**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