Assignment 4

A. Sentiment Analysis

The **cleaned_twitter_data.csv** in **Twitter** folder contains the twitter data from the Assignment 3. The URLs and special characters were already removed in the Assignment 3. The **polarity_twitter_data.py** is a python script which makes bag-of-words for each tweets and compares the words with the list of positive and negative words. The **positive_words.txt** and **negative_words.txt** are the text files containing a list of positive and negative words respectively. These lists have been referenced from Github[1].

The script keeps the count of positive and negative words in a tweet and accordingly assigns the polarity to the tweet. The **twitter_data_polarity.csv** contains tweets with their polarity. Simultaneously, the script stores all the matched positive and negative words in a list. At last, it counts the frequency of the positive and negative words in the list. It stores this information in **words_polarity.csv.** The **tableu.png** and **word_cloud.png** contains the screenshots of Tableu[4] and word cloud.

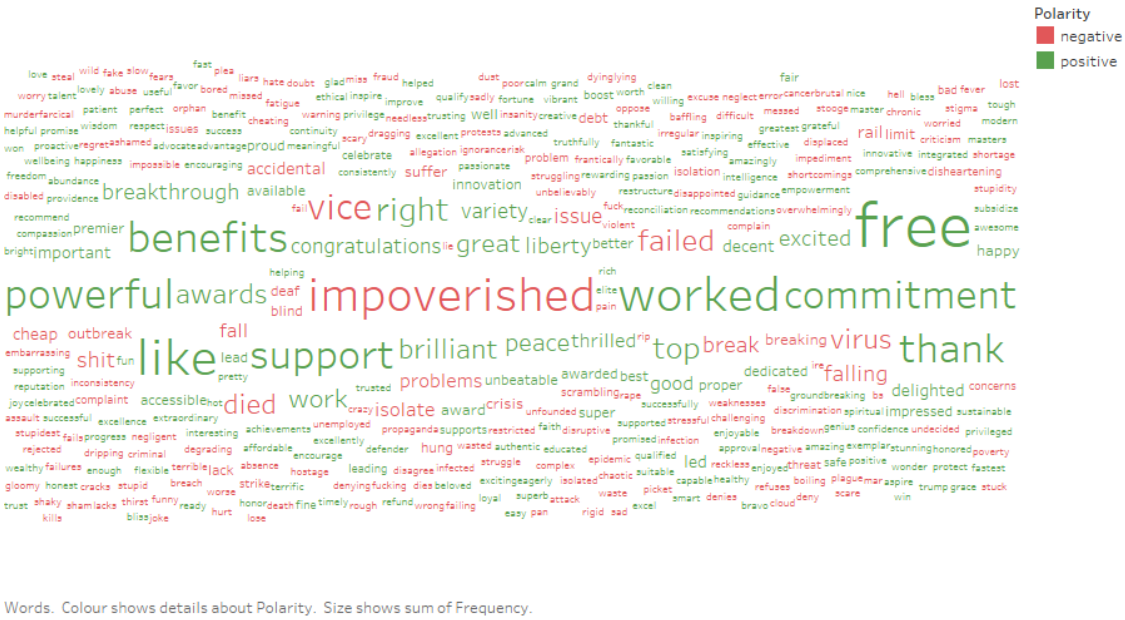
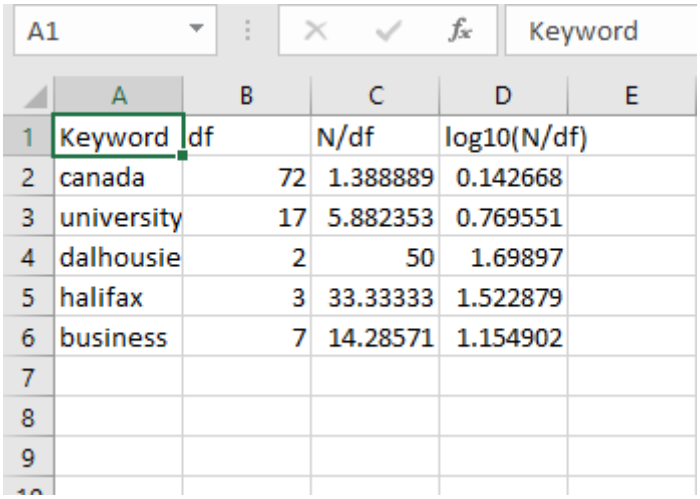


Figure 1 Word Cloud

B. Semantic Analysis

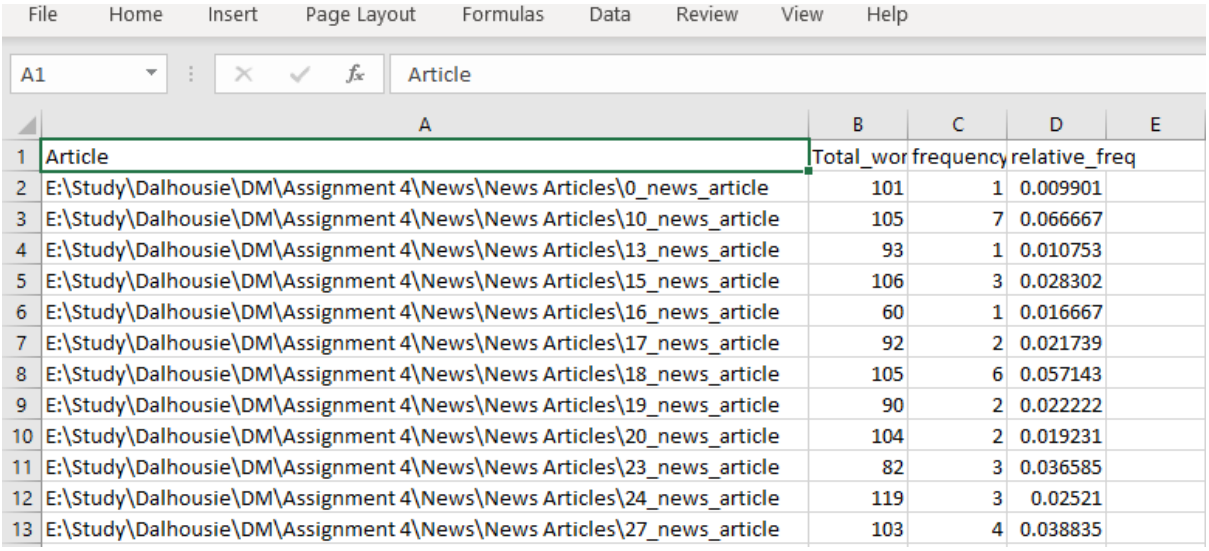
The **cleaned_news_data.csv** contains the news data which is cleaned using **cleaning_news_data.py** script. The script is same which was used in Assignment 3. The **Semantic_Analysis.py** takes each tweet and save it in separate files. These files can be found in **News_Articles** folder. The script also searches for words like 'canada' , 'university' , 'dalhousie university', 'halifax', and 'business' in ‘title’, ’description’ and ‘content’ of every tweet. It keeps count of number of documents these keywords occur in and stores that information in **10.a.csv** file.



	A	B	C	D	E
1	Keyword	df	N/df	log10(N/df)	
2	canada	72	1.388889	0.142668	
3	university	17	5.882353	0.769551	
4	dalhousie	2	50	1.69897	
5	halifax	3	33.33333	1.522879	
6	business	7	14.28571	1.154902	
7					
8					
9					

Figure 2 TF-IDF of keywords

The script also finds the number of occurrences of word “Canada” in documents. It then counts the total words in a document and calculates relative frequency. The **relative_frequency.csv** contains the information about frequency and relative frequency of word “Canada” in every documents it occurs.



	A	B	C	D	E
1	Article	Total_wor	frequency	relative_freq	
2	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\0_news_article	101	1	0.009901	
3	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\10_news_article	105	7	0.066667	
4	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\13_news_article	93	1	0.010753	
5	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\15_news_article	106	3	0.028302	
6	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\16_news_article	60	1	0.016667	
7	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\17_news_article	92	2	0.021739	
8	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\18_news_article	105	6	0.057143	
9	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\19_news_article	90	2	0.022222	
10	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\20_news_article	104	2	0.019231	
11	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\23_news_article	82	3	0.036585	
12	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\24_news_article	119	3	0.02521	
13	E:\Study\Dalhousie\DM\Assignment 4\News\News Articles\27_news_article	103	4	0.038835	

Figure 3 Frequency and Relative Frequency

At last, the scripts prints the article which has the highest relative frequency.

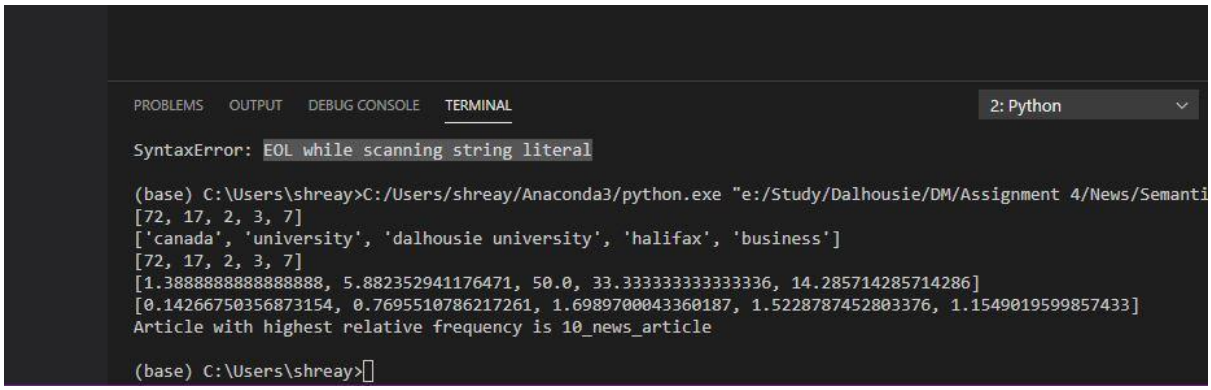


Figure 4 Article with highest relative frequency

References:

[1]"jeffreybreen/twitter-sentiment-analysis-tutorial-201107", *GitHub*, 2020. [Online]. Available: https://github.com/jeffreybreen/twitter-sentiment-analysis-tutorial-201107/tree/master/data/opinion-lexicon-English. [Accessed: 04- Apr- 2020].

[2]"Getting started", Developer.twitter.com, 2020. [Online]. Available:https://developer.twitter.com/en/docs/basics/getting-started. [Accessed: 27-Mar-2020].

[3]"Get started -Documentation -News API", Newsapi.org, 2020. [Online]. Available: https://newsapi.org/docs/get-started. [Accessed: 27-Mar-2020].

[4]"Tableau Desktop", *Tableau Software*, 2020. [Online]. Available: https://www.tableau.com/en-ca/products/desktop. [Accessed: 05- Apr- 2020].