Text Analysis on Social Trends of COVID-19.

Jaeyoung Choi[1]

Library and Information Science,
Sungkyunkwan University, Seoul, South Korea
cjengy@g.skku.edu

Abstract. The outbreak of COVID-19 has changed our lives. This infectious disease was first identified in Wuhan, China, which rapidly spread out to the world, effecting thousands of people in diverse aspects. Due to the influences in the public's health and everyday life, COVID-19 has been accompanied with different word trends. The objective of this project is to determine the social trend associated with the word COVID-19 and Coronavirus. A natural language processing analysis and word cloud visualizations was done to understand the monthly word trend. As a result, from January to May the word focus was mainly about China and East Asia, and from March the trend was altered to global infection, and pandemic. More research and visualizations of the trend regarding further months are required for future work.

Keywords: COVID-19, Crawling, Natural Language Processing, Word Cloud.

1. Introduction

The concerns related to COVID-19 has been amplified throughout 2020. COVID-19 has changed how people behave in various fields, such as areas of health, culture, education and social activities. Also, new terms like social distancing and quarantine has emerged. With these social movements, non contact activities, conferences and lectures, has been emphasized. COVID-19 has caused changes not only in everyday life, but also in the global economic circumstances. Described as an ongoing global crisis, COVID-19 is a continued obstacle to solve. The correlated focus of the pandemic has been modified throughout the migration of the disease. To understand the associated trends regarding COVID-19, we proposed an analysis based on natural language processing.

2. Statement of Goal

Our main objective is to compile the articles from the New York Times online newspaper and then comprehend the word trend regarding COVID-19 to understand how the social flow has changed. We are trying to answer the following question: (1) is there a word trend associated with COVID-19?

3. Methods

The methods section is organized into two sections: Section 3.1 introduces the process of how the dataset was assembled; and Section 3.2. explains the natural language processing analyses and the methods utilized for visualizations of the trend.

3.1. Assembling the dataset

This study uses the headlines and the articles from New York Times. The New York Times is an online American daily newspaper. We collected the content of articles in the duration of 21,January, 2020 to 19, May, 2020. With the utilization of Beautifulsoup from Python, we were able to collect headlines and articles which included the words, COVID-19 or CORONA. We compiled 1134 articles for CORONA, and 796 articles for COVID-19. A total of 1930 articles were amassed.

3.2. Natural language processing analyses and visualization methods

The natural language processing analyses had the following strategy. First, the word preprocess phase, where we made the words of each article to lower cases, deleted unnecessary punctuation, digits, stop-words and single words, and lemmatized the words. Second, to prevent duplicates, we selected unique words from each daily article content. Third, to understand the monthly trend, we combined the daily text into a monthly content corpus. Fourth, to consider only the nouns associated with COVID-19, we extracted the nouns and counted each word. Furthermore, we sorted the words in a descending order regarding the tally. Fifth, we utilized wordclouds for the visualization of each month.

All of the analyses and wordcloud visualization were performed using R version 3.6.1 (R Core Team, 2019). Specifically, we utilized the tm and textstem package to preprocess the words and the wordcloud2 package for graph display.

4. Results

The results section is organized in the monthly order. Each month's wordcloud of interest and bar plot of the word tally, are displayed below.

4.1. Wordcloud and Barplot

This section describes the monthly wordclouds and barplots.

Fig. 1. Wordcloud and Barplot for January

Fig. 1 shows the wordcloud and barplot generated by articles from January. January is when the virus had been first acknowledged. The displayed words are related with the outbreak of the disease, were China and East Asia countries are the main concern. Additionally, the barplot shows the tally of words appearing at January. As the wordcloud, it shows that the main interest is in China, and in words which show the occurrence of the disease such as 'begin' and 'first' have frequently appeared.

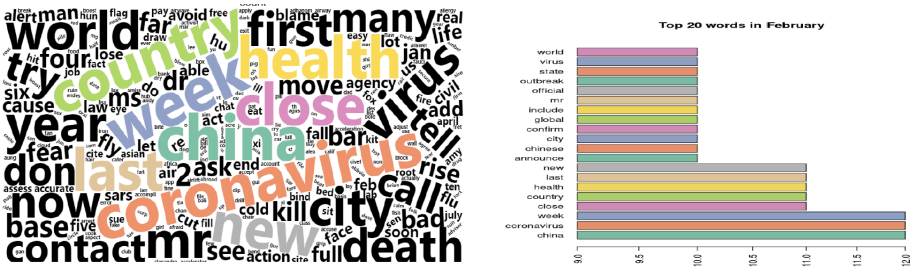

Fig. 2. Wordcloud and Barplot for February

Fig. 2 represents the wordcloud and barplot of February. Similarly with the wordcloud of January, the main interest is in China. But the word 'world' has appeared, indicating the ignition of the global spread. More words regarding the spread was emerging, such as 'move', and 'soon'. Words related to animals like 'bat', and 'fox' show the ignorance regarding the disease's origin. As mentioned in the wordcloud, the words in the barplot are mostly related to China, where the disease was originated from. The overall word tally is similar with January.

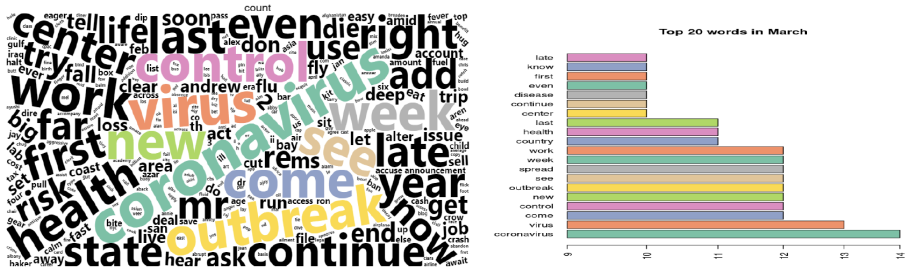

Fig. 3. Wordcloud and Barplot for March

Fig. 3 displays the wordcloud and barplot for March, when the global outbreak was happening. It shows the increased tally among the words: 'world', 'country', and 'state'. We can also see the smaller tally and interest in East Asia countries. As the disease distributed, the word 'health' has appeared with the heightened attention to death. The

barplot shows the outbreak towards the world. More words regarding the global expansion such as 'new', 'come', and 'spread' appear more commonly.

Fig. 4. Wordcloud and Barplot for April

Fig. 4 shows the wordcloud and barplot for April. The word 'pandemic' displays the heightened attention regarding the global situation. The words 'health', 'test', and 'care' seem to appear more often, disclosing the impacts it brought to the public. The barplot shows the more intense concern regarding the pandemic.

Fig. 5. Wordcloud and Barplot for May

Fig. 5 is the wordcloud and barplot for May. The word 'test' has appeared more in the word corpus, implying the severe diffusion of the disease. The overall words seem similar with the wordcloud for April, indicating the global pause due to the pandemic. Additionally, the barplot shows the risen medical impact of COVID-19.

5. Discussion

Is there a trend of words co-occuring with COVID-19? There is a trend that can be seen in the wordclouds and the barplots. The visualization show the movement of the interest shifts from China to a global aspect of the pandemic.

However, as the interpretation of the word trend can be subjective, the trend can be analyzed by different aspects. Additionally, the various methods utilized for preprocessing the words may alter the outcome of the analysis . In detail, the deletion of stop-words, choice of lemmatizing methods, and selection of noun abstraction methods can be an action including subjectivity, which will lead to a difference in the result.

6. Conclusion

The word flow regarding COVID-19 was described by the accumulated data crawled from New York Times. After diverse natural language processing methods, visualization of wordclouds and barplots were constructed. The wordclouds display apparent visual trends, and the barplots show quantitative information of the word tally for each month.

According to the visualizations, January and February show the emergence of the disease, mainly concentrated on China and East Asian countries. But from March, as COVID-19 globally spread out, the word trend indicates more words regarding the world-wide influence. At April, the word 'pandemic' starts to appear more, and is continued until May. With these words is mind, we can conclude that the word trend related to COVID-19 has moved from the concentration of China to the attention of world-wide.

7. Limitations

The interpretation is self-reported, increasing the risk of information bias. Also, some of the word visualization included too much noise. In detail, words like 'dr', 'look', 'even', and 'give' do not provide information about the analysis. However, many of these words were not removed, resulting in wordclouds with several noises included.

8. Future work

It is considered that the inconsistencies of the results might be because of the quantity of the data. Perhaps more accumulation of article data will give more detailed information regarding the social flow of COVID-19. Future analysis with sufficient data will make abundant evidence of the word trend.