CS 410 Project Progress Report

William Shih

1) Which tasks have been completed?

I have completed the task of learning how to use the Twitter API and using NLTK for Sentiment Analysis. I learned how to use the "tweepy" package for Python for interaction with the Twitter API. I created a function to retrieve a number of tweets given a username or search query. Additional information (whether a retweet or reply or quote tweet, number of retweets, number of likes, number of replies, and timestamp) is also retrieved to aid in visualizing or separating the data. I used the "textblob", "vader", and "nltk" package for sentiment analysis. It turns out that for the VADER sentiment analyzer, cleaning the tweets had a minor effect on the sentiment analysis and almost no effect on the textblob sentiment analysis. I created a function to compare the distribution of sentiment of the query compared to a sample of 9999 English tweets using a two-sample t-test.

Learn how to use Twitter API: 5 hours

NLTK for Sentiment Analysis: >10 hours

Flask and d3 for Data Visualization >15 hours

2) Which tasks are pending?

I still need to use flask and d3 to do the data visualization, which I have not yet started. This will take significantly longer than the first half of the project given how many ways the data can be visualized. Also, I still need to write the documentation for the data visualization.

3) Are you facing any challenges?

I can only retrieve 500,000 tweets from Twitter in a given month, so it may be problematic to run a huge number of queries one after another. Also, only the most recent 3,200 tweets can be retrieved from any user without "academic" access which will not apply here. For a given query, only the most recent 7 days of tweets can be retrieved. This will make time series analysis almost useless. So only recent tweets can be analyzed for the data visualization. Also, I have to run a server or something to get my Python code up on the web, which may be challenging to get for free.