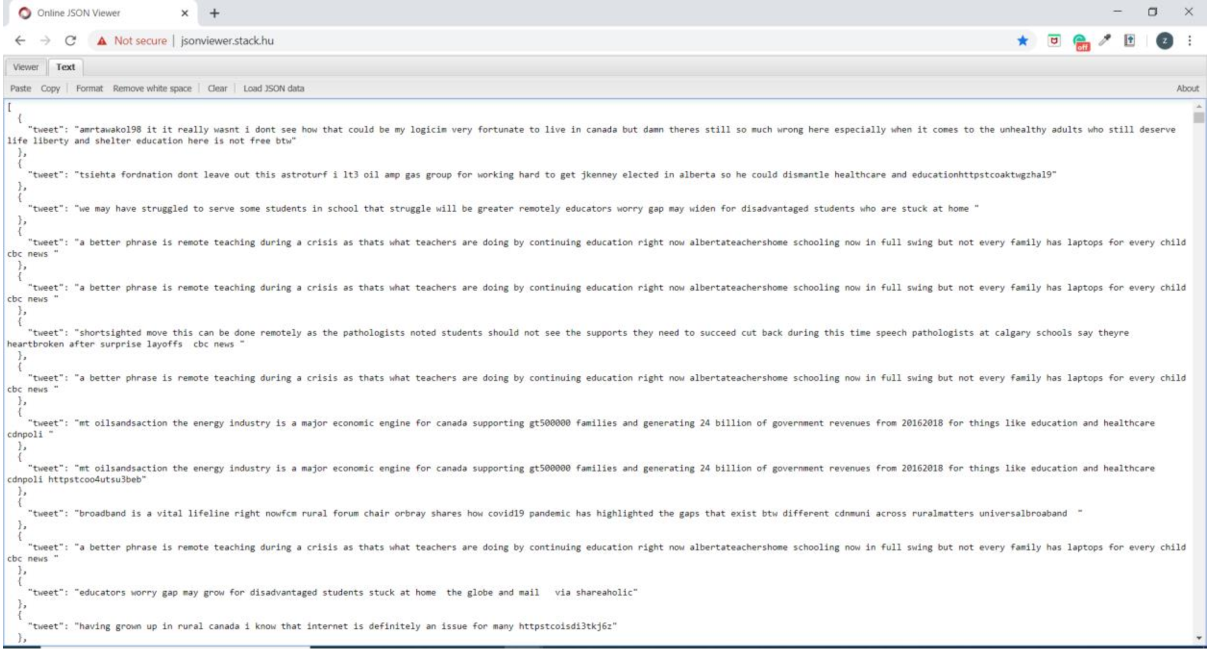
Assignment – 4

A. Sentiment Analysis:

Following steps were followed to perform sentiment analysis on tweets

Step 1: Only the tweets/messages were stored in separate file (tweets.json), which was collected using python script from assignment 3. All the tweets collected using the python script was cleaned. For cleaning purposes all the punctuation, URLs, emojis, and Unicode were removed in the previous assignment. Python script file (twitter.py) can be found in the Script folder.



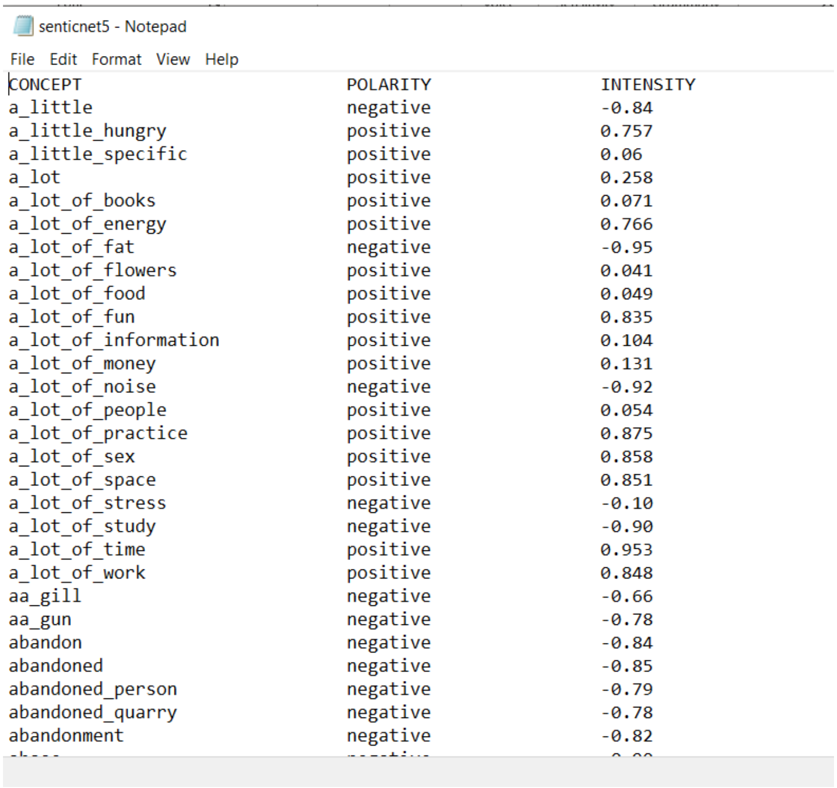
```
[
  {
    "tweet": "amrtawakol98 it it really wasnt i dont see how that could be my logicim very fortunate to live in canada but damn theres still so much wrong here especially when it comes to the unhealthy adults who still deserve life liberty and shelter education here is not free btw"
  },
  {
    "tweet": "tsiehta fordnation dont leave out this astroturf i lt3 oil amp gas group for working hard to get jkenney elected in alberta so he could dismantle healthcare and educationhttpstcoaktwgzhal9"
  },
  {
    "tweet": "we may have struggled to serve some students in school that struggle will be greater remotely educators worry gap may widen for disadvantaged students who are stuck at home "
  },
  {
    "tweet": "a better phrase is remote teaching during a crisis as thats what teachers are doing by continuing education right now albertateachershome schooling now in full swing but not every family has laptops for every child cbc news "
  },
  {
    "tweet": "a better phrase is remote teaching during a crisis as thats what teachers are doing by continuing education right now albertateachershome schooling now in full swing but not every family has laptops for every child cbc news "
  },
  {
    "tweet": "shortsighted move this can be done remotely as the pathologists noted students should not see the supports they need to succeed cut back during this time speech pathologists at calgary schools say theyre heartbroken after surprise layoffs  cbc news "
  },
  {
    "tweet": "a better phrase is remote teaching during a crisis as thats what teachers are doing by continuing education right now albertateachershome schooling now in full swing but not every family has laptops for every child cbc news "
  },
  {
    "tweet": "mt oilsandsaction the energy industry is a major economic engine for canada supporting gt500000 families and generating 24 billion of government revenues from 20162018 for things like education and healthcare cdnpoli "
  },
  {
    "tweet": "mt oilsandsaction the energy industry is a major economic engine for canada supporting gt500000 families and generating 24 billion of government revenues from 20162018 for things like education and healthcare cdnpoli httpstcoo4utsu3beb"
  },
  {
    "tweet": "broadband is a vital lifeline right nowfcm rural forum chair orbray shares how covid19 pandemic has highlighted the gaps that exist btw different cdnmuni across ruralmatters universalbroaband  "
  },
  {
    "tweet": "a better phrase is remote teaching during a crisis as thats what teachers are doing by continuing education right now albertateachershome schooling now in full swing but not every family has laptops for every child cbc news "
  },
  {
    "tweet": "educators worry gap may grow for disadvantaged students stuck at home  the globe and mail   via shareaholic"
  },
  {
    "tweet": "having grown up in rural canada i know that internet is definitely an issue for many httpstcoisdi3tkj6z"
  },
```

Figure 1 tweets in tweets.json opened in http://jsonviewer.stack.hu/

Step 2: Each tweet was read from (tweets.json) file. After that bag-of-word was created for each tweet, which shows the word and its occurrence in tweet.

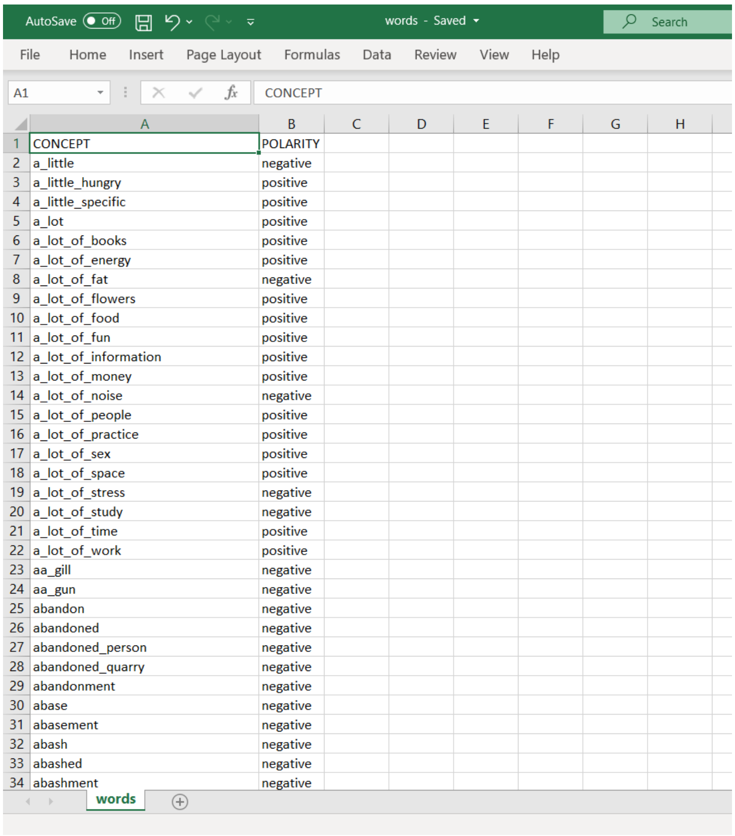
Step3: Text file containing the list of words along with its polarity and intensity was downloaded from (https://sentic.net/downloads/). Senticnet5 was selected as it consists of approximately 1,00,000 words.[1]

This list was modified and stored in separate CSV file (words.csv) showing word and its polarity as we did not require intensity of the word in our analysis.



```
CONCEPT                          POLARITY                 INTENSITY
a_little                         negative                 -0.84
a_little_hungry                  positive                 0.757
a_little_specific                positive                 0.06
a_lot                            positive                 0.258
a_lot_of_books                   positive                 0.071
a_lot_of_energy                  positive                 0.766
a_lot_of_fat                     negative                 -0.95
a_lot_of_flowers                 positive                 0.041
a_lot_of_food                    positive                 0.049
a_lot_of_fun                     positive                 0.835
a_lot_of_information             positive                 0.104
a_lot_of_money                   positive                 0.131
a_lot_of_noise                   negative                 -0.92
a_lot_of_people                  positive                 0.054
a_lot_of_practice                positive                 0.875
a_lot_of_sex                     positive                 0.858
a_lot_of_space                   positive                 0.851
a_lot_of_stress                  negative                 -0.10
a_lot_of_study                   negative                 -0.90
a_lot_of_time                    positive                 0.953
a_lot_of_work                    positive                 0.848
aa_gill                          negative                 -0.66
aa_gun                           negative                 -0.78
abandon                          negative                 -0.84
abandoned                        negative                 -0.85
abandoned_person                 negative                 -0.79
abandoned_quarry                 negative                 -0.78
abandonment                      negative                 -0.82
```

Figure 2 senticnet5.txt showing word along with polarity and intensity



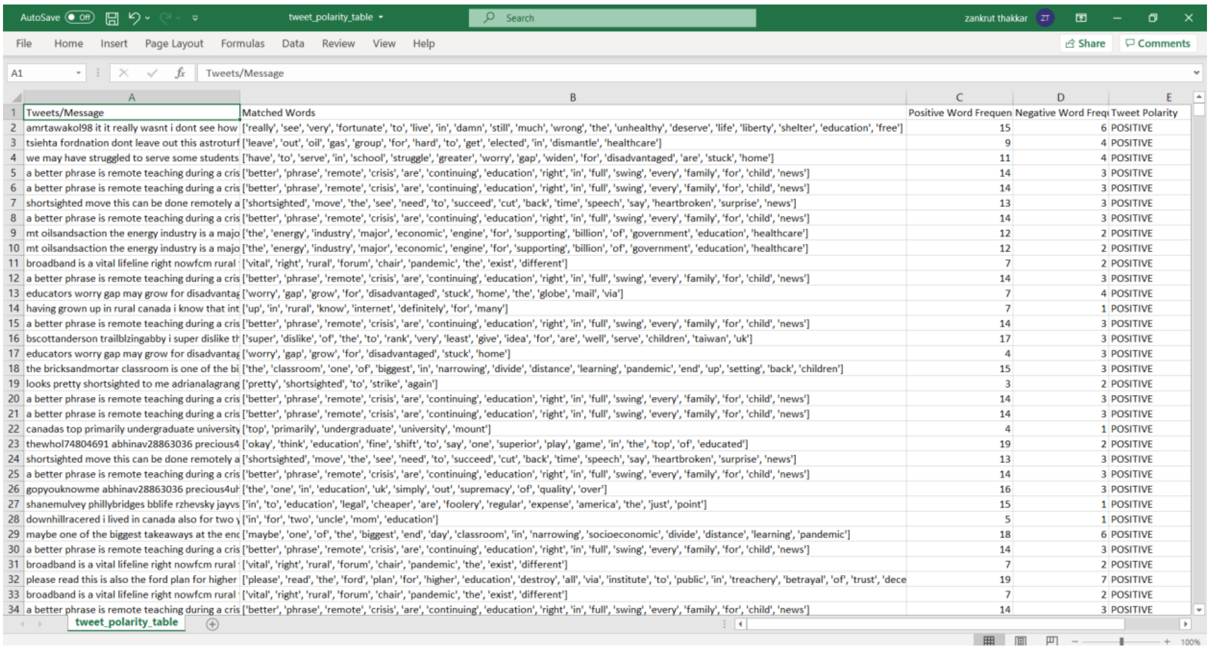
	A	B	C	D	E	F	G	H
1	CONCEPT	POLARITY						
2	a_little	negative						
3	a_little_hungry	positive						
4	a_little_specific	positive						
5	a_lot	positive						
6	a_lot_of_books	positive						
7	a_lot_of_energy	positive						
8	a_lot_of_fat	negative						
9	a_lot_of_flowers	positive						
10	a_lot_of_food	positive						
11	a_lot_of_fun	positive						
12	a_lot_of_information	positive						
13	a_lot_of_money	positive						
14	a_lot_of_noise	negative						
15	a_lot_of_people	positive						
16	a_lot_of_practice	positive						
17	a_lot_of_sex	positive						
18	a_lot_of_space	positive						
19	a_lot_of_stress	negative						
20	a_lot_of_study	negative						
21	a_lot_of_time	positive						
22	a_lot_of_work	positive						
23	aa_gill	negative						
24	aa_gun	negative						
25	abandon	negative						
26	abandoned	negative						
27	abandoned_person	negative						
28	abandoned_quarry	negative						
29	abandonment	negative						
30	abase	negative						
31	abasement	negative						
32	abash	negative						
33	abashed	negative						
34	abashment	negative						

Figure 3 words.csv containing word and its polarity

Step 3: Now, each word from bag-of-words is matched with these words. If the word is present, then it is collected along with its polarity to compute the overall polarity of the tweet.
For each tweet, we calculated total positive frequency and negative frequency from the matching words. Now, to compute the overall frequency of the tweet:

positive frequency > negative frequency than overall polarity of tweet is "POSITIVE"
positive frequency < negative frequency than overall polarity of tweet is "NEGATIVE"
positive frequency = negative frequency than overall polarity of tweet is "NEUTRAL"

Each tweet along with the matched words, Positive Word frequency, Negative Word frequency, and Tweet Polarity is stored in the (tweet_polarity_table.csv) file it can be found in Output folder.

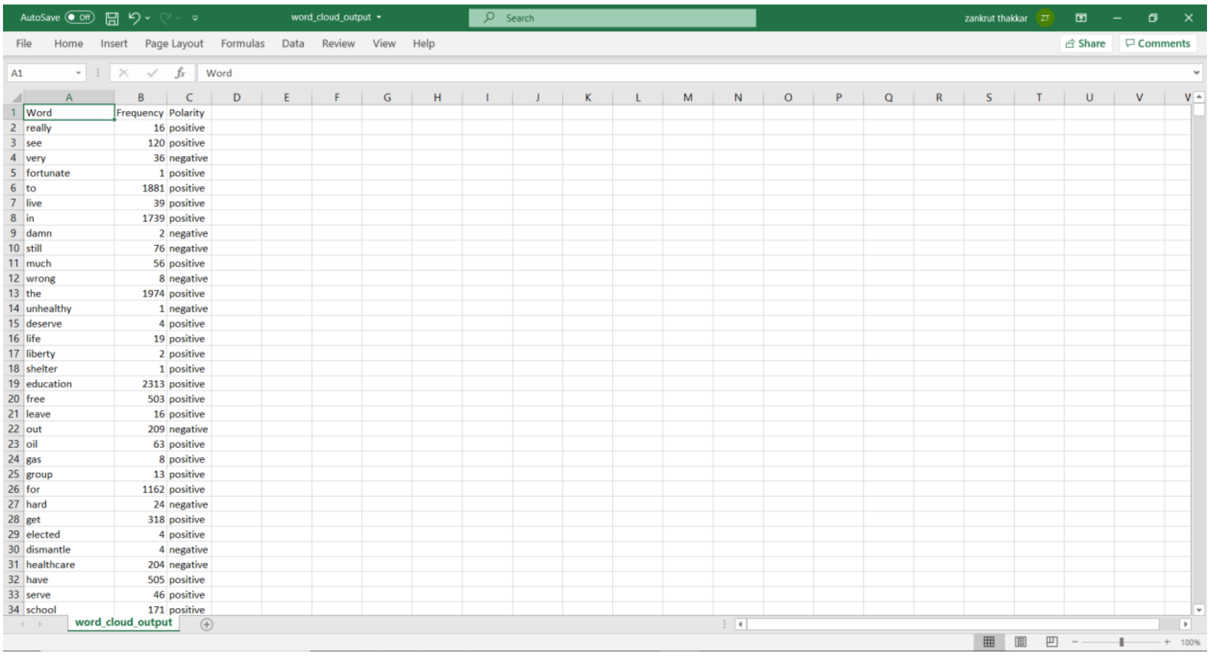


	A	B	C	D	E
1	Tweets/Message	Matched Words	Positive Word Frequen	Negative Word Frequ	Tweet Polarity
2	amrtawakol98 it it really wasnt i dont see how	['really', 'see', 'very', 'fortunate', 'to', 'live', 'in', 'damn', 'still', 'much', 'wrong', 'the', 'unhealthy', 'deserve', 'life', 'liberty', 'shelter', 'education', 'free']	15	6	POSITIVE
3	tsiehta fordnation dont leave out this astroturf	['leave', 'out', 'oil', 'gas', 'group', 'for', 'hard', 'to', 'get', 'elected', 'in', 'dismantle', 'healthcare']	9	4	POSITIVE
4	we may have struggled to serve some students	['have', 'to', 'serve', 'in', 'school', 'struggle', 'greater', 'worry', 'gap', 'widen', 'for', 'disadvantaged', 'are', 'stuck', 'home']	11	4	POSITIVE
5	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
6	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
7	shortsighted move this can be done remotely a	['shortsighted', 'move', 'the', 'see', 'need', 'to', 'succeed', 'cut', 'back', 'time', 'speech', 'say', 'heartbroken', 'surprise', 'news']	13	3	POSITIVE
8	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
9	mt oilsandsaction the energy industry is a majo	['the', 'energy', 'industry', 'major', 'economic', 'engine', 'for', 'supporting', 'billion', 'of', 'government', 'education', 'healthcare']	12	2	POSITIVE
10	mt oilsandsaction the energy industry is a majo	['the', 'energy', 'industry', 'major', 'economic', 'engine', 'for', 'supporting', 'billion', 'of', 'government', 'education', 'healthcare']	12	2	POSITIVE
11	broadband is a vital lifeline right nowfcm rural	['vital', 'right', 'rural', 'forum', 'chair', 'pandemic', 'the', 'exist', 'different']	7	2	POSITIVE
12	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
13	educators worry gap may grow for disadvantag	['worry', 'gap', 'grow', 'for', 'disadvantaged', 'stuck', 'home', 'the', 'globe', 'mail', 'via']	7	4	POSITIVE
14	having grown up in rural canada i know that int	['up', 'in', 'rural', 'know', 'internet', 'definitely', 'for', 'many']	7	1	POSITIVE
15	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
16	bscottanderson trailblzingabby i super dislike th	['super', 'dislike', 'of', 'the', 'to', 'rank', 'very', 'least', 'give', 'idea', 'for', 'are', 'well', 'serve', 'children', 'taiwan', 'uk']	17	3	POSITIVE
17	educators worry gap may grow for disadvantag	['worry', 'gap', 'grow', 'for', 'disadvantaged', 'stuck', 'home']	4	3	POSITIVE
18	the bricksandmortar classroom is one of the bi	['the', 'classroom', 'one', 'of', 'biggest', 'in', 'narrowing', 'divide', 'distance', 'learning', 'pandemic', 'end', 'up', 'setting', 'back', 'children']	15	3	POSITIVE
19	looks pretty shortsighted to me adrianalagrang	['pretty', 'shortsighted', 'to', 'strike', 'again']	3	2	POSITIVE
20	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
21	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
22	canadas top primarily undergraduate university	['top', 'primarily', 'undergraduate', 'university', 'mount']	4	1	POSITIVE
23	thewhol74804691 abhinav28863036 precious4	['okay', 'think', 'education', 'fine', 'shift', 'to', 'say', 'one', 'superior', 'play', 'game', 'in', 'the', 'top', 'of', 'educated']	19	2	POSITIVE
24	shortsighted move this can be done remotely a	['shortsighted', 'move', 'the', 'see', 'need', 'to', 'succeed', 'cut', 'back', 'time', 'speech', 'say', 'heartbroken', 'surprise', 'news']	13	3	POSITIVE
25	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
26	gopyouknowme abhinav28863036 precious4uk	['the', 'one', 'in', 'education', 'uk', 'simply', 'out', 'supremacy', 'of', 'quality', 'over']	16	3	POSITIVE
27	shanemulvey phillybridges bblife rzhevsky jayvs	['in', 'to', 'education', 'legal', 'cheaper', 'are', 'foolery', 'regular', 'expense', 'america', 'the', 'just', 'point']	15	1	POSITIVE
28	downhillracered i lived in canada also for two y	['in', 'for', 'two', 'uncle', 'mom', 'education']	5	1	POSITIVE
29	maybe one of the biggest takeaways at the enc	['maybe', 'one', 'of', 'the', 'biggest', 'end', 'day', 'classroom', 'in', 'narrowing', 'socioeconomic', 'divide', 'distance', 'learning', 'pandemic']	18	6	POSITIVE
30	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE
31	broadband is a vital lifeline right nowfcm rural	['vital', 'right', 'rural', 'forum', 'chair', 'pandemic', 'the', 'exist', 'different']	7	2	POSITIVE
32	please read this is also the ford plan for higher	['please', 'read', 'the', 'ford', 'plan', 'for', 'higher', 'education', 'destroy', 'all', 'via', 'institute', 'to', 'public', 'in', 'treachery', 'betrayal', 'of', 'trust', 'dece	19	7	POSITIVE
33	broadband is a vital lifeline right nowfcm rural	['vital', 'right', 'rural', 'forum', 'chair', 'pandemic', 'the', 'exist', 'different']	7	2	POSITIVE
34	a better phrase is remote teaching during a cris	['better', 'phrase', 'remote', 'crisis', 'are', 'continuing', 'education', 'right', 'in', 'full', 'swing', 'every', 'family', 'for', 'child', 'news']	14	3	POSITIVE

Figure 4 tweet_polarity_table.csv

Visualisation:

Step 4: For Visualisation of data using tableau. All the matching words along with its total occurrences in all the tweets and polarity are collected and stored in separate file (world_cloud_output.csv), that can be found in Output folder. As our file (words.csv) consists of many words, there are many words matching found from the tweets. The polarity of each matched word is based on the data we downloaded.

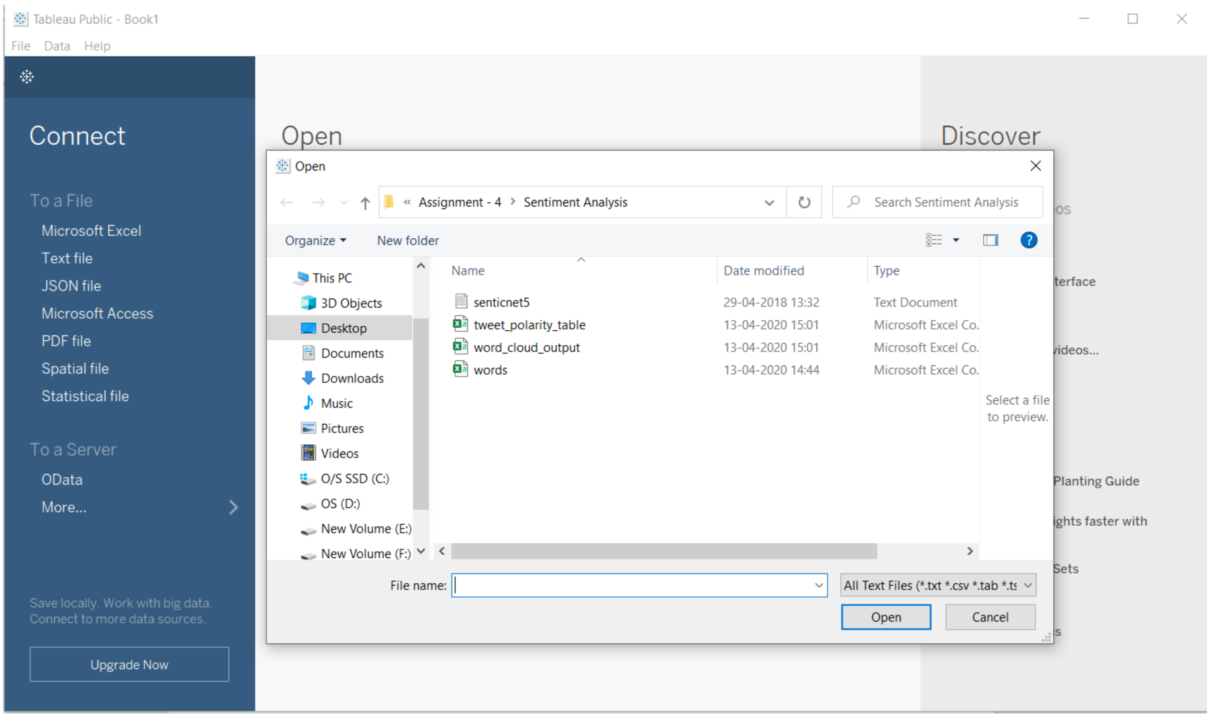


	A	B	C
1	Word	Frequency	Polarity
2	really	16	positive
3	see	120	positive
4	very	36	negative
5	fortunate	1	positive
6	to	1881	positive
7	live	39	positive
8	in	1739	positive
9	damn	2	negative
10	still	76	negative
11	much	56	positive
12	wrong	8	negative
13	the	1974	positive
14	unhealthy	1	negative
15	deserve	4	positive
16	life	19	positive
17	liberty	2	positive
18	shelter	1	positive
19	education	2313	positive
20	free	503	positive
21	leave	16	positive
22	out	209	negative
23	oil	63	positive
24	gas	8	positive
25	group	13	positive
26	for	1162	positive
27	hard	24	negative
28	get	318	positive
29	elected	4	positive
30	dismantle	4	negative
31	healthcare	204	negative
32	have	505	positive
33	serve	46	positive
34	school	171	positive

Figure 5 word_cloud_output.csv

Loading csv to tableau:

Click on File → Text File → Select the CSV file you want to load.



Save locally. Work with big data. Connect to more data sources.

Name	Date modified	Type
senticnet5	29-04-2018 13:32	Text Document
tweet_polarity_table	13-04-2020 15:01	Microsoft Excel Co.
word_cloud_output	13-04-2020 15:01	Microsoft Excel Co.
words	13-04-2020 14:44	Microsoft Excel Co.

Select a file to preview.

Figure 6 Loading data into Tableau

Word Cloud:

All the positive words are shown in green colour and negative words are shown in red colour. And the size of each word depends upon the frequency of word.[9]

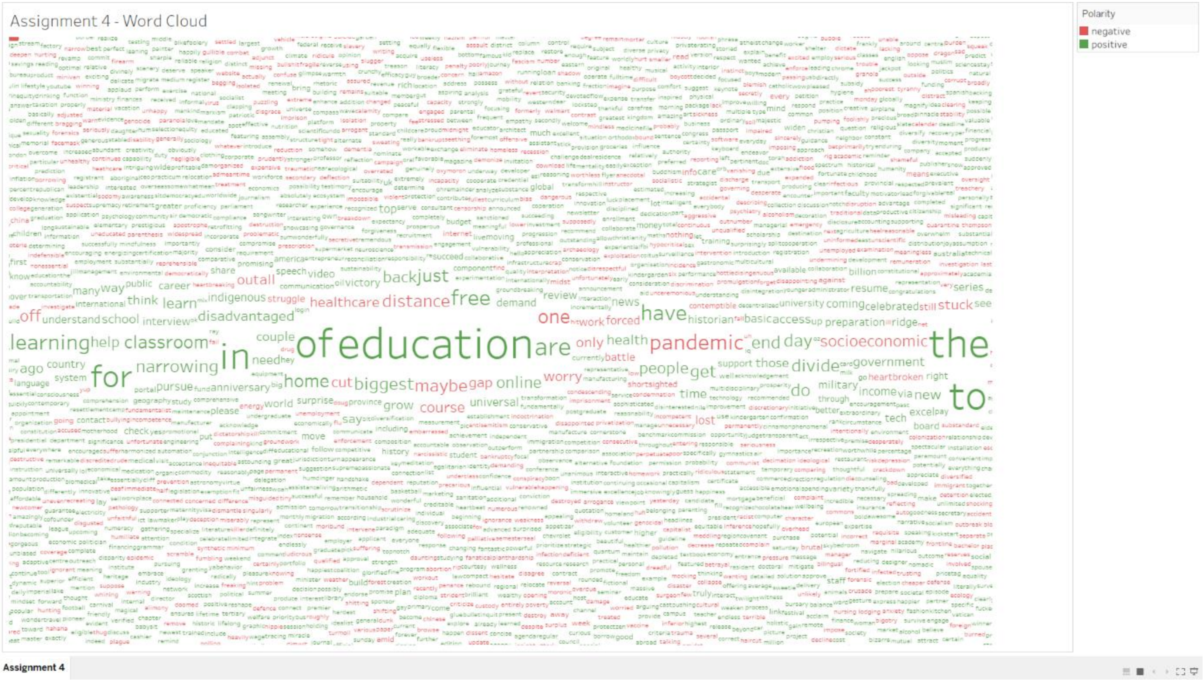


Figure 7 Word Cloud in Tableau

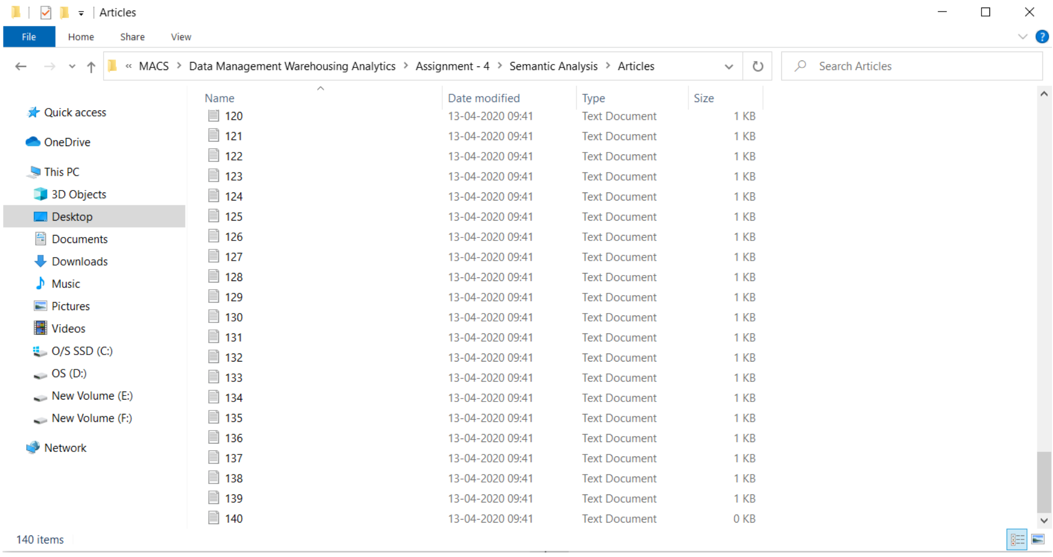
Python script file (sentiment_analysis_tweets.py) can be found in the Script folder.

B. Semantic Analysis:

Following steps were followed to perform semantic analysis on the news data:

Step 1: The news data was collected and cleaned using the python script from assignment 3 from (https://newsapi.org/). In the previous assignment for cleaning purpose all the punctuation, URLs, emojis, Unicode were removed.[5] Python script file (news.py) can be found in the Script folder.

Step 2: Using the python script only the “title”, ”description”, and “content” of the article was collected and stored in different files. Here, there are 140 articles and for each article separate text file is created.

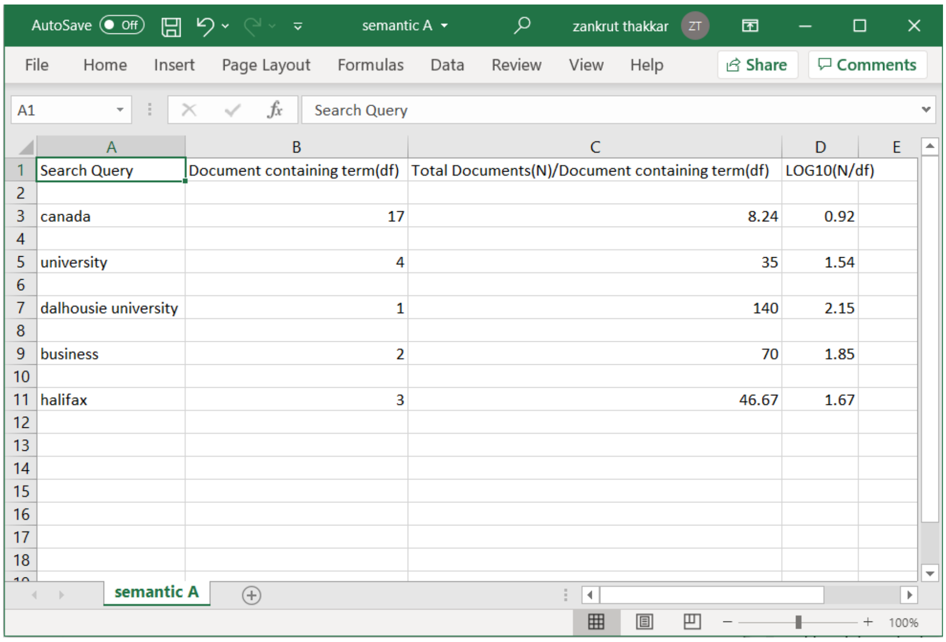


Name	Date modified	Type	Size
120	13-04-2020 09:41	Text Document	1 KB
121	13-04-2020 09:41	Text Document	1 KB
122	13-04-2020 09:41	Text Document	1 KB
123	13-04-2020 09:41	Text Document	1 KB
124	13-04-2020 09:41	Text Document	1 KB
125	13-04-2020 09:41	Text Document	1 KB
126	13-04-2020 09:41	Text Document	1 KB
127	13-04-2020 09:41	Text Document	1 KB
128	13-04-2020 09:41	Text Document	1 KB
129	13-04-2020 09:41	Text Document	1 KB
130	13-04-2020 09:41	Text Document	1 KB
131	13-04-2020 09:41	Text Document	1 KB
132	13-04-2020 09:41	Text Document	1 KB
133	13-04-2020 09:41	Text Document	1 KB
134	13-04-2020 09:41	Text Document	1 KB
135	13-04-2020 09:41	Text Document	1 KB
136	13-04-2020 09:41	Text Document	1 KB
137	13-04-2020 09:41	Text Document	1 KB
138	13-04-2020 09:41	Text Document	1 KB
139	13-04-2020 09:41	Text Document	1 KB
140	13-04-2020 09:41	Text Document	0 KB

Figure 8 Separate text file for each article

Part A:

Step 3: Searched each text file and took count of files in which “canada”, ”university”, ”dalhousie university”, ”business”, and “halifax” are present. Then computed the Total number of documents/ document containing term (N/df). Also, after that $\log_{10}(N/df)$ was calculated using math.log function. Moreover, output was rounded to two decimal points. All the data was inserted in pretty table and converted to output (sematic A.csv) file.

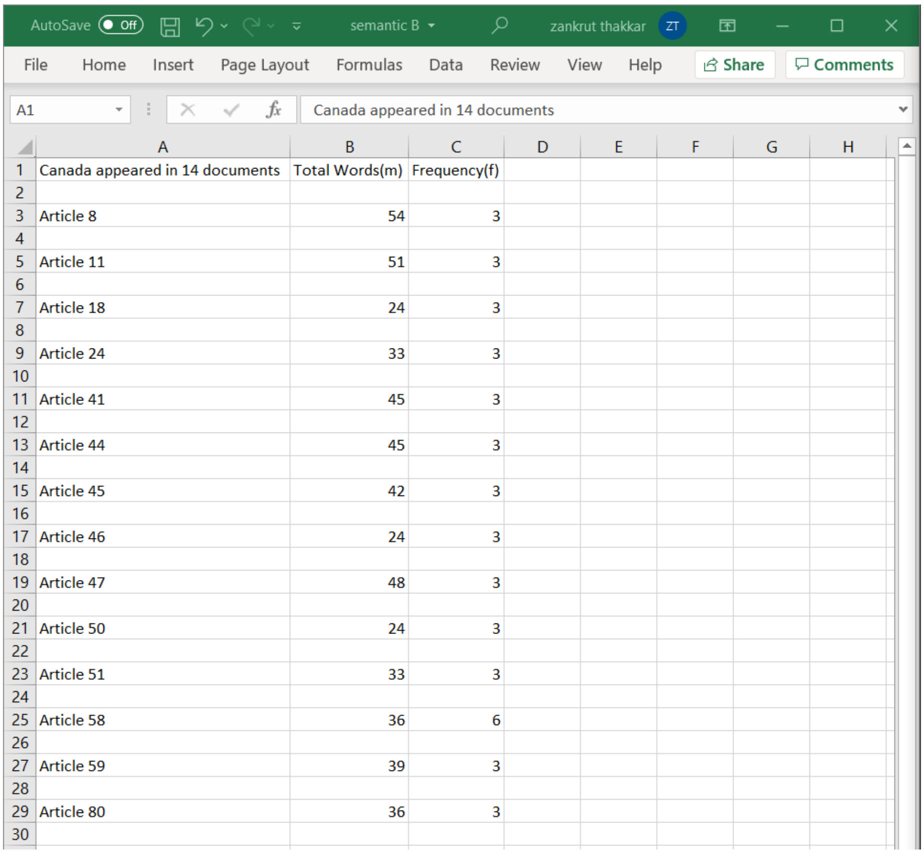


Search Query	Document containing term(df)	Total Documents(N)/Document containing term(df)	LOG10(N/df)
canada	17	8.24	0.92
university	4	35	1.54
dalhousie university	1	140	2.15
business	2	70	1.85
halifax	3	46.67	1.67

Figure 9 semantic A.csv

Part B:

Step 4: After the above-mentioned step each file was searched and counted number of times word “canada” was present in it by appending the data into list and counting word using count() method provided by list. The output was inserted into pretty table and converted to output (semantic B.csv) file.

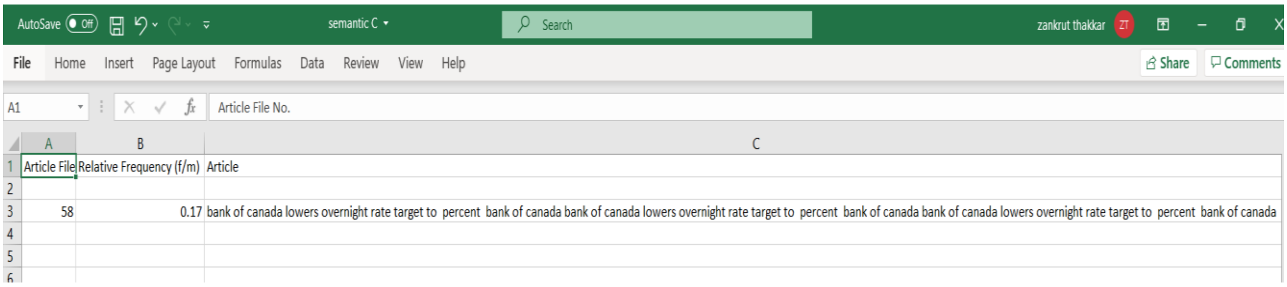


Canada appeared in 14 documents	Total Words(m)	Frequency(f)
Article 8	54	3
Article 11	51	3
Article 18	24	3
Article 24	33	3
Article 41	45	3
Article 44	45	3
Article 45	42	3
Article 46	24	3
Article 47	48	3
Article 50	24	3
Article 51	33	3
Article 58	36	6
Article 59	39	3
Article 80	36	3

Figure 10 semantic B.csv

Part C:

Step 5: After that relative frequency = Total words (m) / Frequency (f) was computed for the articles to in which “canada” was present. And article with highest relative frequency was taken and stored in output (semantic C.csv) file.



Article File No.

	A	B	C
1	Article File	Relative Frequency (f/m)	Article
2			
3	58	0.17	bank of canada lowers overnight rate target to percent bank of canada bank of canada lowers overnight rate target to percent bank of canada bank of canada lowers overnight rate target to percent bank of canada
4			
5			
6			

Figure 11 semantic C.csv

Python script file (semantic_analysis_news.py) can be found in the Script folder. All the three output files semantic A.csv, semantic B.csv, and semantic C.csv can be found in Output folder.

References:

[1] "Downloads « SenticNet", *Sentic.net*, 2020. [Online]. Available: https://sentic.net/downloads/. [Accessed: 13- Apr- 2020].

E Cambria, S Poria, D Hazarika, K Kwok. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: AAAI, pp. 1795-1802 (2018)

[2] “Docs,” *Twitter.com*, 2020. [Online]. Available: https://developer.twitter.com/en/docs. [Accessed: 27-Mar-2020]

[3] "Tweepy Documentation — tweepy 3.8.0 documentation", *Docs.tweepy.org*, 2020. [Online]. Available: http://docs.tweepy.org/en/latest/. [Accessed: 13- Apr- 2020]

[4] "apache/spark", *GitHub*, 2020. [Online]. Available: https://github.com/apache/spark/blob/master/examples/src/main/python/wordcount.py. [Accessed: 13- Apr- 2020]

[5] "News API - A JSON API for live news and blog articles", *Newsapi.org*, 2020. [Online]. Available: https://newsapi.org/. [Accessed: 13- Apr- 2020]

[6]"Python: Count the occurrences of each word in a given sentence - w3resource", *w3resource*, 2020. [Online]. Available: https://www.w3resource.com/python-exercises/string/python-data-type-string-exercise-12.php. [Accessed: 13- Apr- 2020].

[7] J. Bodnar, "Python PrettyTable tutorial - generating tables in Python with PrettyTable", *Zetcode.com*, 2020. [Online]. Available: http://zetcode.com/python/prettytable/. [Accessed: 13- Apr- 2020].

[8] C. line, A. Smith, J. B&#246;cker and A. Klinke, "Convert Python Pretty table to CSV using shell or batch command line", *Stack Overflow*, 2020. [Online]. Available: https://stackoverflow.com/questions/32128226/convert-python-pretty-table-to-csv-using-shell-or-batch-command-line. [Accessed: 13- Apr- 2020].

[9] "Creating a Word Cloud | Tableau Software", *Kb.tableau.com*, 2020. [Online]. Available: https://kb.tableau.com/articles/howto/creating-a-word-cloud . [Accessed: 13- Apr- 2020].