DATA 620
Final Project Proposal

Authors: John Kellogg, Jeff Shamp

Spring 2021

Goal: Predict the speakers of and analyze the similarities between US Presidential speeches with a focus on the top 5 most eloquent speakers.

Data Source: Kaggle (Joseph Lilleberg) gathered the data for us. With the size of this project, we chose an already gathered dataset vs a lengthy web scrape. (https://www.kaggle.com/littleotter/united-states-presidential-speeches)

Methodology:

- Prediction: Using feature extraction coupled with the bag of words technique we hope to train the model to predict the speakers.
    - Can we train a model to detect who wrote the speech?
    - Is it effective when only using modern speakers?
    - Do we need to use speakers from different eras for the model to have real chance in accuracy?
- Analyzation: Using Network mapping we hope to show the similarities in the speeches by word usage.
    - Does word usage change over time?
    - Are their common words or themes in all of them?

Responsibilities:

- John will focus on the prediction models
- Jeff will focus on the Network models
- We will pull together to finalize the report

Concerns:

- **Prediction:** We as a people speak differently than we write and we tend to follow similar speech patterns. Our initial concern is the model being ineffective in accurately predict modern speeches. We may have to bring in speeches from different eras to boost the accuracy of the model.
- **Analyzation:** As stated above, we tend to follow similar speech patterns, while those maybe useful to present, there may not be a huge separation in the network.