**Project Scope**

The project design is to utilize NLP techniques to preform data mining, determine term frequency–inverse document frequency (TF-IDF) values, latent Dirichlet allocation (LDA) estimations, topic modeling, and sentiment analysis of 101 State of the Union addresses from 1791 to 2019.

**Desired Outcome**

Sentiment analysis, topic modeling, TF-IDF and LDA values to derive deeper insights of American politics through the centuries and deepen understanding of NLP processes and results.

**Corpus Development**

Corpus is to be developed from SOTU addresses published to the [State of the Union website](http://stateoftheunion.onetwothree.net/index.shtml). A scoped down assortment of all 243 files was used for speed and simplicity.

**Model**

The NLP modeling will incorporate a variety of scripts and/or Jupyter notebooks from the MSDS 453 Winter 2019 course, those discovered on GitHub, and the [SOTU Kaggle website](https://www.kaggle.com/rtatman/state-of-the-union-corpus-1989-2017).

GitHub credits:

Daniel Bashir, [**db7894**](https://github.com/db7894) **/** [**sentiment-of-the-union**](https://github.com/db7894/sentiment-of-the-union)

Shayne, [**shngli**](https://github.com/shngli) **/** [**SOTU-mining**](https://github.com/shngli/SOTU-mining)