Melih Kurtaran

Introduction

The problem is analyzing the news at the first two weeks of December belonging to 2019 and 2020, and comparing the differences in terms of topic distribution over the time. First challenge was to find a suitable news website that can be filtered by date to scrap the news without too much work. Since The New York Times has a very structured interface which users can filter the news by start date and end date, the source will be NY Times. There are many libraries that can be used for web scrapping such as BeatifulSoup, Scrapy or Selenium.

Data Collection

The data collected from nytimes.com has Title, Author, Topic, Date and Link to the article. BeautifulSoup library has been used for web scrapping, 126 entities from the first two weeks of December 2019 and 126 entities from the first two weeks of December 2020 has been collected. Rather than collecting the data in csv file, they are collected into pandas data frame directly. Moreover, each piece of news has been assigned by unique id. For the news from 2019 has an id in a format of 2019xxx and for the news from 2020 has an id of 2020xxx.

Data Processing

The news data already has its topic which is also scraped from the website. However, there is no topic as COVID-19 or pandemic. Therefore, the news categorized as COVID-19 related or not in the data processing part. The keywords selected as COVID, coronavirus, pandemic and mask etc. If a title has one of the keywords, it has been categorized as related and vice versa. Another work done is that if it is USA elections related or not and again some keywords determined as Trump, vote or Democrat etc.

Data Analysis

First, the data frames have been examined describe function and it is seen that there are 38 unique topics for the first two weeks of December 2019 and 34 for December 2020. It has been realized that there are some duplicated titles. However, when it is checked the title is "Coronavirus Briefing: What Happened Today" and belonging to different dates. Then, the topic distribution visualized by different charts. All topics visualized by bar charts and then most frequent ones have been visualized. Later for a better visualization, they have been visualized next to each other by pie charts as seen in figure I. Moreover, average lengths of title has been calculated and it is observed that average length of the titles are 10% longer in 2020 comparing to 2019.

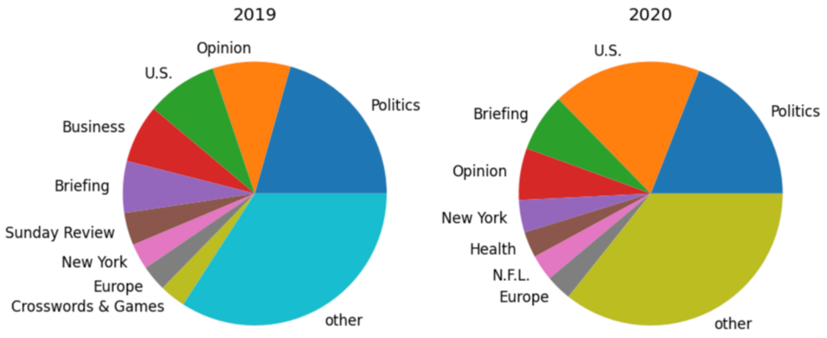


Figure I: Pie charts of topic distribution

Two sets of pie charts have been generated to observe the difference of COVID-19 related news distribution and USA Election news distribution for both 2019 and 2020. There was only one news about COVID-19 in the first two weeks of December 2019 but there are 30 news related to COVID-19 in the first weeks of December 2020 as visualized in figure II. Hence, coronavirus pandemic had an impact on the daily news distribution of New York Times. Another observation is that USA Election related news has been increased from 19 to 28. USA Election held on November 3 in 2020, therefore it was being discussed one year ago at 2019 December, but it continued more, even after one month later than the election.

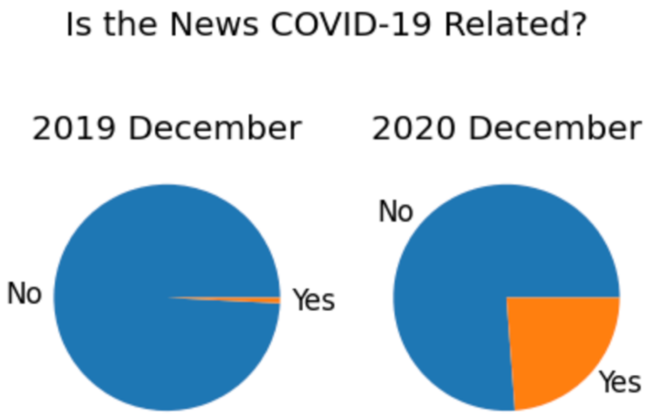
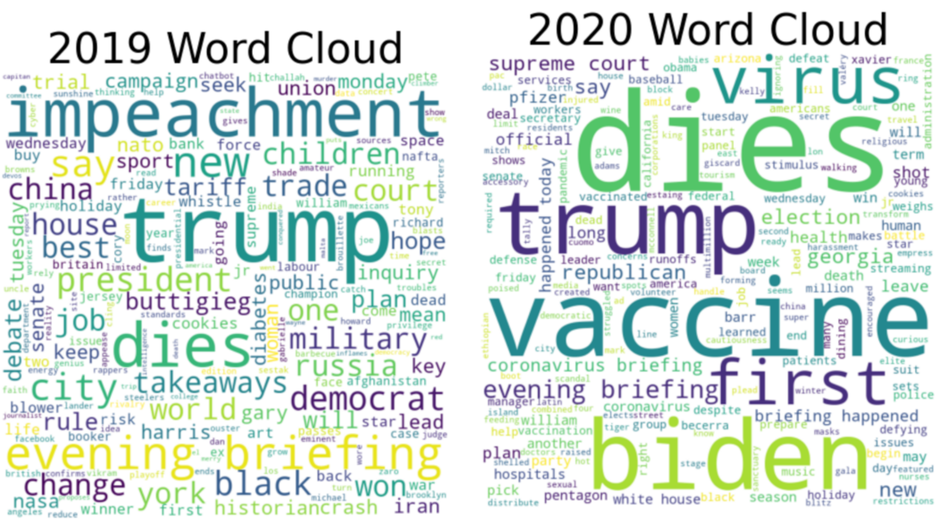


Figure II: Pie charts of COVID-19 Related News Distribution

One other comparative analysis study is implementing a word cloud to observe the popular words in titles both from 2019 and 2020. The word “vaccine” becomes one of the most used word in the first two weeks of December 2020. Because at that time, vaccine against COVID-19 has been produced and countries were starting to vaccinate their citizens. Also, the surname of the newly elected president of USA Biden becomes very popular in 2020 December as it can be seen in the figure III.


2019 Word Cloud

2020 Word Cloud

Figure III: Word Clouds belonging to 2019 and 2020

User Interaction

As part 5 of the project, two demos have been developed for interaction with the user. If a user enters the id of a new, the program returns the link of that specific news. If a user wants to search among the news, the user can enter any word and the program will return all the news with their id and title which are related to keyword. For instance, if user inputs as "Biden", the program will return the news which has "Biden" keyword in the title. However, if user inputs "Bide" then nothing has been returned to the user. The user should enter the words without any spelling mistakes and if the user enters one of the predefined topic news such as Music, Health or Arts then the program returns all the news belonging to that topic.

Conclusion

New York Times website provides only 9 news in a page, to retrieve more news the user needs to click on show more button. Selenium could be used to click to button again and again while web scrapping, but it would make program slower. Instead of using selenium, the website has been scraped with BeautifulSoup by iterating day by day. In that way, 126 news for December 2019 and 126 news for December 2020 has been retrieved.

NY Times has a topic for every news, but it was not providing enough detail to analyze the news and compare with each other. Therefore, the news data has been processed for certain topic relevance. In conclusion, the work has been successfully implemented and can be improved in the future for further analysis.