News Comparison Analysis

Introduction:

News Analytics has grown a lot in developing many solutions like making better business decisions in companies based on statistics and facts, understanding the trends and determining how things happening around us are affecting us. This project aims to compare and analyse the news span between the first two weeks of December 2019 and December 2020 to understand how the focus has shifted from one topic to another over the year.

Data collection:

The data used in this problem is collected from a famous international news agency, 'New York Times. The New York Times website has dedicated pages for every day called Today's Headlines which shows multiple news articles that are separated based on the category of the news. The extracted data includes different news article's title, their author and their category from the period between the first two weeks of December 2019 and December 2020.

Methods:

Web Scraping was used for fetching and extracting the data from the web page. Beautifulsoup, a python package, was installed and used for scraping information from the webpage. The requests function was used to pull the data/source code from the webpage. Lxml parser was used in BeautifulSoup object because of its very fast and lenient.

The base URL used for December 2019 is:

'https://www.nytimes.com/issue/todaysheadlines/2019/12/'

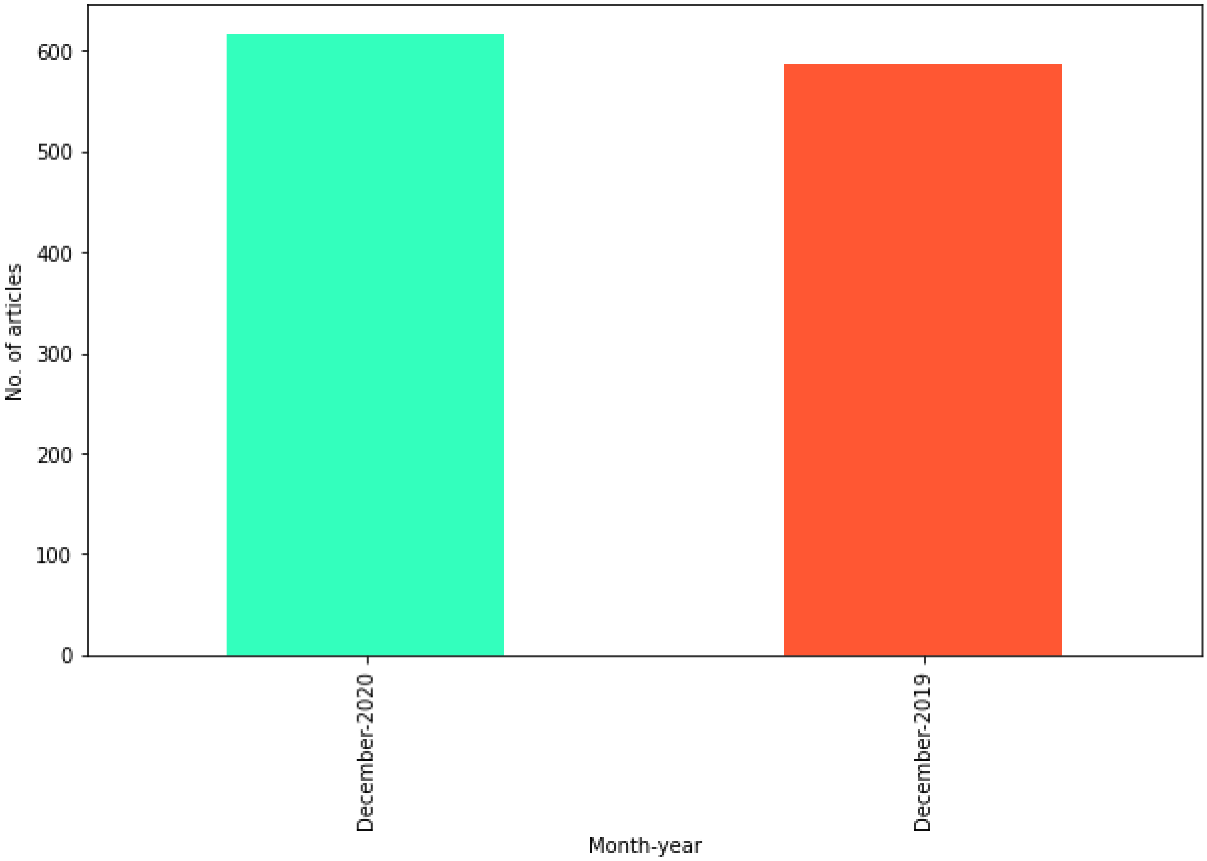
Base URL for December 2020:

'https://www.nytimes.com/issue/todaysheadlines/2020/12/'

Exploratory Data Analysis:

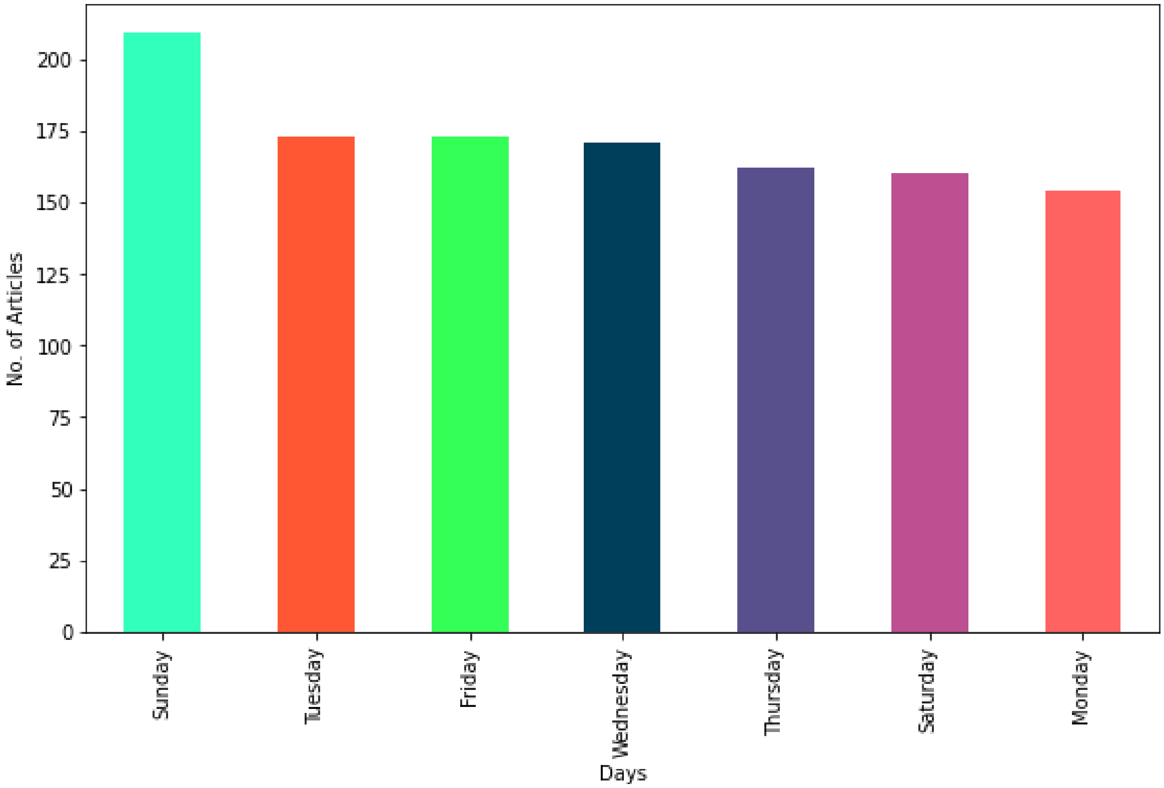
Exploratory data analysis was performed on the collected data to better understand the data and visualize the data to identify patterns or trends in it. In this project, 1202 different news articles in the first 2 weeks of December 2019 and December 2020 were identified. Different visualizations were performed for further understanding and analysis.

1) **Comparing the Number of articles published during the first two weeks of December 2019 and December 2020.**



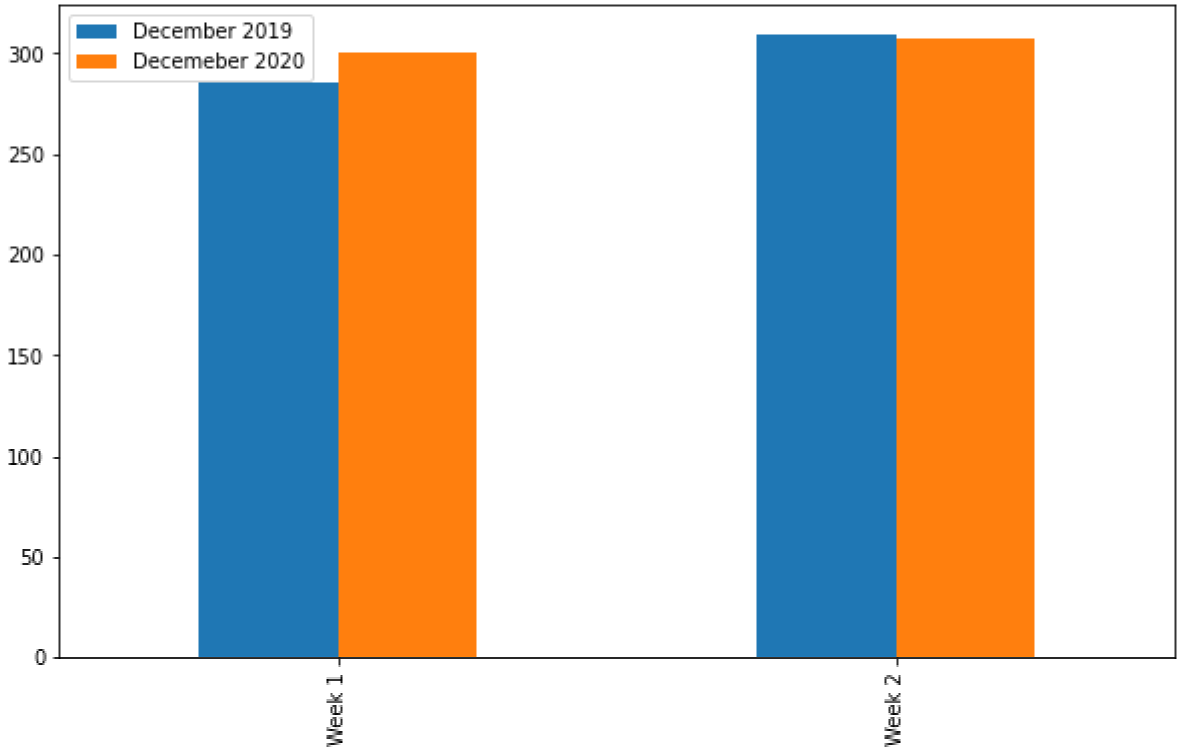
From the above graph, we can see that there is a slight growth in the number of articles published in the next year.

2) Comparing the number of articles published each day for the whole timespan.



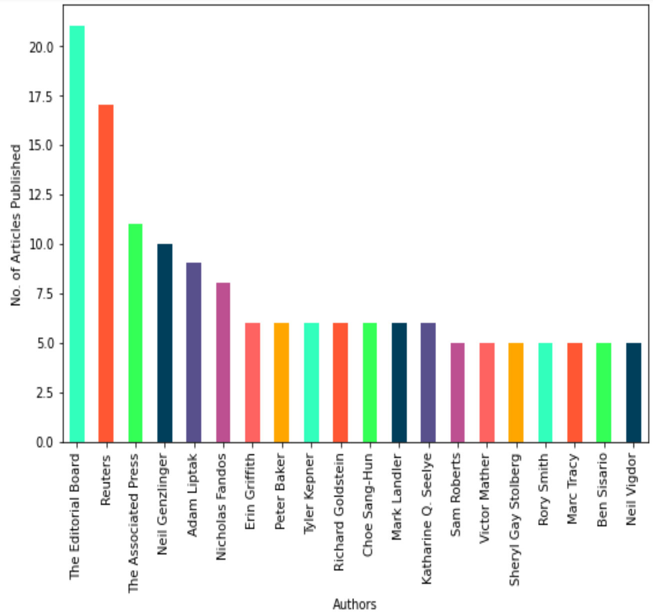
On comparing the number of articles published on a day basis for both the years, it is clear that Sunday has more news compared to other days. And Monday has the least number of articles produced.

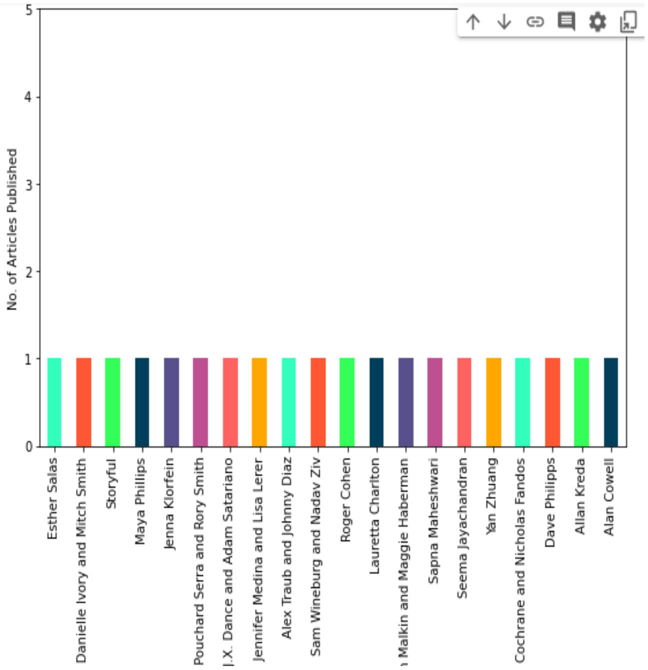
3) Comparing the number of articles published in week one and week two between December 2019 and December 2020.



Taking a look at the first part of the graph, the number of articles published in week 1 is compared between December 2019 and December 2020. The number of articles published in week1 has slightly increased in the next year. Unlike the first part, on comparing the number of articles published in week 2 between December 2019 and December 2020 in the second part, there is a decrease in the number of articles published.

4) Visualising the popularity of the Authors.



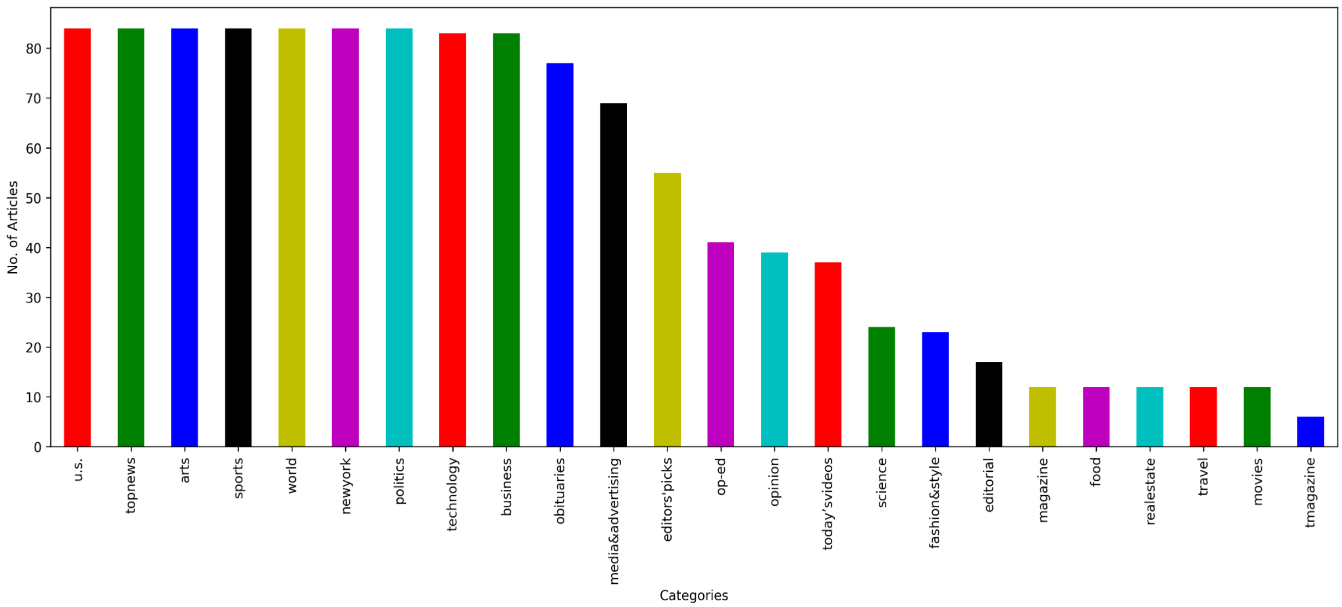


The above two graphs compare the popularity of the news article's authors. The first graph shows the top 20 authors. 'The Editorial Board' is the most popular author with more than 20 published articles. The second graph shows the bottom 20 authors, the least popular authors, who have published only one article.

Comparison analysis study of categories of news articles:

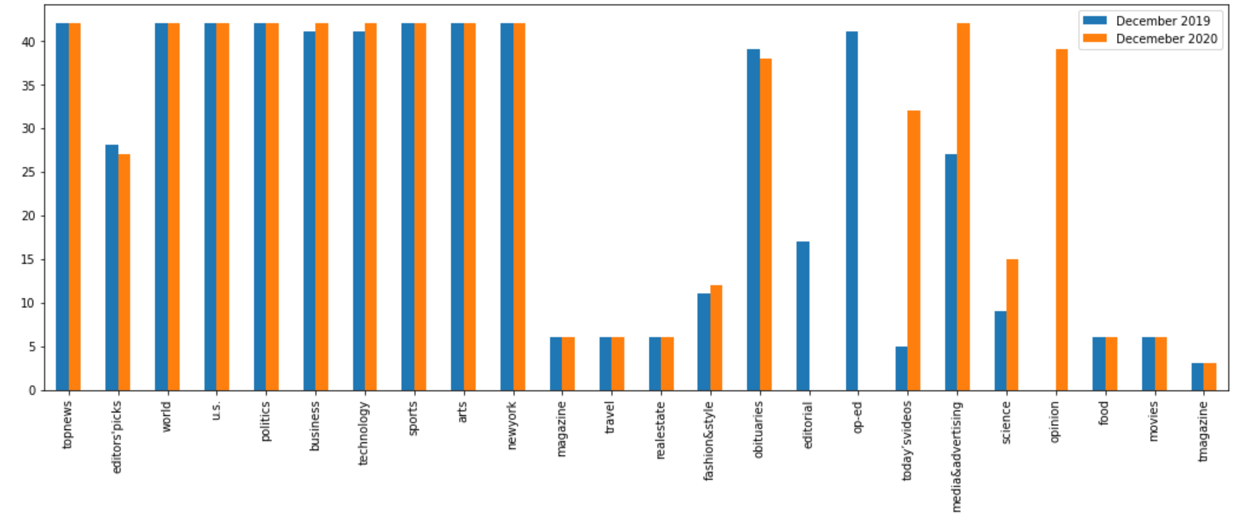
We have already categorized the news articles during the web scraping process now we have visualised and analysed different categories of the news articles to understand their distribution between December 2019 and December 2020.

1) Plotting the popular category of both years



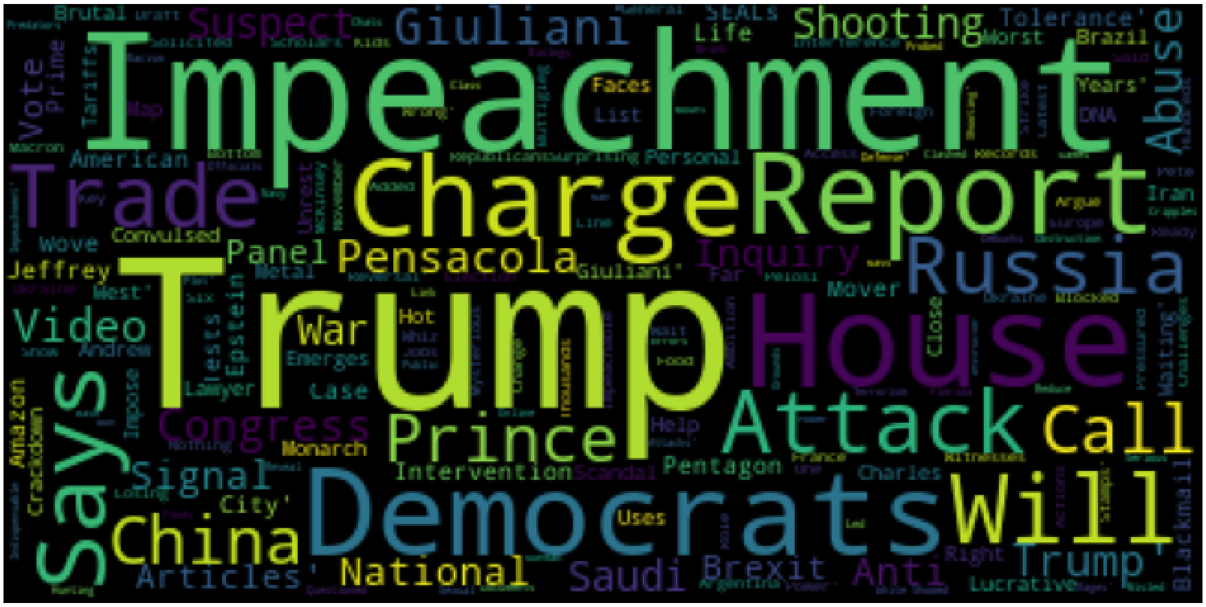
The above bar chart compares and plots the popularity of different categories of news articles for both years. Categories like ‘US’,’ Arts’,’Sports’, ‘World’, ‘New York’, ‘Politics’,’Technology’ and ‘Business’ are the most popular with more than 80 articles throughout 2019 and 2020. Whereas, articles of ‘tmagazine’ have been least popular with less than 10 articles throughout 2019 and 2020.

2) **Comparing the number of articles in each category between December 2019 and December 2020.**

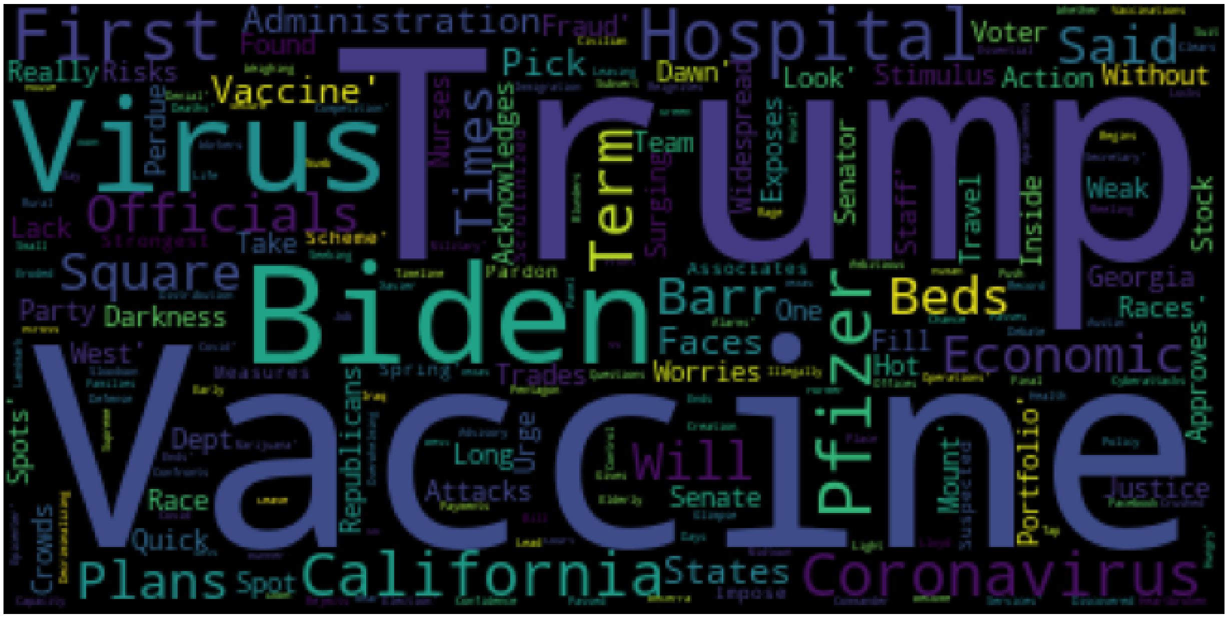


The above bar graph compares each category of the news articles for two different years. Looking at the most popular categories discussed in the previous graph, there isn’t much difference in the number of articles published in both the years. The most significant difference is found in categories like ‘Media & Advertising’ with 28 articles in 2019 and more than 40 in 2020. There are also some categories of news like ‘Editorial’ and ‘Op-ed’ that have been discontinued in the next year. A new category of news called ‘opinion’ have been introduced in December 2020.

3) Visualising the most popular word used in the Top news of the year 2019.



4) Visualising the most popular word used in the Top news of the year 2020.



Analysis of the above two Visualisations:

To deeply understand what and how different the news was in December 2019 and December 2020, the WordCloud technique has been used for extracting the most frequent text in the TopNews articles.

From the first word cloud picture, we can see that 'Trump', 'Impeachment', 'Democrats' are the most prominent topics in December 2019. And all these topics come under the category of 'Politics'. So, it can be said that the category 'Politics' has the highest distribution in December 2019.

From the second word cloud picture, we can see that topics like 'Trump', 'Biden', 'Virus',' Coronavirus' and 'Vaccine' are most prominent in December 2020. We can conclude that in December 2020 most populated news was among the categories of 'Politics' and 'Corona'.

Conclusion

The first bottleneck that I encountered was in collecting the data from ‘NewYorkTimes’. The news web page had the option of filtering the dates of the first 2 weeks of December 2019 and December 2020. But the output showed only the first 11 news articles and a ‘Show More’ button, the articles loaded after clicking on the ‘Show more’ button cant be scraped using the BeautiSoup tool since BeautifulSoup does not have a click function. Later, I found that NewYorkTimes also have a dedicated webpage called Today’s Headlines for each date summarizing different news articles based on different categories.

The second problem was figuring how to collect and organize the data so that it can be easily visualized at further stages. This was solved by collecting and storing the data in a CSV file and reading it in pandas, which is a great data frame for representing and analysing data. At last, I had to compare and visualise the data so that some meaningful full results can be gained from it.

Overall working on the project was a great experience as it taught me how to collect and analyse data so that we can have some meaningful results.