# Assignment 4

## A. Data Processing and Sentiment Analysis

Select 1,000 tweets' text from REST key in "twitter_A3_data.json" (retrieved data from assignment 3). Then, cleaning data by removing RT, special characters, and multiple spaces from the text.

Next, create bag-of-words for each tweet. Then, compare each bag-of-word with a list of positive and negative word from [1]. This method is sentiment analysis in lexicon-based approach by creating dictionary and using key-value pair to count the number of occurrences of the word in the sentence.

Then, the program calculates the score and decides whether the polarity is positive, neutral, or negative. If the word is in the list of positive words, the word will multiply by 1 with the number of occurrences of that word. However, if the word is in the list of negative words, the word will multiply by -1 with the number of occurrences of that word. The number from positive and negative is added and result in score. If the score is more than 0, it is positive. If the score is equal to 0, it is neutral. If the score is less than 0, it is negative. The sample output is in Figure 1: sample of tweet_analysis_1000.csv.

| number | tweets | positive match | negative match | score | polarity |
|---|---|---|---|---|---|
| 1 | This is such a shame My son and I are reading A History of Canada in 10 Maps and | | shame | -1 | negative |
| 2 | Canada What part I Dunno just Canada | | | 0 | neutral |
| 3 | It was 10 in Ontario also Canada Why do people NEED to be contrarian You know full | well | | 1 | positive |
| 4 | as some of you know I have a good e-friend from Canada and when he talks to me a | good\|supported | | 2 | positive |
| 5 | 14 New Coronavirus Cases In Canada Including Community Spread | | | 0 | neutral |
| 6 | I m voting for for Presented by | | | 0 | neutral |
| 7 | Teens who engage in sexting get charged Yet an auxiliary member of got a free pass f | free | | 1 | positive |
| 8 | We re thrilled to have support once again this year for the Spirit of the Capital Just a | thrilled\|support | | 2 | positive |
| 9 | I often think about PM Justin Trudeau and how he came a long at the right time How | right\|fortunate\|patience | | 3 | positive |
| 10 | A class action lawsuit would be sweet | sweet | | 1 | positive |

*Figure 1: sample of tweet_analysis_1000.csv*

This logic is flaw as it ignores the negation (e.g. no, not). From [2], WordStat is a sentiment dictionary that measure negative and positive sentiment by negation rule. Therefore, it is positive if the positive word is not preceded by negation or if the negative word is preceded by negation. Moreover, the above logic also ignores the context. For example, "like" can be positive word or conjunction or preposition.

List of files:

- Data: *Twitter_A3_data.json* (from assignment 3), *tweet_text_1000.txt*
- "opinion-lexicon-english" folder: text file with the list of positive and negative words
- Source code: *A1_tweets_clean.py*, *A2_tweets_analysis.py*
- Output: *tweet_analysis_1000.csv*

### Working on Visualization Tool

Before, using Tableau. Data preprocessing is needed to distinct word and count word frequency for each word. After data preprocessing, I import three files and make connection between them. Next, I start doing the word cloud using Tableau. Word cloud presents text visualization for text analysis which can be used in qualitative data such as online survey and product review. The

more frequent the word is used, the larger the display. For example, word cloud can be used to show frequently used keyword which significantly help in creating marketing messages to attract more traffic using popular keyword.

Figure 2 and Figure 3 show the use of positive words in the tweets.
From Figure 2: Positive Words' Word Cloud, “like” appears the most in positive words in tweets. Followed by “work”, “right”, and “good.” As shown in Figure 3: Positive Words' Graph, “like” appears in 32 tweets from 1,000 tweets. While “work”, “right”, and “good” appear in 19 tweets each from 1,000 tweets.

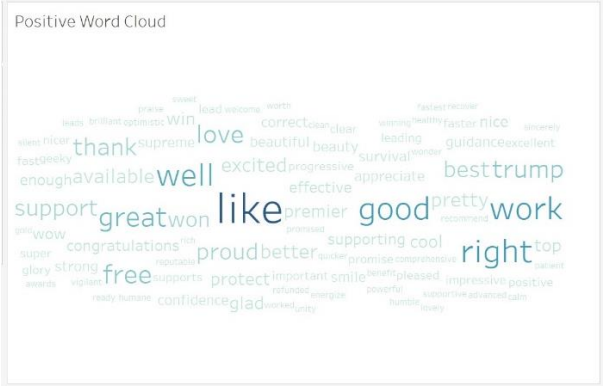


*Figure 2: Positive Words' Word Cloud*

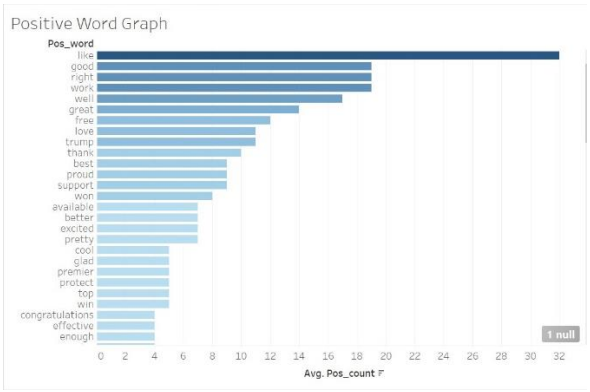


*Figure 3: Positive Words' Graph*

Figure 4 and Figure 5 show the use of negative words in the tweets.
From Figure 4: Negative Words' Word Cloud, “break” appears the most in negative words in tweets. Followed by “epidemic” and “outbreak.” As shown in Figure 5: Negative Words' Graph, “break” appears in 16 tweets from 1,000 tweets. While “epidemic” and “outbreak” appear in 14 tweets each from 1,000 tweets.


*Figure 4: Negative Words' Word Cloud*

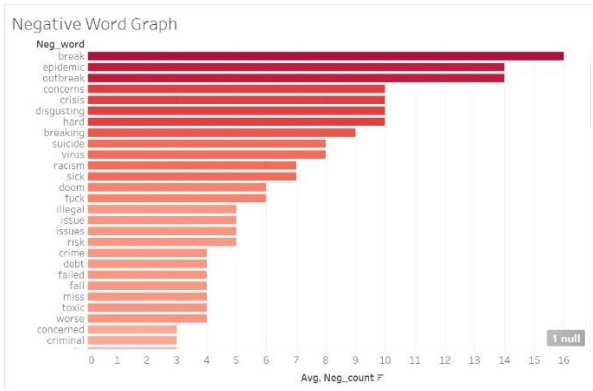


*Figure 5: Negative Words' Graph*

From [3], although word cloud is a powerful tool, it has advantages and disadvantages. Advantages of word cloud are easy to understand and impactful in the presentation. However, there are several drawbacks for using word cloud. First, it only emphasizes the frequency of words, but not the importance of words. Second, it doesn’t provide the context which means the meaning of words may be lost. Therefore, it is more suitable for exploring the qualitative data rather than analyzing complex data. Lastly, if the frequency of words is about the same number, the word cloud is not suitable because all the text will be the same size in text visualization.

List of files:

- Source code: *A3_tweets_analysis_word_count.py*
- Output: *positive_count.csv*, *negative_count.csv*

### B. Data Processing and Semantic Analysis

First, I retrieved the news articles from news API using the specified keywords. Then, extract only "title", "description", and "content" from each news article. Next, write the retrieved data to text file and store in "documents" folder.

From [4], term frequency-inverse document frequency (TF-IDF) is used to evaluate the relevant of keyword in a document in document repository. TF-IDF is done by measuring 2 things which are document frequency and term frequency.

Document frequency measures the number of documents in document repository that has a term appears. As shown in Figure 6: Document Frequency, the retrieved data from news API have documents that has the term "Canada" in 137 documents from 500 documents. "Log 10 (N/df)" column shows that the term is more common in document repository if the value is closer to 0. "Canada" is the most common term from five term in document repository.

| Total Documents | 500 | | |
|---|---|---|---|
| Search Query | Document containing term (df) | Total Documents (N) / number of documents term appeared (df) | Log 10 (N/df) |
| Canada | 137 | 500/137 | 0.56 |
| university | 20 | 500/20 | 1.4 |
| Dalhousie University | 13 | 500/13 | 1.59 |
| Halifax | 49 | 500/49 | 1.01 |
| business | 92 | 500/92 | 0.74 |

*Figure 6: Document Frequency*

Term frequency measures the number of times the term appears in a document that has that term. As shown in Figure 7: Term Frequency, "term frequency (f/m)" column shows that the term appears more in the document when the value is higher.

| Term | Canada | | |
|---|---|---|---|
| Canada appeared in 137 documents | Total Words (m) | Frequency (f) | Term Frequency (f/m) |
| news_1.txt | 101 | 5 | 0.05 |
| news_2.txt | 99 | 2 | 0.02 |
| news_3.txt | 78 | 3 | 0.04 |
| news_4.txt | 61 | 3 | 0.05 |
| news_5.txt | 112 | 3 | 0.03 |

*Figure 7: Term Frequency*

As shown in Figure 8: print news that has the highest term frequency on the console, it is a result of searching the term "Canada" from document repository. The first document (news_31.txt) is about Air Canada which is the Canadian flag carrier and airline. Therefore, if the user searches for "Canada" which means a country, this result is not relevant at all. While the second document (news_44.txt) is about China donates equipment to Canada which is relevant to Canada as a country.

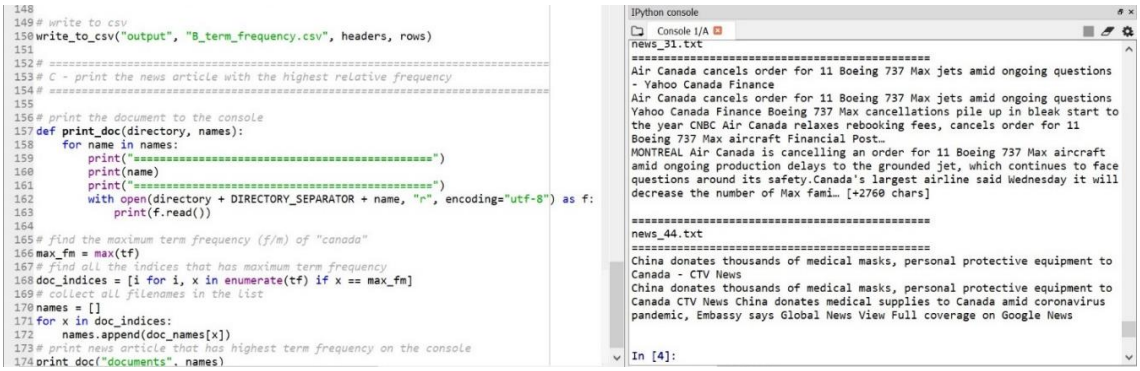


*Figure 8: print news that has the highest term frequency on the console*

TF-IDF is useful for information retrieval and keyword extraction. It helps in document search and deliver the results that relevant to the term. However, using TF-IDF alone is not going to give users the most relevant document because it ignores context. For example, the result in Figure 8 shows that the first document is more relevant to "Air Canada" which is a Canadian Airline, more than "Canada" as a country. While the second document is relevant to "Canada" as a country. Therefore, TF-IDF is suitable to use for screening the documents, but further analysis is needed to find the relevant information by understanding the context of that document.

List of files:

- Source code: *B1_extract_news.py, B2_news_TF-IDF.py*
- Output:
  - "documents" folder with 500 text files (e.g. *news_x.txt*)
  - "output" folder: *A_document_frequency.csv*, *B_term_frequency.csv*, *C_print_doc.jpg*

**Reference**

[1] B. Liu and M. Hu, "Opinion Mining, Sentiment Analysis, and Opinion Spam Detection," 15 May 2004. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. [Accessed 2 April 2020].

[2] O. Davydova, "Sentiment Analysis Tools Overview, Part 1. Positive and Negative Words Databases," Medium, 13 July 2017. [Online]. Available: https://medium.com/@datamonsters/sentiment-analysis-tools-overview-part-1-positive-and-negative-words-databases-ae35431a470c. [Accessed 2 April 2020].

[3] S. McKee, "Presenting Qualitative Survey Data with Word Clouds," Survey Gizmo, 2 June 2014. [Online]. Available: https://www.surveygizmo.com/resources/blog/qualitative-data-word-cloud/. [Accessed 5 April 2020].

[4] B. Stecanella, "What is TF-IDF?," MonkeyLearn, 10 May 2019. [Online]. Available: https://monkeylearn.com/blog/what-is-tf-idf/. [Accessed 6 April 2020].