W200 - Project 2 - Proposal
Proliferation of Fake News
Github Repo: https://github.com/UC-Berkeley-I-School/mids-200-project-2-Kenz-JM
JM Stilb, MaKenzie Muller

Datasets:
Information is useful and dangerous. Additionally, the complexities of today's information ecosystem are not well understood despite their severe consequences. We plan to explore several of the complexities in this ecosystem to gain a better understanding of how certain distribution dynamics can create a feedback loop and how influential figures can amplify certain information. Specifically we will be analyzing the symbiotic relationship between conservative news outlets such as Fox News and our former President Trump, and furthermore, how an influential figure like President Trump amplifies certain news topics.

Primary

- Trump tweets - https://www.kaggle.com/austinreese/trump-tweets
- News Articles - https://www.kaggle.com/snapcrack/all-the-news?select=articles1.csv

Supplemental

- Google Trends - https://trends.google.com/trends/?geo=US
    - Join on Date columns
- Twitter Trends - https://storywrangling.org/
    - Join on Date columns

Data Structure:

Trump's Tweets
Trump's tweets from May 2009 through mid June 2020 were scraped from Twitter and consolidated into a .csv file of ~43,000 rows. Each row has the following column attributes:
- Tweet ID (a unique identifier for the tweet)
- Link (link address to the tweet)
- Content (the text written in the tweet)
- Date (the date the tweet was sent, in a MM/DD/YYYY HH:MM format)
- Retweets (the number of times the tweet was retweeted)
- Favorites (the number of times the tweet was 'favorited' or 'liked')
- Mentions (the twitter username mentioned in the tweet, denoted with an @Username format)
- Hashtag (any hashtags used, denoted with #Hashtag format)

News Articles

News articles from popular publications including CNN, Fox News, and others were scraped and consolidated into 3 separate csv files of ~ 145,000 rows. Each row has the following column attributes:

- Id
- Database id (unique identifier for the database the data was pulled from)
- Article title
- Publication name (the name of the media network)
- Author name
- Date of publication (the date the news article was published)
- Year of publication
- Month of publication
- Url (link to the news article)
- Article content (the copy from the news article)

Initial Exploration:

- Line Chart: Growth rate for different news topics
  - Indicate when Trump began tweeting about it
  - Indicate relevant events
- Line Chart: Growth rate (rate of decay) of useful information compared to dangerous information
  - Indicate when Trump began tweeting about it
  - Indicate relevant events
- Bar Chart: # Times the news outlet was the origin versus Trump
- Bar Chart: # of tweets about news topic compared to number of articles from varying news outlets

Plan for Final Report:
Here are the topics/questions we plan to address in the final report:

- How do symbiotic relationships affect the propagation of information?
- Where does dangerous information originate?
- How quickly can dangerous information be amplified?
- Can influential figures dictate what is spoken about on politically-unaligned networks and social media?
- What happens to important information when dangerous information is amplified?