Takaki Kato
Professor Zhi Li
2020SP-01: Problem Solving & Software Design
26 March 2020

Assignment 2: Text-Mining

1. Project Overview

I have decided to analyze reddit data, particularly Coronavirus subreddit for this assignment. Firstly, I mined the raw data from reddit using 'praw' package and 'pushshit' API. Then processed the raw data into 'DataFrame' format using 'pandas' package', to clean and prepare the data for further analysis. Finally, I analyzed the data using word frequencies, and natural language processing using 'nltk' package, and visualized the data using 'mathplotlib' package. Main learning objectives for this assignment was to learn how to text-mine, process and clean the data, analyze the data through natural language processing, and visualize the data.

2. Implementation

Main components of my code can be divided into three parts, mining, processing, and analyzing the data.

First problem I faced in implementation of my code was choosing how to mine the data from reddit. The 'praw' package was suitable for mining the current data, such as the new or hot submissions currently on reddit; however, it was not suitable for mining the historical data with specific date of submission. Thus, I used 'praw' to mine the data to analyze the current hot submissions. For the historical data, 'pushshift' API was more suitable as it can be used to request submissions using 'after' and 'before' attributes to set the date range. Thus, I used 'pushshif' to mine the data to analyze the trend over time.

Another design decision I had to make in implementation of my code was the choice of data format to use for processing the data. After some research, I learned that it is more efficient to clean the data using 'DataFrame' format from 'pandas' package in comparison to using dictionaries and lists as it is easier to apply function to each submission. If I had used dictionaries or lists to store the data, I would have had to iterate over each item in the dataset to process the data, but with 'DataFrame' it can be achieved in one line using apply function. Conveniently natural language processing with 'nltk' package, and visualization with 'mathplotlib' package can be done easily with 'DataFrame' format as well.

Two negative points in using 'DataFrame' for this project was that in order to conduct word frequency analysis, and find the average sentiment scores, I had to convert the data back to list and dictionary format, because it is easier to conduct word frequency analysis by creating a list of all words, and result from sentiment analysis is returned in dictionary format.

3. Results

Below are the results for most frequently used words and average sentiment scores for 1000 hot submissions in r/all and r/Coronavirus at the time of writing. The purpose of this part was to the general trend in what words are most frequently used and what the sentiment of people are. r/all subreddit was used to gauge the mood of general population and r/Coronavirus subreddit was used to gauge the mood of population who are likely more aware and are concerned about the recent corona virus outbreak throughout the world. You can see that r/all the only key words that jumps out are related corona virus such as '19', coronavirus', and 'covid', which tells that corona virus is being talked about a lot in the public. Additionally, the compound sentiment score is 0.0586, which is very close to 0, meaning the sentiment among these submissions are still positive, but not very much so. In r/Coronavirus words that jump out are 'case', 'new', 'death', 'positive', which indicates people are generally talking about the new cases and deaths related to corona virus the most. As for the sentiment, it is unsurprisingly negative.

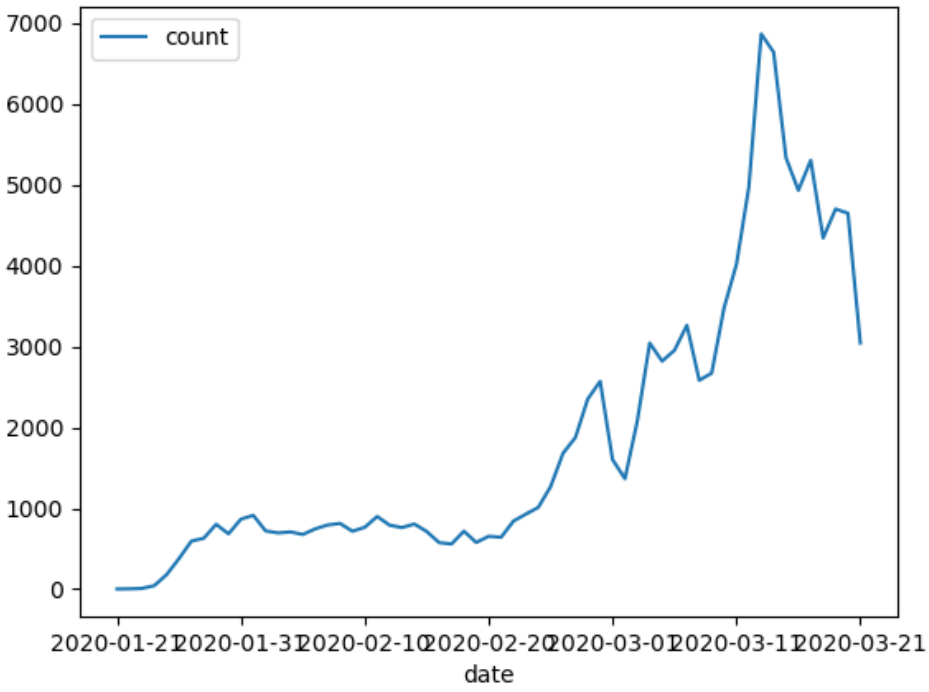
```
r/all f

Most frequently used words are:
day 22
year 21
make 20
one 17
new 16
get 16
19 14
coronavirus 14
covid 14
time 14

average sentiment scores are:
{'neg': 0.09905400000000006, 'neu': 0.7276910000000004, 'pos':
0.16025400000000004, 'compound': 0.05866469999999996}

r/Coronavirus

Most frequently used words are:
coronavirus 278
covid 154
19 141
case 124
new 95
death 64
test 53
positive 53
000 51
say 46
```

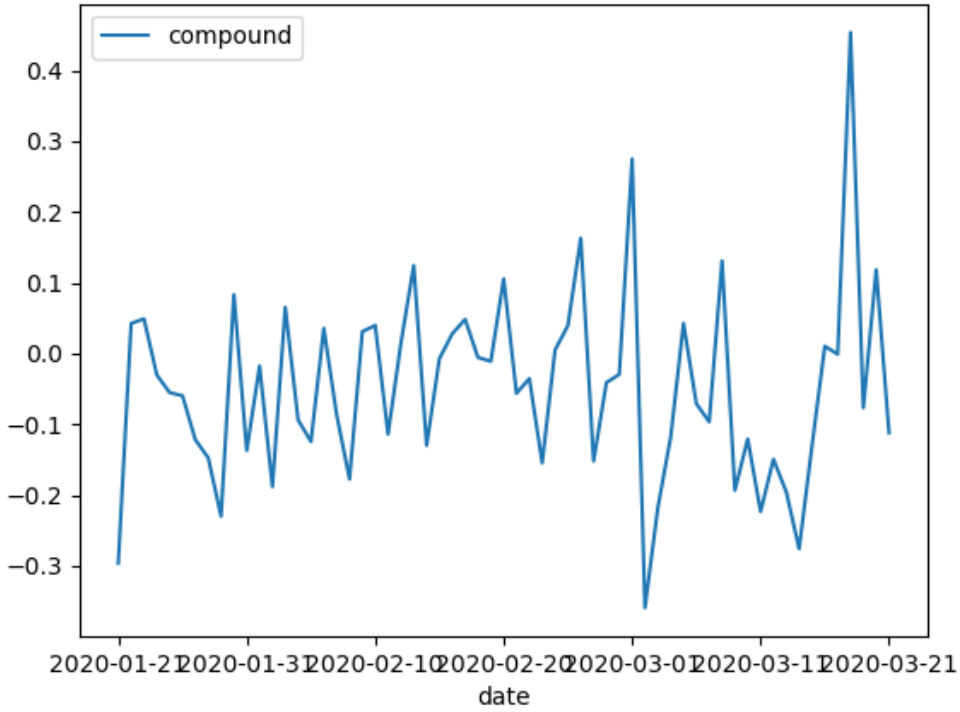
```
average sentiment scores are:
{'neg': 0.13620604099244893, 'neu': 0.7723495145631067, 'pos':
0.09144120819848978, 'compound': -0.08035059331175833}
```

Below are the average sentiment scores and number of submissions in r/Coronavirus between 1/21/2020 and 3/21/2020. I have printed out the first and last five dates. From this table you can already see that the number of submission have dramatically increased over the period; however, the trend for sentiment score is not very clear.

	date	neg	neu	pos	compound	count
0	2020-01-21	0.355000	0.645000	0.0000	-0.296000	1
1	2020-01-22	0.077333	0.851667	0.0710	0.042517	3
2	2020-01-23	0.009100	0.917300	0.0736	0.048990	8
3	2020-01-24	0.089900	0.768900	0.1412	-0.030190	42
4	2020-01-25	0.103700	0.838700	0.0577	-0.055110	178
..	...	...	...	...	...	...
56	2020-03-17	0.070800	0.846800	0.0824	-0.000910	5304
57	2020-03-18	0.035000	0.704000	0.2610	0.453590	4343
58	2020-03-19	0.158300	0.735500	0.1062	-0.076480	4701
59	2020-03-20	0.118700	0.716000	0.1653	0.118240	4649
60	2020-03-21	0.105500	0.846300	0.0482	-0.111940	3044

To better visualize the trend over time I have created two line graphs comparing date and number of submission, and date and compound sentiment score. You can see that there is dramatic rise in number of submissions starting around end of February. This is likely the period when the corona virus has started to spread in countries outside of China and global population have become more aware and concerned about the virus. Similarly, for the compound sentiment score it was gradually increasing, but there is decline starting around the same date. Additionally, the recovery of compound sentiment score in more recent dates can likely be attributed to submissions concerning the positive measures taken against corona virus and people starting to recover in some countries.


count
7000
6000
5000
4000
3000
2000
1000
0
2020-01-21
2020-01-31
2020-02-10
2020-02-20
2020-03-01
2020-03-11
2020-03-21
date


compound
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2020-01-21
2020-01-31
2020-02-10
2020-02-20
2020-03-01
2020-03-11
2020-03-21
date

4. Reflection

From process point of view the entire project flow very well through each step as I have taken the time to plan out the general structure and components of the project. Some thing that I could improve is to make the code cleaner and more efficient as there may be some unnecessary conversions between data formats, which lead to unnecessarily longer codes. I believe that my project was appropriately scoped, but I could have decreased the scope a little more as there were many new things, I had to learn to complete this project. One of the things I wish I had known before starting was that ‘praw’ was not suitable to mining historical data, as trying to find alternative method of mining the data took the longest for me. There are many applications in the future for things I have learned through this project, but I believe the ‘pandas’ dataset was would be most useful in the future as it allows me to do many things in more efficient manners as compared to using iterations on lists and dictionaries to process large amount of data.