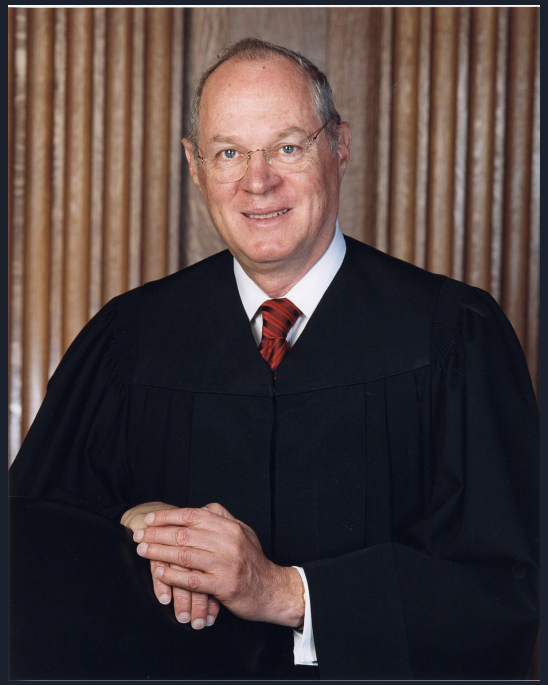
Capstone 2:

Predicting how Supreme Court Justice Anthony Kennedy Votes

Overview

- Introduction
- Data Extraction and Wrangling
- Preliminary Findings
- Machine Learning Strategy
- Machine Learning Results
- Conclusion and Future Work



Introduction

- Justice Anthony Kennedy has been of interest to SCOTUS (Supreme Court of US) watchers since 2005
  - Sandra Day O'Connor was the last 'swing vote' on the court
- Justice Kennedy has been a key vote on important social and political issues such as same-sex marriage and gerrymandering
  - 'Key vote' in this context means he breaks the 4-to-4 tie of the other 8 justices

Data Extraction

- Learned multiple new python packages and libraries to parse the SCOTUS website and download necessary files
- Parsed PDF files and transformed them into text documents for data wrangling
- Used the Washington University in St. Louis (WUSTL) Supreme Court Database to tag each case and extract further information about each case

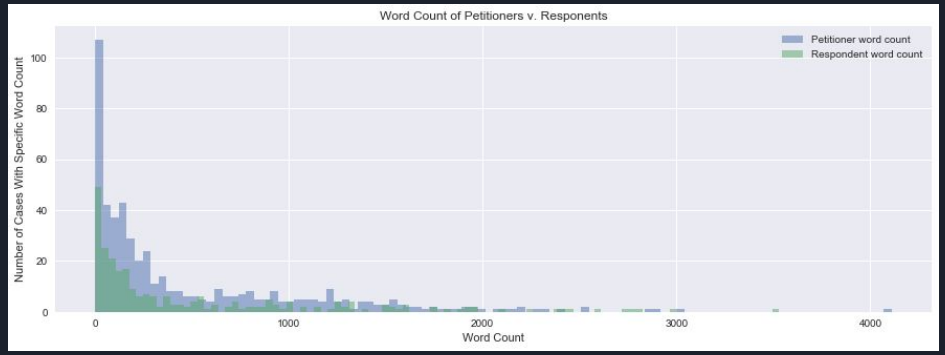
Data Wrangling

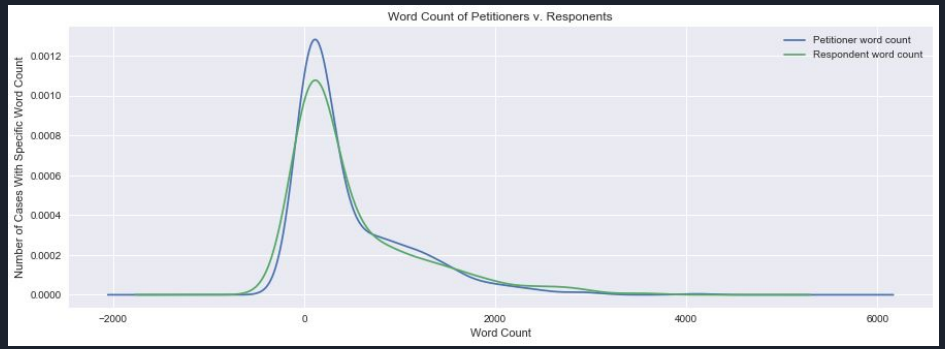
- Text had to be extensively cleaned due to markings and transcription peculiarities that produced irregularities in the output text files
- A specific script was written to clean the text, pre-process it using the Snowball Stemmer, and joined with the WUSTL database

Data Wrangling

- Cases from 2000 - 2003 could not be used because the court reporter did not mark which Justice was speaking when she took down her transcripts
- Cases from 2013 and 2014 were discarded because their PDF files could not be parsed (they returned text without spacing)

Notable EDA Results





- Word count, an engineered feature, did not show any correlation between the Petitioner winning or the Respondent winning
- Kennedy votes with the winning side a vast majority of the time; he rarely dissents
- On the whole, Kennedy has a slight conservative bent in his voting record
- State and Federal governments are usually parties to a SCOTUS case

Machine Learning Strategy

- Text was broken up according to the Issue Areas
  - This was to test the hypothesis that there is more predictive power when examining specific subject areas rather than all speech text indiscriminately
- The following pre-processing steps were applied to the text before being fed into ML models:
  - Stopwords were removed
  - Text was stemmed with the Snowball Stemmer
  - Tf-Idf was applied as the final vectorizor before being inserted into ML models
- Text was fed into several ML classifiers:
  - Multinomial Naive Bayes, SVM, Decision Tree, Random Forest, AdaBoost,
  - Each classifier was joined with a GridSearch pipeline that found the best possible hyperparameters for a given model

Machine Learning Results

	Baseline	Multi. Naive Bayes	SVM	AdaBoost	Decision Tree	Random Forest
Criminal Procedure	65.1%	68.8%	64.2%	51.4%	52.2%	66.9%
Civil Rights	73.6%	81.5%	76.9%	76.9%	70.8%	78.5%
Economic Activity	70.1%	77.2%	77.2%	59.5%	58.2%	72.2%
Federalism	56.4%	45.0%	45.0%	50.0%	50.0%	45.0%
First Amendment	69.0%	66.7%	52.4%	66.7%	66.7%	66.7%
Judicial Power	58.5%	52.8%	52.8%	52.8%	52.8%	54.7%

Machine Learning Results

- The initial hypothesis that there are differences in predictive power between issue areas and models was confirmed
- Highest Predictive Power:
  - Civil Rights & Multinomial Naive Bayes: + 7.9%
- Lowest Predictive Power:
  - First Amendment & SVM: -16.6%

Conclusion and Future Work

Conclusion:

- There is some predictive power in correctly selecting the winning side for certain types of cases and models
- Justice Kennedy's word count during oral argument is not strongly correlated with either winning or losing; however, this may be different for other Justices

Future Work:

- Apply this analytical framework to all of the Justices and create an aggregate model of Court predictions
- Test the MNB model which had the highest predictive power in the
- Engineer new features, such as speech interruptions, and join them to the current features to see if there is improvement in model performance