Zoe Cheng
MIS 3640
Professor Zhi Li
3/27/2020

Assignment 2 Write Up

Project Overview

For the assignment I decided to use Twitter as a data source because it offers an insight into the most up to date trending topics around the world and what people are saying about it. To analyze the data, I pickled and unloaded the tweets into a list then into a text file as well. I was hoping to see if there were any certain keywords repeated that are associated with the topic I decided to search on Twitter – Hong Kong.

Implementation

The first step in the project for me was to decide what topic to search for in Twitter to use as my data and analyze it. After deciding on using "Hong Kong', I went ahead to pickle the data. I knew that I wanted to have the tweets broken down into word by word in a text file for the easiest and most effective way to analyze the text themselves.

The Twitter data as itself once pickled included a lot of metadata and not just the tweets themselves. This meant that I would need to isolate the tweets and "get" them. The unloaded tweets in the list were each tweet as its own item so I therefore had to break the tweets down into singular words in order to perform any sort of text analysis. I decided to take the list and write it into a text file because I felt more comfortable with the data that way. Since we learned how to analyze the Project Gutenburg book in class, I found that it was this way I would be able to understand my own processes and code the best since I was familiar with it.

After writing the tweets by word into a text file I wanted to analyze the text first by the frequency of words to understand if there were any words that were associated to the search. I also wanted to see how many different words there were to compare the variety between tweets. I made sure to update my “detele_words.txt” file to include more relevant terms like the repetition of the word “RT” so that it will not be included in the analysis. In addition to the basic word frequencies, I also conducted a sentiment analysis on the data using NLTK. I wanted to see if there was an overall mood surrounding the topic of Hong Kong.

Results

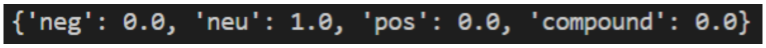
In my text analysis, I printed the total number of words, the number of different words, as well as the most common words found within tweets related to the search “Hong Kong”. There was a total of 1954 words collected from my data, 683 of which are different words. From my results, I found that many people who are tweeting about Hong Kong included the words “coronavirus”, “cases”, “65”, “opposition”, “sedition”, and “politician” as taken from the most common words used in the tweets. The table below shows the count of these words.

Word	**Frequency**
Coronavirus	21
Cases	19
65	19
Sedition	11
Opposition	11
Politician	11

Based on the most common words found in my search, the frequency of these words appearing within tweets is very interesting and tells a lot about what people are currently talking about when it comes to the topic of Hong Kong. First, there has been a spike in coronavirus cases in Hong Kong and as of today (3/27/2020), there were 65 new cases. This points to the top three most common words.

On the other hand, the words “sedition”, “opposition”, and “politician” were mentioned a few times as well because a recent news story surrounding a Hong Kong opposition politician surfaced where she was arrested under the offence of sedition. Both the coronavirus and the politician news stories are relevant to what is happening at the moment. This shows how Twitter’s content is always current and up to date to the latest developing news stories.

As for the sentiment analysis, I wanted to identify what the overall emotion and sentiment was within the tweets. This analysis would take the text from the tweets and gauge if it were a negative, neutral or positive sentiment. In this case, results showed that the tweets gathered from my search were neutral as shown in the picture. This was interesting to me because Twitter is often known to be a platform for people to voice their opinions, pointing to a more biased view of things. A neutral sentiment analysis of the text would suggest that the tweets I gathered were relatively objective, without positive or negative emotions linked to it.



```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

Reflection

Going into this project, I think I had a pretty good idea of what I wanted to do in terms of the process – find my data source (Twitter) and downloading the tweets, then analyzing it with ways that we learned in class. However, when it came to the actual data, it was a lot more work than I anticipated and it did not turn out the way I expected it to. One of the downsides with Twitter data is that tweets can vary in topic since people have different opinions and things to say, some even irrelevant to what I was searching.

Moving forward and taking what I've learned, I think I would reconsider using Twitter as a data source. It is interesting to see what people are tweeting about but at the same time, it is not as consistent as a book or Wikipedia page.