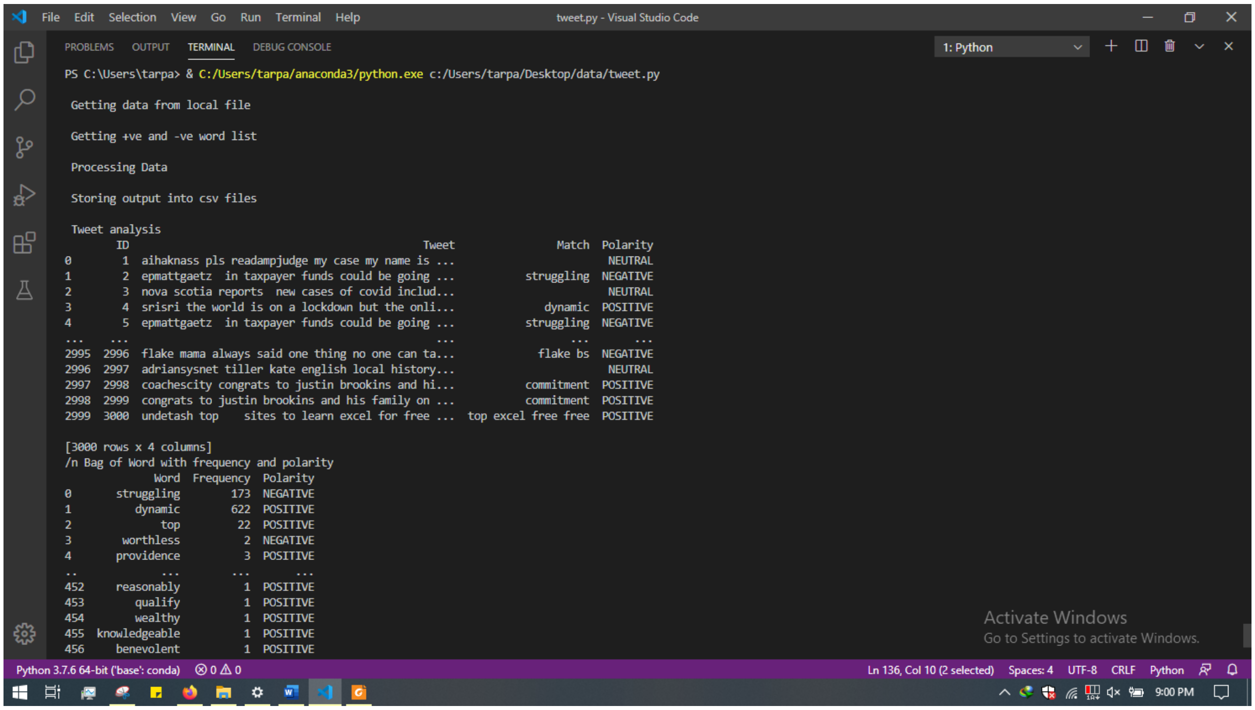
Assignment 4

A. Sentiment Analysis

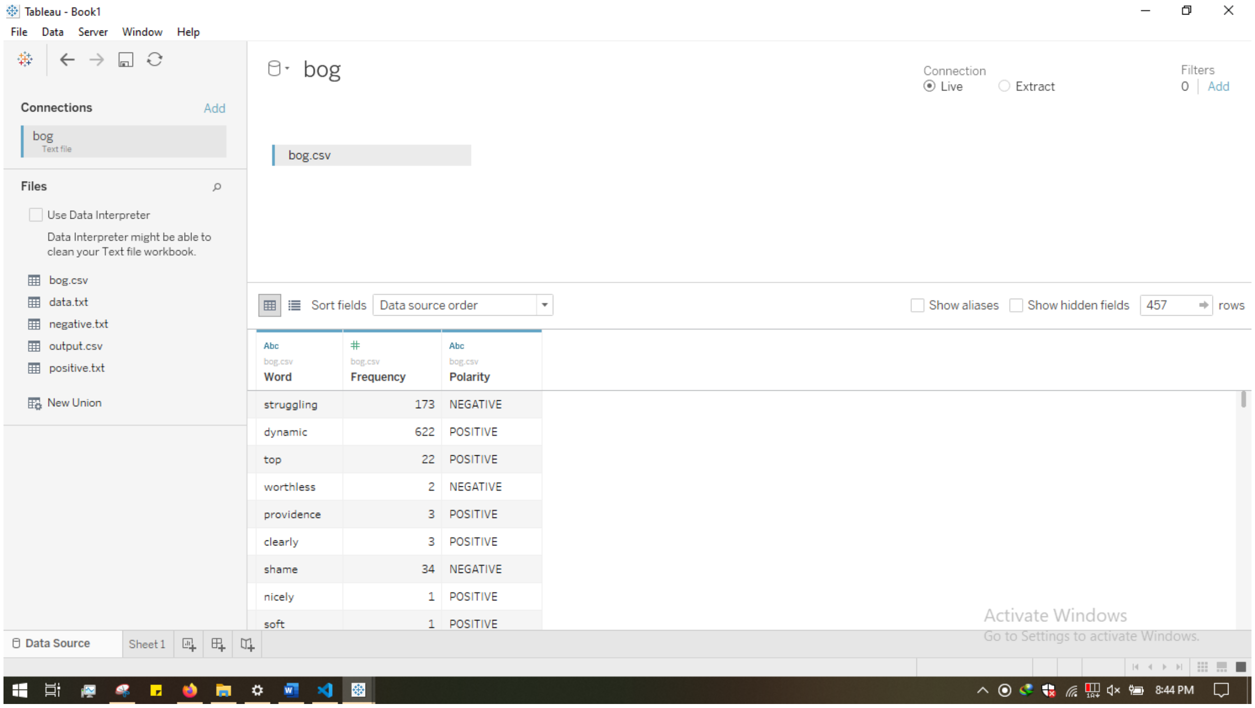
- “Tweet.py” python script fetched data and process them as given instructions.
- “getdata()” function fetched all the tweets then cleans it and store it in a text file – “data.txt”, which is commented in the submitted script.
- Script then reads the cleaned data from “data.txt” and process it as per given instructions.
- Text file of positive and negative words list is obtained from links [1][2].
- Each tweet text is converted into list of words and then compared to positive and negative word list, if the count of positive words is greater than negative word count then tweet’s polarity is “POSITIVE”.
- Similarly, if negative word count is greater than positive word count then it is marked as “NEGATIVE” tweet. If both count is equal, then tweet is marked as “NEUTRAL”.
- At the same time, frequency of each word is recorded into a dictionary – bag of word, which is then converted into dataframe to print it properly.



```
PS C:\Users\tarpa> & C:/Users/tarpa/anaconda3/python.exe c:/Users/tarpa/Desktop/data/tweet.py

 Getting data from local file

 Getting +ve and -ve word list

 Processing Data

 Storing output into csv files

 Tweet analysis
        ID                                          Tweet               Match  Polarity
0        1  aihaknass pls readampjudge my case my name is ...                      NEUTRAL
1        2  epmattgaetz  in taxpayer funds could be going ...          struggling  NEGATIVE
2        3  nova scotia reports  new cases of covid includ...                      NEUTRAL
3        4  srisri the world is on a lockdown but the onli...             dynamic  POSITIVE
4        5  epmattgaetz  in taxpayer funds could be going ...          struggling  NEGATIVE
...    ...                                                ...                 ...       ...
2995  2996  flake mama always said one thing no one can ta...            flake bs  NEGATIVE
2996  2997  adriansysnet tiller kate english local history...                      NEUTRAL
2997  2998  coachescity congrats to justin brookins and hi...          commitment  POSITIVE
2998  2999  congrats to justin brookins and his family on ...          commitment  POSITIVE
2999  3000  undetash top    sites to learn excel for free ...  top excel free free  POSITIVE

[3000 rows x 4 columns]
/n Bag of Word with frequency and polarity
             Word  Frequency  Polarity
0      struggling        173  NEGATIVE
1         dynamic        622  POSITIVE
2             top         22  POSITIVE
3       worthless          2  NEGATIVE
4      providence          3  POSITIVE
..            ...        ...       ...
452    reasonably          1  POSITIVE
453       qualify          1  POSITIVE
454       wealthy          1  POSITIVE
455  knowledgeable         1  POSITIVE
456    benevolent          1  POSITIVE
```

Figure 1: Output of tweet.py

- These dataframes are saved as CSV files named **“bog.csv”** and **“output.csv”**.
- Bog.csv file is loaded as data source in Tableau to create word cloud.



□ Use Data Interpreter

Connection ● Live ○ Extract

Filters 0 | Add

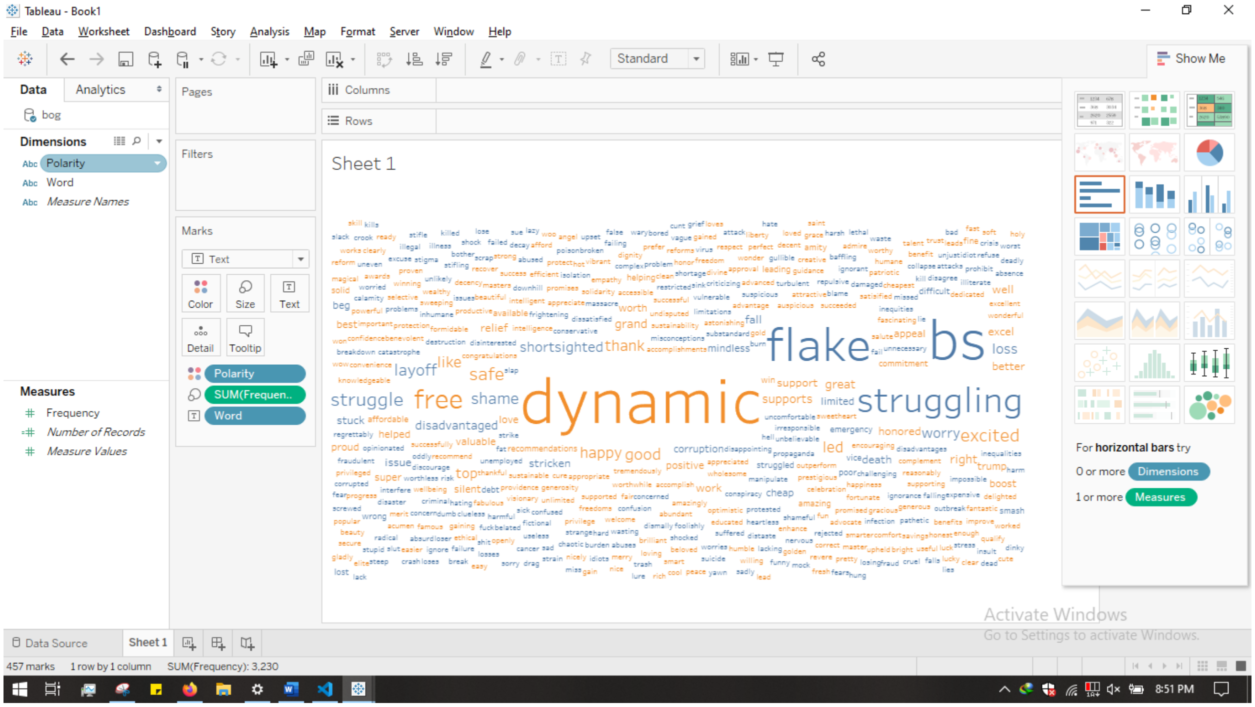
Sort fields: Data source order

□ Show aliases □ Show hidden fields 457 rows

Word	Frequency	Polarity
struggling	173	NEGATIVE
dynamic	622	POSITIVE
top	22	POSITIVE
worthless	2	NEGATIVE
providence	3	POSITIVE
clearly	3	POSITIVE
shame	34	NEGATIVE
nicely	1	POSITIVE
soft	1	POSITIVE

Figure 2: Load csv into Tableau

- In Tableau word cloud, frequency of words is selected as size and polarity as color of text in chart, as shown in image. [3][4]



Marks: Text

- Color: Polarity
- Size: SUM(Frequency)
- Text: Word

457 marks 1 row by 1 column SUM(Frequency): 3,230

Figure 3: Create Word cloud

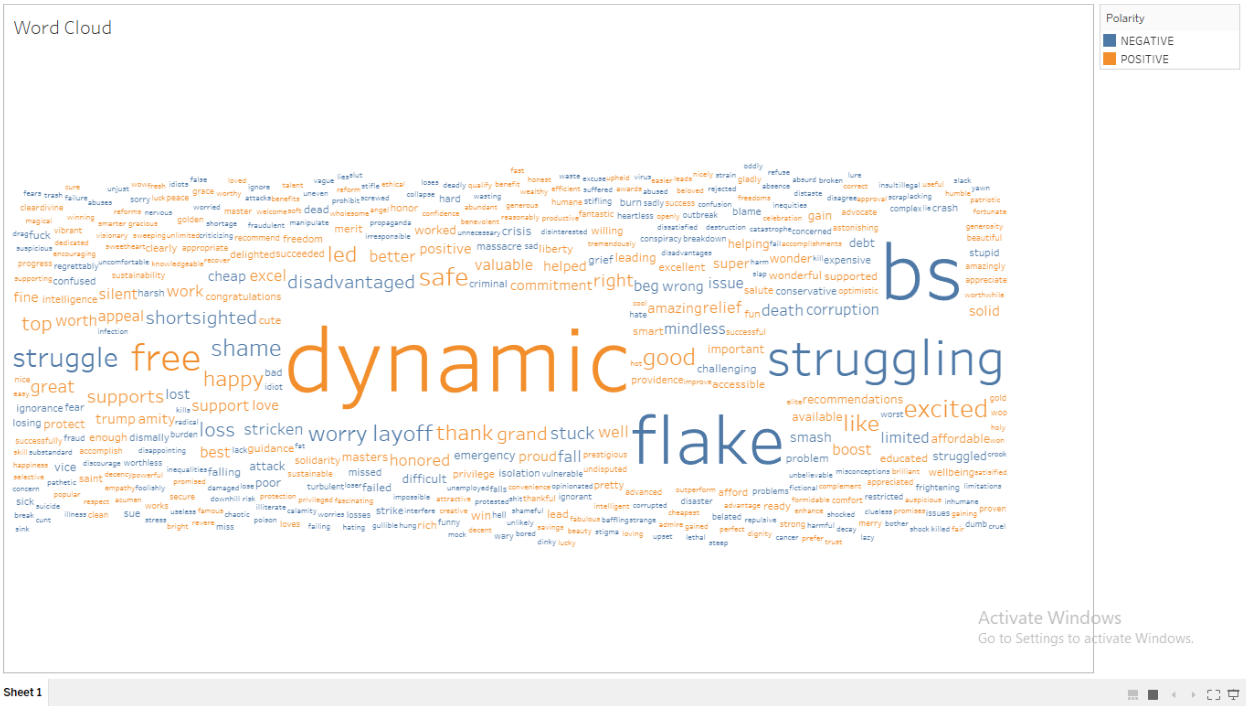
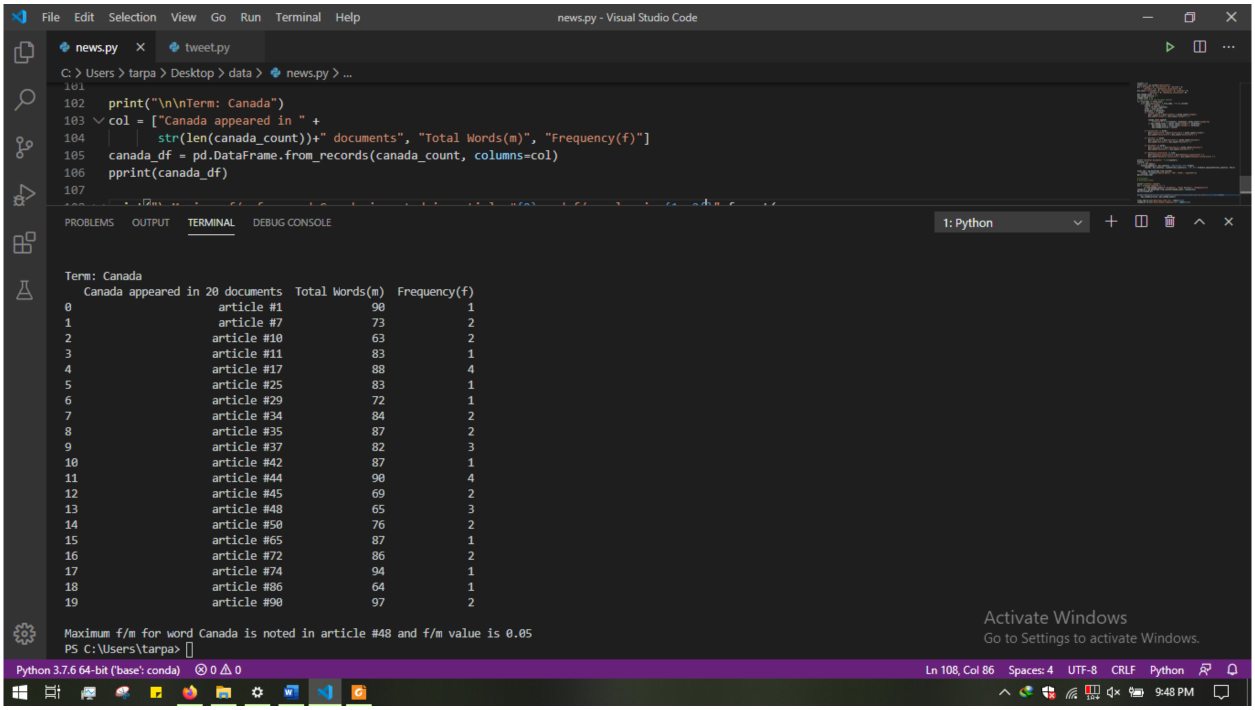


Figure 4: Word cloud

B. Semantic Analysis

- Created a script named “news.py”, that reads “news.json” file. It fetches title, content and description of every news article and converts into one string,
- This string is then cleaned and stored into new file in “news” folder.
- After creating new file for every news article, script reads each file one by one and do the processing as per given instructions.
- Script finds the number of text files in “news” folder and records the frequency of words given in instruction.
- At the same time, it records the article name in which “Canada” appears, and records it in separate list, to create Canada frequency count per document table, as shown in following image.



File Edit Selection View Go Run Terminal Help
news.py - Visual Studio Code
news.py
tweet.py
C: > Users > tarpa > Desktop > data > news.py > ...
PROBLEMS OUTPUT TERMINAL DEBUG CONSOLE
1: Python
Python 3.7.6 64-bit ('base': conda) ⊗ 0 ⚠ 0
Ln 108, Col 86 Spaces: 4 UTF-8 CRLF Python
9:48 PM
Activate Windows
Go to Settings to activate Windows.

```python
print("\n\nTerm: Canada")
col = ["Canada appeared in " +
        str(len(canada_count))+" documents", "Total Words(m)", "Frequency(f)"]
canada_df = pd.DataFrame.from_records(canada_count, columns=col)
pprint(canada_df)
```

```
Term: Canada
   Canada appeared in 20 documents  Total Words(m)  Frequency(f)
0                       article #1              90             1
1                       article #7              73             2
2                      article #10              63             2
3                      article #11              83             1
4                      article #17              88             4
5                      article #25              83             1
6                      article #29              72             1
7                      article #34              84             2
8                      article #35              87             2
9                      article #37              82             3
10                     article #42              87             1
11                     article #44              90             4
12                     article #45              69             2
13                     article #48              65             3
14                     article #50              76             2
15                     article #65              87             1
16                     article #72              86             2
17                     article #74              94             1
18                     article #86              64             1
19                     article #90              97             2

Maximum f/m for word Canada is noted in article #48 and f/m value is 0.05
PS C:\Users\tarpa> 
```

Figure 5: Output of news.py

- Outputs are stored as CSV files named – **“news_tfdf.csv”** and **“Canada_frequency.csv”** that can be found in “data” folder.

References:

[1] *Ptrckprry.com*, 2020. [Online]. Available: http://ptrckprry.com/course/ssd/data/negative-words.txt. [Accessed: 14- Apr- 2020]

[2] *Ptrckprry.com*, 2020. [Online]. Available: http://ptrckprry.com/course/ssd/data/positive-words.txt. [Accessed: 14- Apr- 2020]

[3] "Word Clouds in Tableau: Quick & Easy.", *Medium*, 2020. [Online]. Available: https://towardsdatascience.com/word-clouds-in-tableau-quick-easy. [Accessed: 14- Apr- 2020]

[4] "Thanks for choosing a free trial of Tableau Desktop.", *Tableau Software*, 2020. [Online]. Available: https://www.tableau.com/en-ca/products/desktop/download. [Accessed: 14- Apr- 2020]