Presidential Sentiment Analysis
MSDS692 – Data Science Practicum 1

Progress Report for Week *3*

Jeremy Beard

Project Details

This project is centered around a dataset which contains speeches from the State of the Union of all presidents. The project will utilize a variety of natural language processing techniques in order to answer questions that have been created surrounding the dataset. Sentiment analysis will be utilizing, general word commonality will be explored and word frequency will be analyzed. The final output will be a visualization comparing all the presidents to each other within the lens of the State of the Union.

Project Timeline:

Week 1 – Project definition and submit proposal (DONE)

Week 2 – Datasets selected and former related work collected (DONE)

Week 3 – Initial data loading and data cleaning performed (DONE)

Week 4 – Initial output and analysis for single speeches completed

Week 5 – Output and analysis for all speeches / all presidents completed

Week 6 – Result congregated, summary visualizations created, presentation began

Week 7 – Presentation completed, dry runs completed

Week 8 – Final Project Presentation

Planned Work for the Week:

This past week, I began loading in the data with a Python script and organizing it according to examples I've used in the past, collected previously. I wanted to begin to finalize the list of question I'd like to answer during this project. So far, the questions I have are:

- *What is the quantitative positive/negative sentiment between all the speeches?*
- *Which presidents use the widest variety of words?*
- *What is the quantitative positive/negative sentiment between all the presidents?*
- *What are the themes or buzzwords among the different speeches / presidents?*
- *What were the most common words used in each speech?*
- *Which presidents gave the longest speeches?*
- *Which presidents gave the shortest speeches?*
- *Which presidents used the most unique words?*
- *Which presidents used the least unique words?*

This next week, I plan to begin answering the questions put forth by myself previously. The data has been loaded into a python script and a list has been created which consolidates all the presidents' State of the Union speeches. This will assist with the analysis per president. The next steps are eliminating stop words, and beginning the frequency analysis and sentiment analysis.

Progress for the Week:

This week the python script was created which loads in the State of the Union speech data and begins to organize it. The list of questions I would like answered was also built out further and is mentioned above. This is likely the chosen set of questions until the project enters its final phase. In the python script, the data from all the presidents' State of the Union speeches was loaded into a list of dictionaries with the 'year', 'president', and 'text' fields. This was also sorted and a new list was created with the 'presidents' field consolidated. Since presidents have 4+ State of the Union speeches, this new list concatenates all of the respective presidents' speeches so each president only has 1 entry, complete with all of his speeches.

Roadblocks/Issues:

No roadblocks or major issues to report yet. Next week we will start to create some initial output and visualization. This may create some blockers or roadblocks but this is to be determined. No current issues are happening.

Plan for next Week:

This next week, I plan to begin answering the questions put forth by myself previously. The data has been loaded into a python script and a list has been created which consolidates all the presidents' State of the Union speeches. This will assist with the analysis per president. The next steps are eliminating stop words, and beginning the frequency analysis and sentiment analysis. Next week these initial outputs will be created.

Resources for the Week:

State of the Union Corpus (1790 - 2018). (2018, October 19). Kaggle.

https://www.kaggle.com/datasets/rtatman/state-of-the-union-corpus-1989-2017