CS410 Project Progress Report

Twitter Sentiment Analysis

Rudy Rath

Tasks Completed

The following tasks have been completed –

- Creation of Twitter Developer Account
- Generation of Twitter api key and access token for application
- Proof of Concept of scraping Twitter data using **Tweepy** package
- Research of **NLTK** features
- Research of **VADER** and **Textblob**

Tasks Pending

The following tasks are pending –

- Design decisions regarding the challenges explained in the next section.
- Coding for Scraping Twitter Data – Need refinement from what was done for POC
- Coding for Sentiment Analysis with NLTK and VADER – only have skeleton code so far from the examples
- Project Documentation

Challenges

So far, I have faced the following challenges that I need to overcome in the final design and implementation –

1. While developing the POC for scraping Twitter data, I have hit the Rate Limit error as documented in https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits several times. I need to factor that in the design of the application and decide if scraping is done at run time or if should be done in advance. While doing the scraping in real time by taking the input string from the user during run time, I have found that the optimal number of tweets so far is 1500. My preference is to do the scraping in real time after user input, so I will explore this rate limit issue further.

2. Since the Twitter api key and access token are linked to my developer account and are not supposed to be shared publicly, I need to figure out how to provide instructions to run the code when I submit the project. I will probably use **.gitignore** to put the config file in the GitHub repo but anyone trying the run my code in their own environment will still need the config file. If I'm not able to figure out the resolution on my own, I will seek advice from course instructors.

Project Proposal Meta-Reviews Feedback

One question in the Meta-Reviews the project proposal was about the dataset –

Whether I will be crawling for the dataset or use existing ones.

The answer to that is that I plan to scrape the twitter data and build my own dataset. An example of such a data set is in the "**extracted_tweets.csv**" file that I've uploaded to the GitHub repo. The phrase for this extraction was "**Biden**"