Supreme Court Oral Arguments Outcome Prediction Team First Checkpoint

Chanteria Milner, Federico Dominguez, Jessup Jong, and Michael Plunkett

Presumed Rubric:

- Load data and preprocess the text.
- Conduct basic exploratory analysis (descriptive statistics such as counts, sums, averages.
- If you filter the dataset and only use a subset of transcripts (e.g., only cases heard by the Roberts Court), provide justification. You only need to report the above details for your filtered subset.
    - *We did not filter the dataset, so this does not currently apply to us.*

Repository URL: Link

- Instructions on how to run it and the general repository functionality are within the `README.md` file and code documentation strings, respectively.

Statistics File: Link

Project Summary:

The project uses historic United States Supreme Court cases to train natural language processing models to predict case rulings.

Project Structure:

The project is structured in a way so that all functionality is discreetly packaged and runnable via `make ____` commands (`make get-data`, `make clean-data`, `make describe-data`). It is also possible to run the whole application from downloading the external files to doing the initial statistical analysis in one command from the command line via the `make run` command.

All data is stored in this file space, presuming you are looking from the repository base directory: `./supreme_court_predictions/data/`. There are four folders within that folder that store specific data that is specified by their name: `clean_convokit`, `convokit`, `models`, and `statistics`. For each folder, there is a `README.md` file that explains what files are stored within them. The storing of this data is done locally and ignored via the `.gitignore` file so that large amounts of data that are easily accessible through the code mentioned above are not saved in the repository. To put it another way, any data needed to run the `make ____` functions, requires the user to run the `make get-data` command on the command line.

Initial Statistical Analysis:

We partitioned our preliminary analysis into five sections, with our metrics being counts and averages.

Advocates:

- For this analysis, we provided unique counts of advocates that were on the following sides: for the petitioner, for the respondent, unknown, or for amicus curiae. We also provided counts of the total unique advocates across all the cases, the unique roles that attributes can have, and a count of aggregate roles that advocates can have. We determined these aggregate roles by doing visual text analyses of a subset of the roles listed in the dataset. This led to the following aggregate roles that advocates could take: inferred, for the respondent, for the petitioner, and for amicus curiae.
- The majority of advocates worked on the side of the petitioner (approximately 39%), and the majority of the advocate roles were inferred by the dataset creator (about 37%).

Cases:

- For this analysis, we provided descriptives statistics of the count of cases on each win side (for the petitioner, for the respondent, unclear, or unavailable); a count of total cases; the count of courts cases span; the total number of years the cases dataset span; the total number of unique petitioners; and the total number of unique respondents.
- About 65% of cases ruled in favor of the petitioner.

Speakers:

- This analysis describes the speakers present across the cases. Speakers took on 1 of 3 roles: advocates, justices, or unknown. The majority of speakers were advocates, with only 35 speakers being justices. The dataset also included approximately 8900 unique speakers.

Voters:

- For this analysis we counted the sum total of positive, negative, and absent votes from justices along with the total number of justices that voted over the span of 65 years (1955 to 2019).
- Within that timespan 35 justices were accounted for in the records, roughly 59% of votes were positive, 40% of votes were negative, and 1% of votes were not cast as either positive or negative.

Utterances:

- For this analysis we counted the average number of utterances per case and the average number of speakers that showed up per case.
- The average number of utterances per case was approximately 253 and the average number of speakers per case was 10.