**Presidential Data Analysis**

**Introduction**

There is a vast amount of public data about everything every president has done, but there is not as much in the way of analysis and application of this data. The existing data is mostly separated by President, making it harder to observe similarities and patterns. The goal of this project is collect all of this available data about our presidents, which will enable us to analyze the speeches and elections of Presidents from Washington to Trump. These insights will allow us to gain a greater understanding and context for each president, in addition to their respective political climates, i.e. how the American public was feeling. It will also give us better insight as to how each candidate won their election.

**Proposed Work**

The American Presidency Project is a non-profit, non-partisan website dedicated to archiving presidential documents. We will be scraping through their data archives and downloading all data they have to offer about past presidents. We will spend a part of this project sifting through the data that we obtain and determining what to keep for analysis. We will have to do a lot of preprocessing to get the data, some of which is unformatted, into a form usable by python and Pandas.

We have particular interest in the state of the union addresses given every year by the president. We believe that analysis of the words used in each address will grant us insights into how the American public felt and what events happened during each corresponding year.

We found a project online that builds Markov chains using words from Donald Trump’s twitter and generates random tweets in the style of Donald Trump. We plan to research this project and making markov chains of our own trained on speech data for different presidents. We think we will be able to generate create parts of new speeches based on the training data from the old speeches.

We plan on doing more EDA on the state-by-state election results from each presidential election and using the insights we gain to create machine learning classifiers.

**Timeline**

**Week 1:**

Scraping and cleaning data from state of the union addresses, citizen sentiment, and other seemingly important data. Putting this data into data frames and creating an organization scheme to start doing EDA on our data.

**Week 2:**

Continue developing EDA for our data, as well as scraping more speeches, rather than just the state of the union addresses. See if we can find patterns in speeches across presidents; consider Markov Chains for predicting “new” presidential speeches.

**Week 3:**

Explore data to find a way to use machine learning classifiers, e.g. predicting elections from state election data, using newspaper endorsements to predict elections, etc.

**References**

<http://www.presidency.ucsb.edu/showelection.php?year=2016>

<https://filiph.github.io/markov/>

<http://www.presidency.ucsb.edu/index.php>