

Data Management, Warehousing and Analytics

Assignment 4

A. Sentiment Analysis

I have used python script *sentiment_analysis.py* so perform all the tasks in this section. I have used *JSON/tweets.json* file to get the tweets. These are the same tweets that were generated in the pervious assignment. The script creates a bag of words for each tweet and stores it in file *JSON/bag_of_words.json*. Figure 1 shows the contents of the file.

```
    "bow9": {
      "How": 1,
      "was": 1,
      "I": 1,
      "shuffled": 1,
      "into": 1,
      "ja": 1
    }
  },
  {
    "bow10": {
      "CANADA": 2,
      "INC": 1,
      "Canada": 2,
      "Look": 1,
      "at": 1,
      "this": 1
    }
  },
  {
    "bow11": {
      "So": 1,
      "like": 1,
      "we": 1,
      "gonna": 1,
      "talk": 1,
      "about": 1,
      "how": 1,
      "it": 1,
      "s": 1,
      "am": 1,
      "where": 1,
      "Justin": 1,
      "is": 2,
      "in": 1,
      "Canada": 1,
      "and": 1,
      "he": 1,
      "": 1
    }
  }
}
```

Figure 1

I have used a list of positive and negative words from University of Illinois at Chicago website [1] [2]. Positive words are stored in *TXT/positive-words.txt* and negative words are stored in *TXT/negative-words.txt*. I, then, compared each tweet with these lists to find the sentiments of a tweet. Figure 2 shows table showing each tweet tagged as “positive”, “negative” and “neutral”.

1	Tweet	Message	polarity	match
2	1	RT And this article provides a great example of the kind of community spirit that we need to get through this a spirit that w	positive	['great']
3	2	RT Did I see you Americans are getting a one time check to cover COVID unemployment Because Canada just announced	neutral	
4	4	RT Canadians Avoid non essential travel outside of Canada If you choose to travel despite our advice you may have difficul	negative	['difficulty']
5	5	Second Wave of Retail Bankruptcies Expected in Canada Amid COVID Pandemic Expert	neutral	
6	6	RT m marandi Govts that block steal Iranian assets at the behest of Trump or on their own initiative are directly aiding the U	neutral	
7	7	RT NEW If you are returning home from outside of Canada the government will now be using the Quarantine Act to enforce	neutral	
8	10	CANADA INC Canada Look at this CANADA Canada	neutral	
9	11	So like we gonna talk about how it s am where Justin is in Canada and he is just now going to bed SRRRRR	positive	['like']
10	12	RT Did I see you Americans are getting a one time check to cover COVID unemployment Because Canada just announced	neutral	
11	13	Sigan consumi ndole a los gringos	neutral	
12	16	RT Different people in Canada and the U S report massive movement of heavy military equipment Is this the real reason for	neutral	
13	17	RT It s now clear albeit with the benefit of hindsight that the government indeed all governments in Canada underesti	positive	['clear', 'benefit']
14	20	RT n Pandemics with Chinese Characteristics Buy all medical supplies around the world Make domestic sales illegal Donat	negative	['illegal']
15	22	RT KashmirUnderModiCorona Prevent the next Genocide in India by Three emails if you live in Canada Please find the lette	negative	['Genocide']
16	25	RT Did I see you Americans are getting a one time check to cover COVID unemployment Because Canada just announced	neutral	
17	26	RT How countries are responding to the Coronavirus England Paying of people s salary Denmark Paying of people s	neutral	
18	27	RT nouveaux relev s d tat civil sont maintenant sur Filae O Ain Ardennes C te d Or Dordogne Maine et Loire M	neutral	
19	28	BLAST FROM THE PAST Canada Annan nd XV team who took on Montreal Barbarians Smile for the camera StayIndoors St	positive	['Smile']
20	29	RT Did I see you Americans are getting a one time check to cover COVID unemployment Because Canada just announced	neutral	
21	30	Yes I think it would be a great idea for Harry Meghan and Archie to travel from Canada to come and stay with their elderly ar	neutral	
22	31	RT It may suffice to mention that no one would have to blackmail anyone to show empathy to Justin if he contracts the virus	neutral	
23	32	To my dear friends in Canada please take head Ismaili Ismailimail	neutral	
24	33	RT This is a COVID Canadian freelance artist resource list If you know of any other resources please tweet me And plea	negative	['plea']
25	34	RT f First Pakistani Winter Olympian Brings Pride to a Small Village Skier Muhammad Abbas made a little sporting history by	positive	['Pride']
26	35	Acetech is a leading software developmentcompany Acetech has over a decade of rich experience in delivering critical	positive	['leading', 'rich']
27	36	RT Did I see you Americans are getting a one time check to cover COVID unemployment Because Canada just announced	neutral	
28	38	Hi sir people in Canada has full geo location map details of Covid patient found before he is isolated Data for Telangana mig	neutral	
29	39	But in China Clean street High technology Delicious foods and developed cities how dare the fking India always compared	positive	['Clean', 'Delicious']

Figure 2

Figure 3 shows a word cloud for all the positive words in the tweets. I have used file *TXT/positive_tableau.txt* to construct the word cloud in tableau [3].



Figure 3

Figure 4 shows a word cloud for all the negative words in the tweets. I have used file *TXT/negative_tableau.txt* to construct the word cloud in tableau.

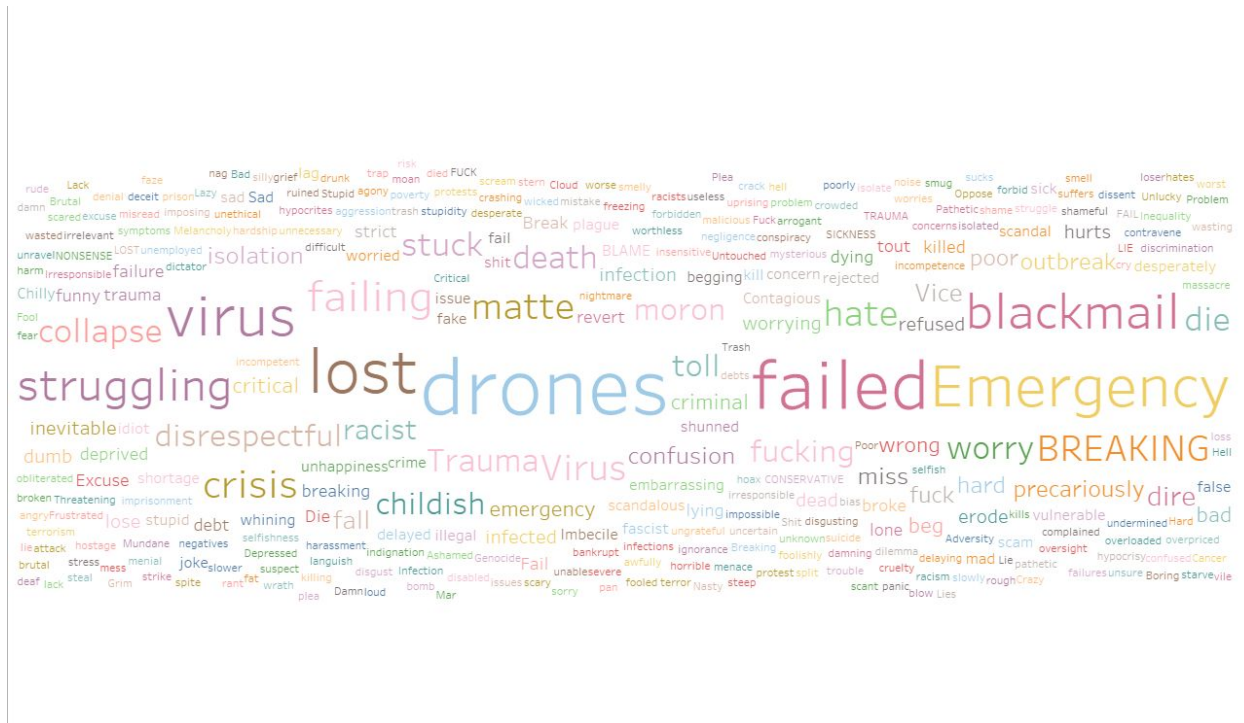


Figure 4

B. Semantic Analysis

I have used the script *semantic_analysis.py* to perform all the operations in this section. I have used *JSON/news_articles.json* for all the news articles. This is the same file generated in the previous assignment. The script converts all the news articles into separate files such that each file contains one news article and contains information “title”, “description” and “content”. Figure 5 shows the required table generated from the news articles files and calculates the frequency-inverse document frequency of each keyword.

	A	B	C	D
1	Query	Documents_containing_term	n/df	log(n/df)
2	Canada	137	5.11	0.71
3	University	67	10.45	1.02
4	Dalhousie University	10	70	1.85
5	Halifax	55	12.73	1.1
6	Business	34	20.59	1.31
7				

Figure 5

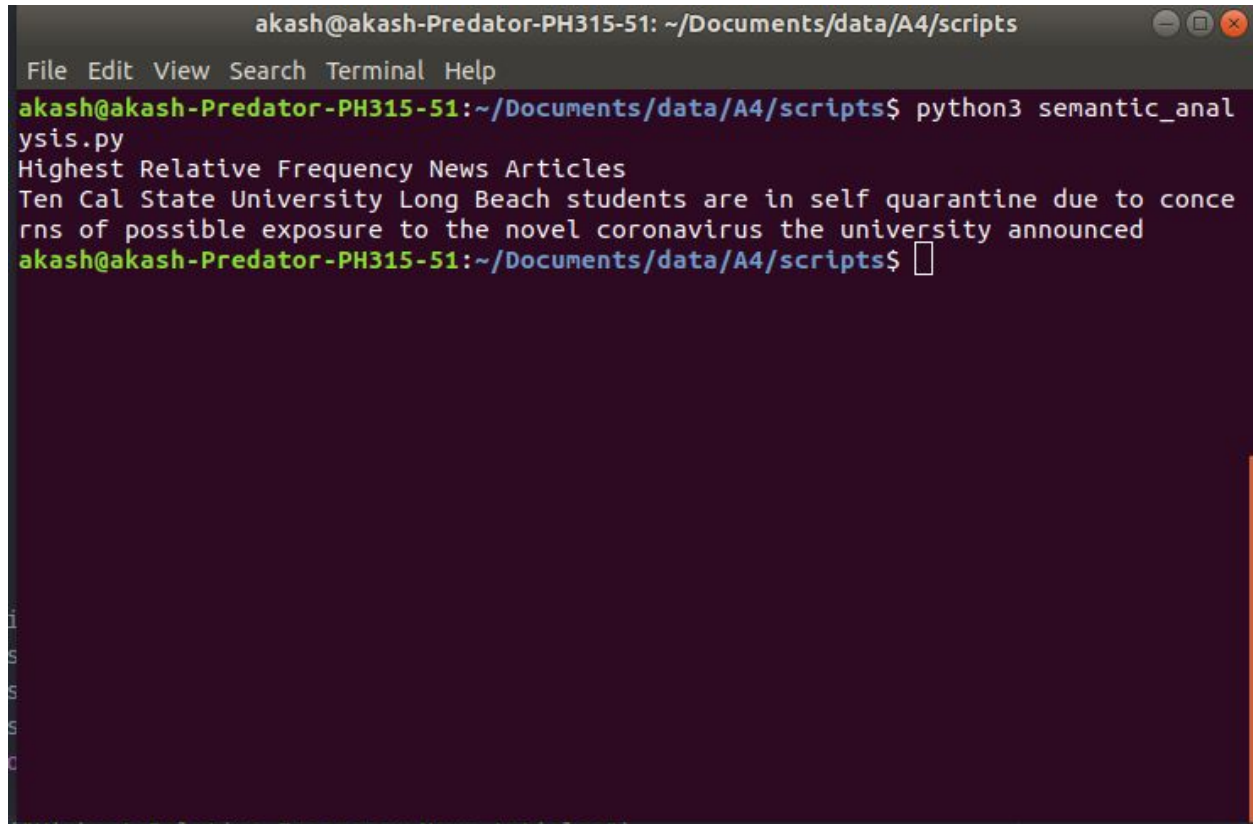
Figure 6 shows a table having all the articles containing the keyword “Canada” along with total words and frequency of the keyword.

	A	B	C	D
1	Article	frequency	Total words	
2	article_91	1	40	
3	article_657	1	41	
4	article_47	1	43	
5	article_247	1	38	
6	article_613	1	40	
7	article_12	1	44	
8	article_16	1	44	
9	article_77	2	46	
10	article_496	1	44	
11	article_63	1	42	
12	article_434	1	42	
13	article_674	1	46	
14	article_93	1	40	
15	article_239	1	42	
16	article_533	1	43	
17	article_33	1	51	
18	article_1	1	43	
19	article_494	1	44	
20	article_15	1	46	
21	article_528	1	45	
22	article_31	2	42	
23	article_9	1	42	
24	article_676	1	45	
25	article_636	1	38	
26	article_5	1	49	
27	article_482	1	44	
28	article_17	1	43	
29	article_14	1	40	

Figure 6

The corresponding table for other keywords can be found in directory *Semantic_analysis_table*.

Figure 7 shows the output of the script displaying the news article with the highest relative frequency.



```
akash@akash-Predator-PH315-51: ~/Documents/data/A4/scripts
File Edit View Search Terminal Help
akash@akash-Predator-PH315-51:~/Documents/data/A4/scripts$ python3 semantic_analysis.py
Highest Relative Frequency News Articles
Ten Cal State University Long Beach students are in self quarantine due to concerns of possible exposure to the novel coronavirus the university announced
akash@akash-Predator-PH315-51:~/Documents/data/A4/scripts$
```

Figure 8

References

- [1] *Ptrckprry.com*, 2010. [Online]. Available: <http://ptrckprry.com/course/ssd/data/positive-words.txt>. [Accessed: 11-Apr-2020].
- [2] *Ptrckprry.com*, 2010. [Online]. Available: <http://ptrckprry.com/course/ssd/data/negative-words.txt>. [Accessed: 11-Apr-2020].
- [3] “Creating a Word Cloud | Tableau Software,” *Tableau.com*, 2014. [Online]. Available: <https://kb.tableau.com/articles/howto/creating-a-word-cloud>. [Accessed: 11-Apr-2020].