News trend during COVID-19

By Team **^(?=no one)us.*$**

Data used:
- https://www.kaggle.com/jannalipenkova/covid19-public-media-dataset/data# (over 50,000 news articles with full text scraped from online websites from January to April)
- GDELT Project dataset (around 80,000 news articles per day, only themes and url, not full text)

Using the above news articles datasets, our group explored the social impacts brought by COVID-19 throughout the months of January to April.

From a basic, overall analysis, we noticed that keywords like "trump", "pandemic", "china", etc. appear frequently, and a network graph is centered at "coronavirus", meaning it frequently appears together with other words. We also notice an obvious shift in focus of articles throughout the period in response to events like COVID-19 spread, outbreaks, public announcements, testing, and other government plans.

Using the sentimentr package, we notice that the mean sentiment score of news articles became higher (more positive) during March, as shown in the figure. We verified this difference to be significant by ANOVA and determined the cause to be a group of strongly positive news flooding in. So we decided to investigate this change: **why is news sentiment becoming more positive as the situation becomes worse and worse? What is this group of positive news? Where do they come from?**

Going through the same basic analysis, we notice a lot of "announce" and business-related terms, and we already see positive words like "support" and "expand". Moreover, the new network graph is centered at "announce" and surrounded by many positive words. This suggests the rise in positive sentiment in March is due to organizations and government taking actions and planning to overcome the worsening situation.

Next, we vectorized the data, reduced its dimension using PCA, and performed k-means clustering. Over time, old clusters like "China", "cruise", "airlines" disappear and new ones like "action" and "medical" show up. The sizes and centering of clusters tell us what topics are trending and agree with our previous sentiment analysis.(cool interactive clustering plot: https://linus810.github.io/interactive/)

From a quantitative perspective, we first looked at the trend and composition of general themes like: disease, employment, government, etc. Second, we filtered the themes that increased and grouped them by how much they increased. This reveals how strongly each was affected by the pandemic.

We conclude with two ending stories: COVID-19 "brings peace" to the views on war, weapon, terrorism due to decreased news frequency referencing those topics, and it unites people, as shown by decreasing uses of "I" and increasing uses of "we".

References:
https://python-graph-gallery.com/13-percent-stacked-barplot/
https://python-graph-gallery.com/124-spaghetti-plot/
https://www.kaggle.com/jannalipenkova/covid-19-media-overview
https://www.kaggle.com/maksimeren/covid-19-literature-clustering
https://www.kaggle.com/zhu701/covid-19-cbc-news-exploration
https://www.tidytextmining.com/