Team Members: Alfredo Gomez, Bryan Uribe

Question:
How are various news outlets talking about COVID-19?

Data:
This would involve scraping some of my data and can rely on some old code of mine to do this. As a preliminary, here are some links we would have as a starting point:

https://www.foxnews.com/category/shows/the-five/transcript
http://transcripts.cnn.com/TRANSCRIPTS/

Web Scraping will most likely be done though Selenium through a google browser history.
We would begin by considering a small time frame (a few days) and then extend our scraping once we find we need more data for further analysis.

Methods:
Once we have our data, we will find some preliminary statistics regarding the sources. Initially we would like to start with a count of words they are using to talk about the virus. i.e. instance of “Coronavirus” instead of “COVID-19”.

The next steps would be to run different NLP algorithms in order to gain some insight about words and sentiment used throughout the transcripts.

Possible NLP algorithms:

Latent Dirichlet Allocation (same as tweets) - topic modeling. The medium article is a good review of how the technique works.

LexRank - A process to summarize text by finding highest ranking sentences. This finds the sentence that most closely relates to other sentences in the article. The highest ranking sentence is assumed to be of greatest importance

We can use PCA to confirm how well LDA recognizes topics.

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d
https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c

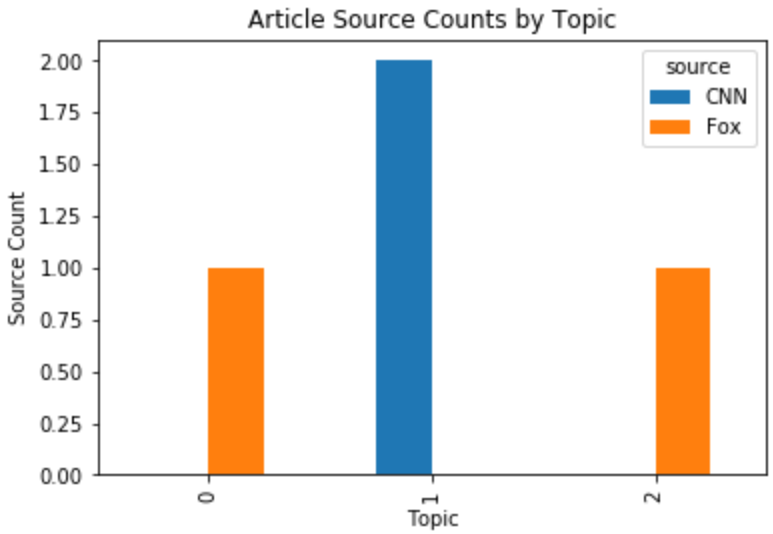
Initial Test:
As a starting point, we took 4 transcript from CNN and Fox from the following sources

http://transcripts.cnn.com/TRANSCRIPTS/2005/04/acd.01.html
http://transcripts.cnn.com/TRANSCRIPTS/2005/05/acd.01.html
https://www.foxnews.com/opinion/gutfeld-on-limiting-coronavirus-lawsuits
https://www.foxnews.com/opinion/gutfeld-trump-media-covid-pandemic

This was mostly a test to get everything working with applying topic models to our own data. After running the topic model, we obtained the following set of topics

Topic 1: ['post', 'rising', 'need', 'think', 'maybe', 'starting', 'means', 'didn', 'hospitals', 'safety', 'leave', 'hand']
Topic 2: ['going', 'like', 'president', 'said', 'got', 'know', 'time', 'day', 'right', 'think', 'mean', 'pandemic']
Topic 3: ['like', 'trump', 'question', 'worse', 'run', 'appreciate', 'sure', 'agree', 'sunday', 'hall', 'fox', 'press']

Counting them by transcript source provides us with the following plot



Although our sample size is quite small it is interesting to see that both CNN articles got labelled the same topic. A key word I noticed in the topic was “president” in contrast to “Trump” being utilized in the third topic.

Questions:
One thing that crossed my mind while making this topic model and the data we collected thus far is that one is a shortened excerpt from a broadcast while the one consists of an entire segment, hence the length is fairly inconsistent between the two sources. Are there any ways we could “correct” for this issue? Or would it be better to look for slightly more similar “segments”?

Web Scraping:
The web scraper visits the website below and collects the text of each article in a pandas dataframe which is then exported to a csv file.

Functionality I need to add:
Use Selenium to press the “show more” button so that we webscrape more than just the 5 most recent articles.

A downside is that it collects all unwanted text, such as picture captions.
https://www.foxnews.com/category/shows/the-five/transcript