on two sets of coordinates to filter geo enabled Tweets. The python script creates files at 1000 Tweet increments to prevent data loss due to file errors, but later we will run another script that combines 100,000 Tweets into one file to make it easier for hadoop to handle this data. We downloaded about 25 million Tweets total, and transferred them to the servers in 1 million Tweet increments due to size constraints of the server. Each Tweet is stored in one line and includes the Tweet itself stripped of any new line characters that would break our storing structure, and information related to the Tweet such as the time created, username, and more. These columns are delimited by groups of semicolons as they will be unique enough for us to use as separators. Screenshots will be included below.

Member Contributions

- Jonathan: Ran script to collect data intermittently over the period of two weeks.
- Jesse: Wrote python script and ran script to collect data intermittently over the period of two weeks.
- Leo: Ran script to collect data intermittently over the period of two weeks.

A terminal window with a dark background and light text. The title bar shows the user 'jgome026' and the path '/home/jgome026'. The prompt is '[jgome026@z7 jgome026]\$. The command 'du -h data/' has been entered and executed. The output is '5.1G data/'. The prompt is now '[jgome026@z7 jgome026]\$' with a cursor.

```
jgome026@z7:/home/jgome026
[jgome026@z7 jgome026]$ du -h data/
5.1G    data/
[jgome026@z7 jgome026]$
```

Size on disk

```
jgome026@z7:/home/jgome026
[jgome026@z7 jgome026]$ cat ./data/24317000 | tail -3
24316997;;;2016-04-25 13:39:55;;;princesskaayb;;;146;;;en;;;period . https://t.co/ask2yB7PMF;;;None;;;Carson, CA;;;via Twi
for Android
24316998;;;2016-04-25 13:39:56;;;W_Angels_Wings;;;64137;;;en;;;@resultsneeded1 Thanks so much for the RT, and have a magnificent
ayi :);;;;None;;;KlrKland, WA;;;via Twitter Web Client
24316999;;;2016-04-25 13:39:57;;;alysaanavarro_;;;401;;;en;;;lol didn't bother to ask or give an explanation which goes to show
t;;;None;;;Riverside, CA;;;via Twitter for iPhone
[jgome026@z7 jgome026]$
```

Example