Ismael Rodriguez
CIS 731
Final Project Proposal
30 SEP 2020

Subject and Purpose. The Supreme Court has suddenly become a topic of great interest. Over time, the justices of the court have pronounced judgements, agreeing, concurring, or dissenting on court cases. The text of these opinions are not only well-suited for sentiment analysis, but also useful as the basis for exploratory trend analysis of those sentiments. While it should be possible to group and sort data by a variety of factors, this project will focus on sorting cases by opinion authors, in order to visualize associated sentiment patterns.

Team Member. I am the only member of this research project. While I am open to a group project, I feel the scope of this project is within the reach of a single investigator.

Problem Description. While at a fundamental level, the trends gleaned from the Supreme Court Dataset should be interesting on their own right, I hold ulterior motives. I intend to use this project to explore the use of machine-learning-based sentiment analysis techniques on large, complex text files. In turn, I want to track the sentiments of supreme court justices over time. From an analytic perspective, I would like to identify when and where unusual outcomes occur. I recognize that these strange outcomes may only emerge through the process of exploratory analysis. However, it will take significant effort to prepare the data set for this analysis.

Data Cleaning. I intend to break down the project using project design techniques that focus heavily on data cleaning that ensures each record has all identified opinions, and associated attributes.

Preliminary, Exploratory and Explanatory Analysis. Preliminary and exploratory analysis will group and sort the dataset by justices over time, excavating for temporal patterns in the dataset. Explanatory analysis will build upon calculations of sentiment for the text of each opinion so that those sentiment measurements can then plot over time.

Data Set. The Supreme Court has produced a huge library of cases. A cursory review of Jacob Boysen's SCOTUS dataset, available on Kaggle.com, suggests that it an extraction of data from 1946-2017. These records came from an API sponsored by courtlistener.com (2017). Alternatively, Boysen also provides a link to an even larger source of data. Hosted by Washington University Law, these cases stretch from the early days of the nation until the court term that ended in 2019. Further split into two datasets, the first of which consists of all cases up to 1945, while the other consists of all cases from 1946 to 2019. In total, the dataset includes a wide variety of information, with each record holding sixty attributes that, multiplied over 228 years of records, results and in over 157,000 specific cells of data (Spaeth et al, 2020). Of all this information, most important to this project will be data on the votes and opinions of individual justices, as well as descriptive data associated with dates, origins, and issues areas. I will attempt to merge the records from 1791 to 2020, gleaning only those attributes of interest for the project. If that task proves too difficult, I will resort to using only post-1945 results found in the Kaggle dataset.

Methods.

Data Preparation. I intend to use PySpark to process the data set in a distributed manner. I will examine the structure of the dataset to determine how best to extract and transform key information. At a minimum, the condensed record features will require opinion author, publication date, and opinion text.

Sentiment Analysis. Once a record of suitable opinion text exists, the next step will be to calculate the sentiment of the opinion text. More specifically, I would prefer to use a pre-existing model that measures ideological sentiment.

Temporal Trend Analysis. With sentiment scores in hand, the next step will be to visualize trends over time. I would like to employ python-based visualization techniques. However, if this presents a time challenge, I will resort to the use of Tableau for final visualizations, as I have gained solid experience both academically and professionally in the use of this tool.

Project Template. One inspiration for this type of trend analysis is a project by Greg Rafferty to measure and visualize the emotions and sentiments of each chapter of each book in the *Harry Potter* series. More specifically, the project uses a variety of sentiment analysis techniques to identify better performing options. As such, it is a good template for both the sentiment analysis and trend analysis coding patterns (2018, 2020).

Evaluation. While much of the focus of this project is on the extraction of patterns over time, I will still seek to evaluate the use of sentiment analysis. As this project does not intend to predict outcomes through machine learning, detailed evaluation may prove difficult. As such, I will investigate using performance measures to gauge the computation of visualizations, as suggested in a conversation with Professor Hsu on September 30, 2020.

References

Boysen, J. (2017). *SCOTUS Opinions Corpus,* Kaggle, https://www.kaggle.com/jboysen/scotus-corpus, accessed 30 September 2020.

Boysen, J. (2017). *scotus_file.gz* [gz], Kaggle, https://www.kaggle.com/jboysen/scotus-corpus, accessed 30 September 2020.

Rafferty, G. (2020). *Raffg / harry_potter_nlp / sentiment_analysis.ipynb* [ipynb], Github, https://github.com/raffg/harry_potter_nlp/blob/master/sentiment_analysis.ipynb, accessed 30 September 2020.

Rafferty, G. (2018). *Sentiment Analysis on the Texts of Harry Potter*, Towards Data Science, https://towardsdatascience.com/basic-nlp-on-the-texts-of-harry-potter-sentiment-analysis-1b474b13651d, accessed 30 September 2020.

Spaeth, H., Epstein, L. et al. (2020). *2020 Supreme Court Database*, Version 2020 Release 1, http://Supremecourtdatabase.org, accessed 30 September 2020.