Project Progress Report: CS 410
Topic: Sentiment Analysis
Team Canyon

Which tasks have been completed?

Following Tasks have been completed:

- Data Streaming/Scrapping
- Data Cleaning and
- Analysis

Data Streaming/Scrapping:

We used the twitter to stream real time data during the presidential elections 2020. The raw data contains following three columns “Author,Date,Tweet” tweets with hashtags “'Elections2020', 'ElectionNight', 'Elections', 'Trump', 'Biden'”

To real time stream data from Twitter, we have created an app in Twitter. Using “tweepy” library in Python and using the “consumer_key”, “consumer_secret”, “access_token”, and “access_token_secret” from the app created in Twitter, we have streamed the live data during the event to fetch around 658280 Tweets.

Data Cleaning:

We used the MeTA analysis toolkit to clean the raw data that was scraped from Twitter. This involved using ‘stemming’ to treat base words as the same, in order to reduce the amount of noise in the analysis.

Analysis:

Current data is analyzed to create time-series of n-gram counts charts are plotted and embedded into html files. These Plots contain the top-20 most popular n-grams. We have computed for 2,3,4 and 5-grams. As per the proposal we have created the initial Sentiment Analysis for Presidential Elections 2020.

Which tasks are pending?

- Evaluation of the test results using Precision/Recall Measures are pending.
- Need to test the generation of charts on a different dataset to test the streamline working of the project.

Are you facing/faced any challenges?

There is no cap on the amount of data that needs to be collected to generate the correct sentiments. In other words, when should we stop scrapping the data from Twitter? This was one of the challenges we faced while scrapping the data from Twitter. In addition to the above, the limitations from Twitter to the number of calls that can be made to collect the data.

Data cleaning is an issue as all the data from twitter is not text. It includes images/gifs/smiles and videos too. Cleaning such data was not straight forward.

Still need to integrate the Evaluation of results using Precision and Recall measures. We are hoping to complete this in time.