Presidential Sentiment Analysis
MSDS692 – Data Science Practicum 1

Progress Report for Week *4*

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 utilized, 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 (DONE)

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:
From last week, my planned work for this week was to start outputting all the charts and data relative to the questions I have come up with. This list of questions is a living list but right now the questions are as follows:

- *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 week I had planned to finish the cleaning of the text and begin answering the questions I had set out for myself. This involves eliminating stopwords and possibly lemmatization. I am also thinking about whether to get involved with n-grams or not as I have done this in the past during my studies at Regis. Regardless, I will start with eliminating stopwords and see where that gets the data. Following that, I was planning to start just simple analysis and then evolve to the more involved questions on the list.

Progress for the Week:
If you will recall, last week I had started the Python script which loads in the data and begins organizing it. This week, I built out that script to 320 lines (with lots of comments), multiple functions, and a handful of visualizations already to begin answering the questions I had wanted to answer. Following last week's work, this week I added code to remove stopwords, create a couple simple overview plots, calculate word counts and words per speech, and more. I created a 'word_substance' feature which takes a ratio of the cleaned word_count / total word_count of speeches to determine the 'wordiness' of the speech. I also added code to save four (4) plots, capturing this word count per speech analysis and word substance analysis.

Roadblocks/Issues:
No roadblocks or major issues to report yet. I have a feeling that it will be difficult to focus my results. Maybe not, but I'm starting to visualize what a final product may look like and I'm having trouble determining the medium. Will this just be a report and some code? Should I built it into a web interface of sorts? How far should I take this?

Another thought I had was if what I'm doing is 'enough'. I'm doing a lot of analysis and sentiment analysis but not much prediction. Is this OK? I'm starting to feel the pros and cons of the freedom of this course 😊

Plan for next Week:
Next week I plan to finalize my analysis and begin to flesh out the draft of what the final product of this analysis may look like. How can I make my visualizations look more compelling? How can I make the biggest impact from this analysis of this particular dataset? Any ideas are appreciated!

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