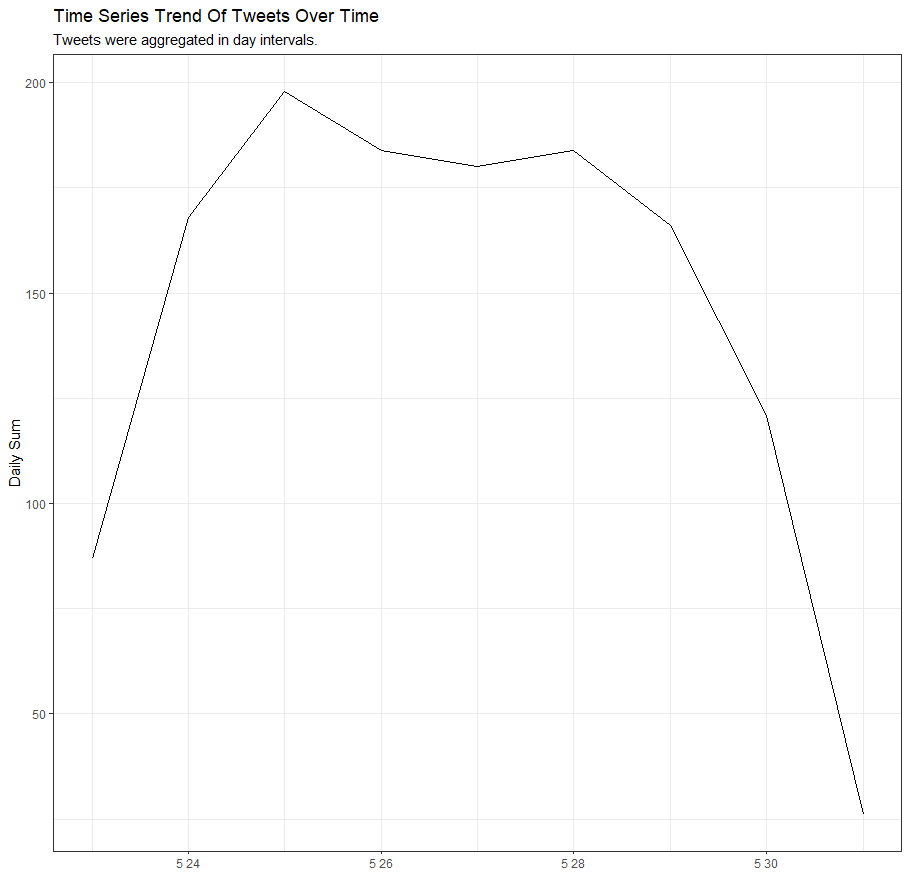
**Second Major Assignment**

20182943 Baek Seonyoung

1. Your R codes used in the workflow of text processing, tokenization, analysis, and visualization (20 points).

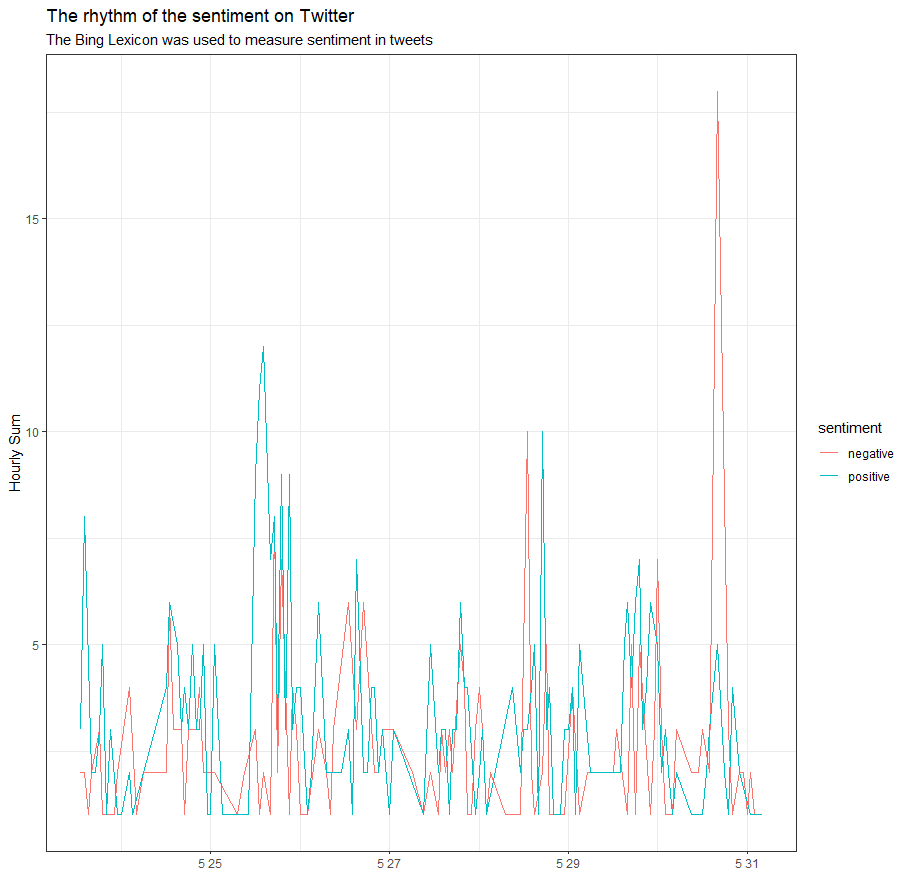
Submitted separately

2. The graph showing the time-series trend of tweets posted over time and your written interpretation of the result (15 points).



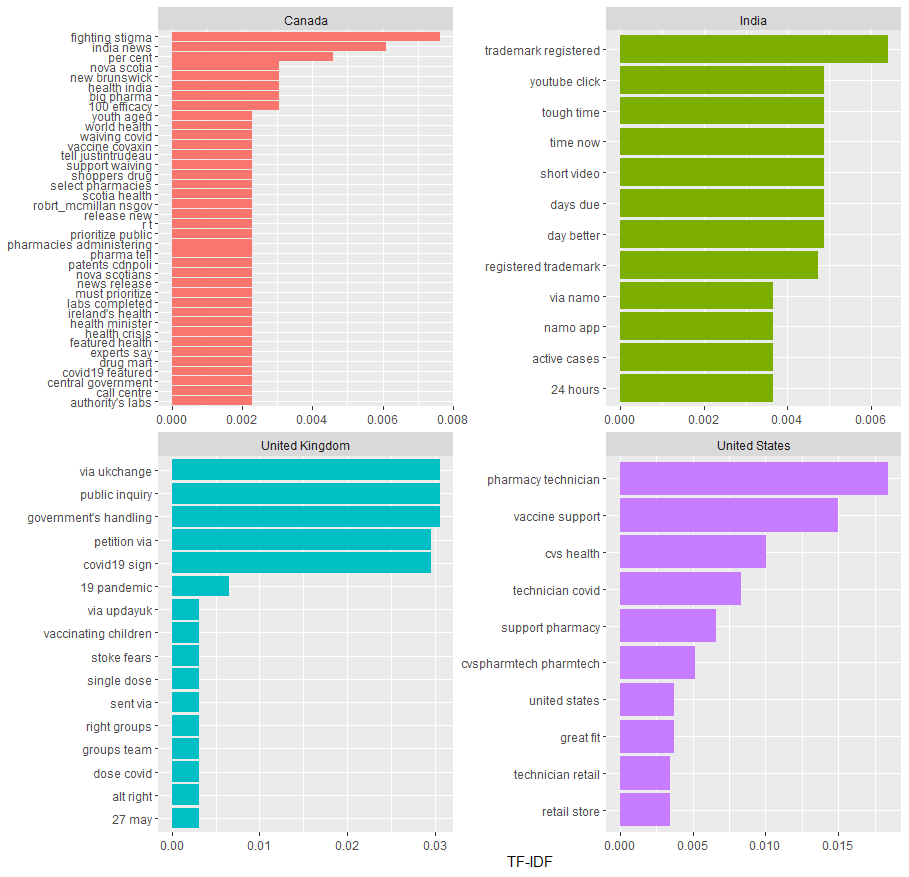
Tweets remained generally high, but highest on 25 May and lowest on 30 May. The levels were also low on 23 and 31, but were not taken into account because only parts of the 24-hours was collected. On the 24th, it was reported that the moderna vaccine proved 100 percent effective in teenage children. It is interpreted that more discussions regarding the Corona vaccine have occurred. In addition, on May 30, there are no social factors that can cause a sudden drop in tweets, but the fact that fewer media articles and fewer tweets are on weekends(especially Sunday) seems to have affected it.

3. The graph showing either the geo-location map or the time-series trend of tweet sentiments in the U.S. and your written interpretation of the result (15 points).



The most notable is the negative sentiment, which soared Sunday evening, May 30. On the evening of the 30th, there were a lot of words such as die, impression, and so on. No direct social issues have been identified with regard to Covid19, but indirectly, concerns about U.S. inflation at the time seem most likely.

4. The tables to compare the most prominent bigrams in terms of TF-IDF among the tweets from the U.S., India, the U.K., and Canada, and your discussion on the similarity or difference in the bigrams among the four countries (15 points).



All four countries show similar levels of tf-idf within their graphs, except for a few bigrams. This means that since it is an English-speaking country, there are not many words of significant frequency except for similar words compared to other countries. Especially in Canada, there are many words that are very similar in tf-idf compared to other countries, and the overall tf-idf level is very low. In India, the specificity of india can be seen through the ‘namo app’. ‘Trademark registered’ and ‘registered trademarked’ are also noticeable, which appears to have been an issue at that time. In the case of the UK, the word ‘via ukchange’ stands out, and many discussions on climate change seem to have been made at the time. In the United States, words such as ‘pharmacy technician’ and ‘vaccine support’ stand out. Those words relative to covid-19 are showed only in America, compared to other countries.

5. Four semantic networks of word co-occurrences in the tweets from the four countries and your interpretation and discussion on some salient topics emerging from the networks (15 points).



In the upper-center cluster, #pondicherry and #pdyfightagainstcorona are linked to #covid\_19 and #vaccination, suggesting that there have been discussions about the inflecton and vaccination of covid-19 in Pondicherry, India. In the lower left cluster, @mohfw\_india (The Ministry of Health and Family Welfare) and lakh (a unit in the Indian numbering system) can be interpreted as the MoHFW's announcement of some performance on covid19 response. In the cluster to the right, discussions on India's prime minister's response to the crisis appear to have been made through ‘Narendra (modi)’, ‘India’ and ‘Minister’. Looking at the largest cluster in the center, it appears that there has been a discussion about the supply of vaccines or drugs (e.g., Tylenol) through ‘cvs’, ‘retail’, ‘store’, ‘health’, ‘application’, ‘pharmtech’, ‘pharmacy’, etc. It is also linked to other clusters such as ‘link’, ‘click’, and ‘just’, which can be interpreted as discussions regarding vaccination reservations and applications have been made. In the leftmost cluster, there may have been discussions on identifying and treating corona infections for migrants through ‘migrated’, ‘curred’, ‘total’, and ‘confirmed’.

6. Your short written responses (less than 10 sentences) to the following questions (10 point each)

6.1) What do you think the most prominent two or three advantages and disadvantages of lexicon-based sentiment analysis?

The biggest advantage of lexicon-based analysis is that it can infer the context in which people are discussing. Word clouds, tf, and tf-idf can give people a sense of what discussions are being made, but there is a limit to inferring how they are being discussed. On the other hand, lexicon-based analysis allows the public to more accurately infer how a particular topic is being accepted and the direction in which the situation is progressing. Of course, the entire context cannot be perfectly taken into account, resulting in some positive analysis, even though it has been negatively used, such as positive used in positive tests, or breakthrough used in breakthrough infections. Although it is not possible to grasp all the nuances of the entire sentence, lexicon-based sentimental analysis is significant in that it can see what trends social issues are moving in broad perspective. However, as mentioned earlier, only single words can be emotionally analyzed, which has a major disadvantage in being unable to understand the context. Therefore, it is necessary to use a combination of bigram or co-occurrences analysis to complement this method of analysis.

6.2) What do you think any meaningful differences between analyzing bigrams and word co-occurrences? Which method do you find more helpful to understand public opinions about COVID-19 vaccines in the four countries?

The difference between Bigrams and word co-ocucurrences (network) lies in the "coverage of how well each word is connected". Of course, the graphs obtained from this project differ in tf-idf or 'did you divide them by each country' etc., but in terms of each analysis method alone, the network is much broader in terms of pairs of words that can see connectivity. The integration of word connections is also better in network because in bigram, the values of ‘A and B’ and ‘B and A’ are different, but in co-occurrences, they do not appear distinctly, but only in lines. In addition, unlike bigram, which requires a numerical view of the degree, network can be seen at a glance by the thickness of the connected line. The fundamental reason for this difference is that in the case of bigram, only consecutive words are viewed, and in the case of network, all of the words represented in one sentence are viewed. Co-occurrences seems much easier to infer, given that more word connections can be seen in understanding public opinion in four countries.