SFWRENG 4NL3 Assignment 2

Sumanya Gulati

7 February 2025

Contents

List of Tables

List of Figures

1 Dataset

The dataset I chose consists of transcripts of US Supreme Court Nomination Hearings held by the Congress from 1971 till 2024. For the scope of this assignment, three women, namely, Justice Sandra Day O'Connor, Justice Ruth Bader Ginsburg and Justics Sonia Sotomayor have been included along with four men - Justice Samuel Alito, Justice Anthony Kennedy, Justice John Roberts and Justice Antonin Scalia.

In 1981, then US President, Ronald Reagen appointed the first woman to the US Supreme Court. This was a century after women in the US fought for their right to practice law as attorneys. To put things into perspective, Ruth Bader Ginsburg, despite graduating from Harvard Law School could not find a job as a practicing attorney at any law firm because of gender-based discrimination. Through her unwavering determination, she co-founded the Women's Rights Project at the American Civil Liberties Union (ACLU).

With such blatant discrimination on the basis of sex in the domain of law, I was curious to find whether the society's attitude towards women has changed over the past few decades. Using the transcripts from the supreme court nomination hearings, I wish to analyze the attitude and mindset of the senators questioning these nominees. The difference in the lenses through which men and women with equal qualifications are judged for the same job is pretty obvious. While women are questioned about their committment towards their family, how they plan on maintaining a work-life balance, their personality and whether they're amiable; the men are judged based on their professional qualifications with more focus put on their competency.

Through this assignment, I hope to see results that are consistent with this widespread belief and show the inherent bias so many people still carry when interviewing equally competent individuals for the same job.

1.1 Data Collection and Splitting

This dataset has been curated using publicly available data on the Supreme US Government Information website.. For each individual in both the categories, the following transcripts from the hearing have been included -

For each nomination hearing, the following files have been included as individual documents:

- Statements of Committee Members (including Prepared Statements)
- Statements and/or Testimony by the Nominee
- Questions and Answers

These PDFs (each corresponding to one individual document) were manually downloaded and saved in separate folders. This data can be found in the `female-nominees` and `male-nominees` folders, respectively.

1.2 Description

The `text_extract.py` script was then used to extract all the data from the PDFs and save them as .txt files which can be found in the `female-nominees-processed` and `male-nominees-processed` folders. For the purpose of this assignment, each text file in the folders is considered a document. Additionally, two categories - *female nominees* and *male nominees* have been created. This means that, for the former category, there are 136 documents and for the latter, 147 documents in total.

Table 1.2 lists the average number of tokens and the average number of unique tokens per category.

Category	**Average Number of Tokens**	**Average Number of Unique Tokens**
Female Nominees	4282.378	835.897
Male Nominees	2024.15	593.68

Table 1: Average Number of Tokens Per Category

This is also visualized in figure 1.2 which evidently displays that the number of tokens per document for the female nominees is twice as much as that of male nominees. The difference between the average number of unique tokens for the two categories is not as sigficant, however.

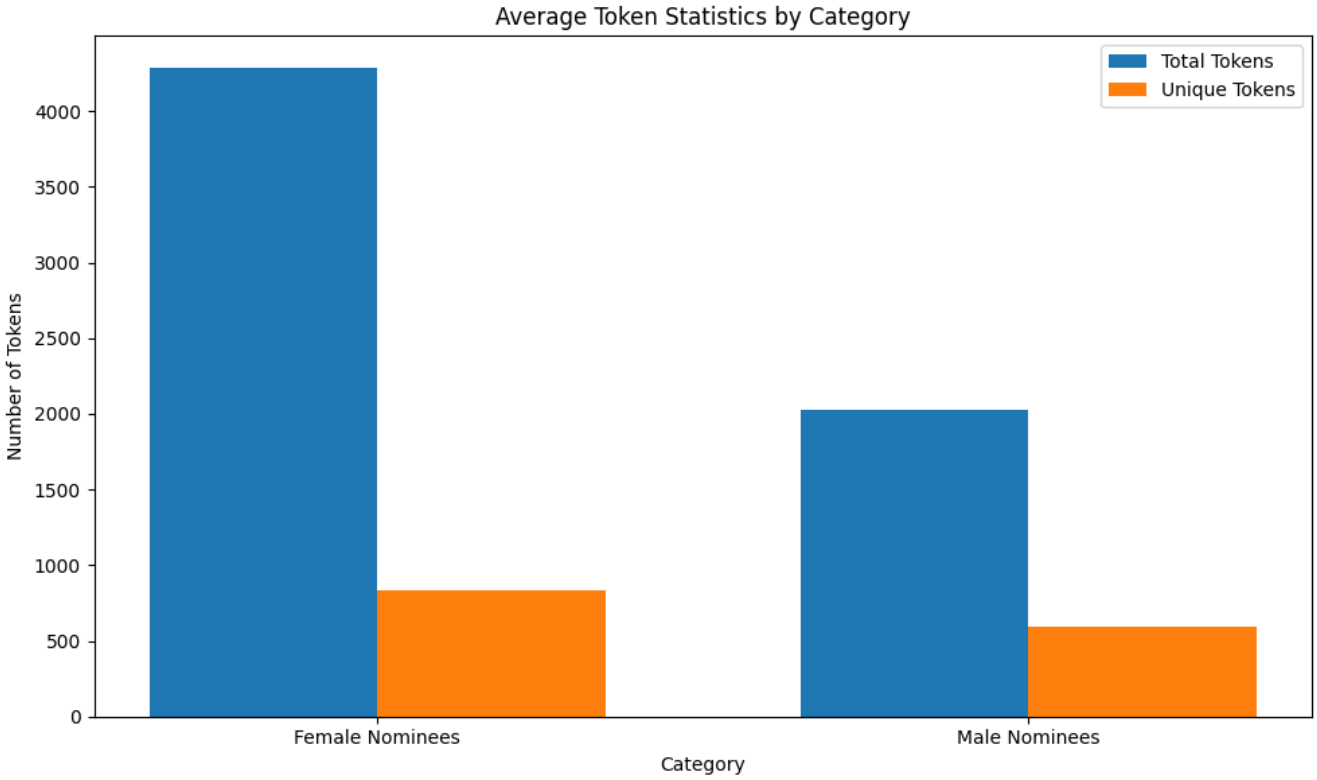

Figure 1: Tokens Per Category

2 Methodology

2.1 Preprocessing

Since the `data_extract.py` file only extracts the data from the PDF and store it in a text file, the first step in preprocessing the dataset was to clean the text. This was accomplished using the following methods through iterative testing:

- Replacing all new lines with a blank space.
- Removing all words that have a hyphen followed by a blank space. For example, long- term to longterm.
- Removing all leading and trailing spaces along with the replacement of multiple blank spaces with a single space.
- Removing all digits to streamline the frequency count and ranking of alphabetic tokens. This was implemented because with the digits in the corpus, the years and other random numbers referencing legal cases and such populated the top 10/top 25 counts corrupting the data analysis.
- Removing all punctuation for more accurate tokenization.

2.1.1 Naive-Bayes

Further preprocessing for Naive-Bayes included the following steps:

- Using the list of common stopwords from the *nltk* library to remove words that do not reflect the actual content.
- The list of stopwords was expanded to account for commonly found words in the dataset that did not convey the sentiment or the content of the sentence. This includes words such as `said, would, one, jan, feb`.
- To further refine our results, commonly used legal jargon was also added to the list of stopwords. This includes words such as `hearings, committee, senator, judge, case`.

2.1.2 Latent Dirichlet Allocation (LDA) Preprocessing

Further preprocessing for LDA included the following steps:

- Adding additional stop words including 'sgpohearings' and 'sgpohearingstxt' that are not actual words with defined meanings.
- Using NLTK's lemmatizer removing any tokens starting with 'sgpo'.

2.2 Bag-of-Words Representation

As shown in tutorials, the `CountVectorizer()` function was used to implement Bag-of-Words on the dataset. Probability calculations given in the instructions PDF for this assignment were then used to implement the `calculate_word_probabilities()` function. The output from this function was then to calculate the Log Likelihood Ratio.

2.3 Naive-Bayes Model

Given the fact that the dataset is made up of real transcripts of legal hearings, the used of names of the judges and/or the senators interviewing them are over-represented in the dataset. This skews the results of analyses performed on the processed dataset.

To combat this issue and to maintain the focus on the goal of this assignment, classifications were establish so the focus of the analyses would remain on words that align with the characteristics of the nominee. 4 different classifications - Legal Interpretation, Background Experience, Personal Characteristics and Competency Qualification were created.

Each of these classifications was populated with a list of related words that speak to that particular classification, such as - words like 'ruling', 'precedent', 'principle', 'rights' and more were added to the Legal Interpretation classification. By analyzing the frequency of these words in the vocabulary for the female and male nominees, any potential bias towards a nominee can be spotted. If the focus of the hearing is on 1-2 classifications as opposed to a general overview of all of them, this would highlight implicit bias.

2.4 Topic Modelling

The `gensim` and `pyLDSvis` libraries were used to implement LDA topic modelling as shown in the tutorial. The `get_topic_words()` and `get_document_topic()` functions were then used to get information on the topics and documents.

2.5 Experimentation

In addition to importing all the functions from the `processing.py` file, two new functions were added to implement text normalization variation and LDA variation. `PorterStemmer()` was used to implement stemming as a part of the preprocessing of the data. Another function was then implemented to create a TF-IDF matrix using the `TfidfVectorizer()` function. The code for this was based on the tutorials.

3 Results and Analysis

The results and insights gather from the analyses has been summarized in the following subsections.

3.1 Naive-Bayes

Gender	Category	Word	LLR Score
female	background experience	backgrounds	1.931906639972084
female	background experience	experienced	1.0156159080979297
female	background experience	experiences	1.9682742841429581
female	background experience	framework	2.52789007207838
female	background experience	services	1.4982706548972224
female	personal characteristics	biases	1.6645918703447364
female	personal characteristics	ethnicity	2.6250538205320293
female	personal characteristics	gender	1.0750667224516768
female	personal characteristics	genderbased	2.10180567676748
female	personal characteristics	grace	1.2751271035830118
female	personal characteristics	identity	1.1209764237557547
female	competency qualification	applicability	1.8941663119892365
female	competency qualification	availability	1.46928311802397
male	legal interpretation	judicially	1.1407866747180364
male	legal interpretation	overruling	1.2014112965344719
male	background experience	practiced	1.4245548478486825
male	background experience	worker	1.0762481535804653
male	background experience	workload	1.4817132616886308
male	personal characteristics	beliefs	1.3067718121922978
male	personal characteristics	embraced	1.7693953341404107
male	personal characteristics	unbiased	2.0570774065921924
male	competency qualification	disability	1.209779546204988
male	competency qualification	wellqualified	2.0570774065921924

Table 2: Results from the Naive-Bayes Analysis

As shown in table 3.1, the analysis shows that when interviewing female nominees are interviewed, greater focus is put on their personal characteristics with conversations about biases, gender, grace and identity taking up a significant amount of attention. In terms of their background experiences and competency qualifictaions, the focus is put on their availability and the top words from the category across documents are neutral in tone. None of the words added to the legal interpretation category were found in a significant frequency for the female nominees.

As for the male nominees, however, words like 'judicially' and 'overruling' are highlighted showing the focus on the conversation around the nominee's career in law and their judgments. The conversations around personal

characteristics seem to be more focused on the individual as opposed to their gender or other aspects of their identity. The most starking difference can be seen in the tone of the competency qualifications since the word 'wellqualified' seems to have a very high probability showing the appreciative tone of the hearings.

3.2 Topic Modelling

The complete list of topic words along with their labels can be found in the `topic_words.csv` file. The averages for the top 20 of these topics split by the gender can also be found in the `category_topics.csv` file.

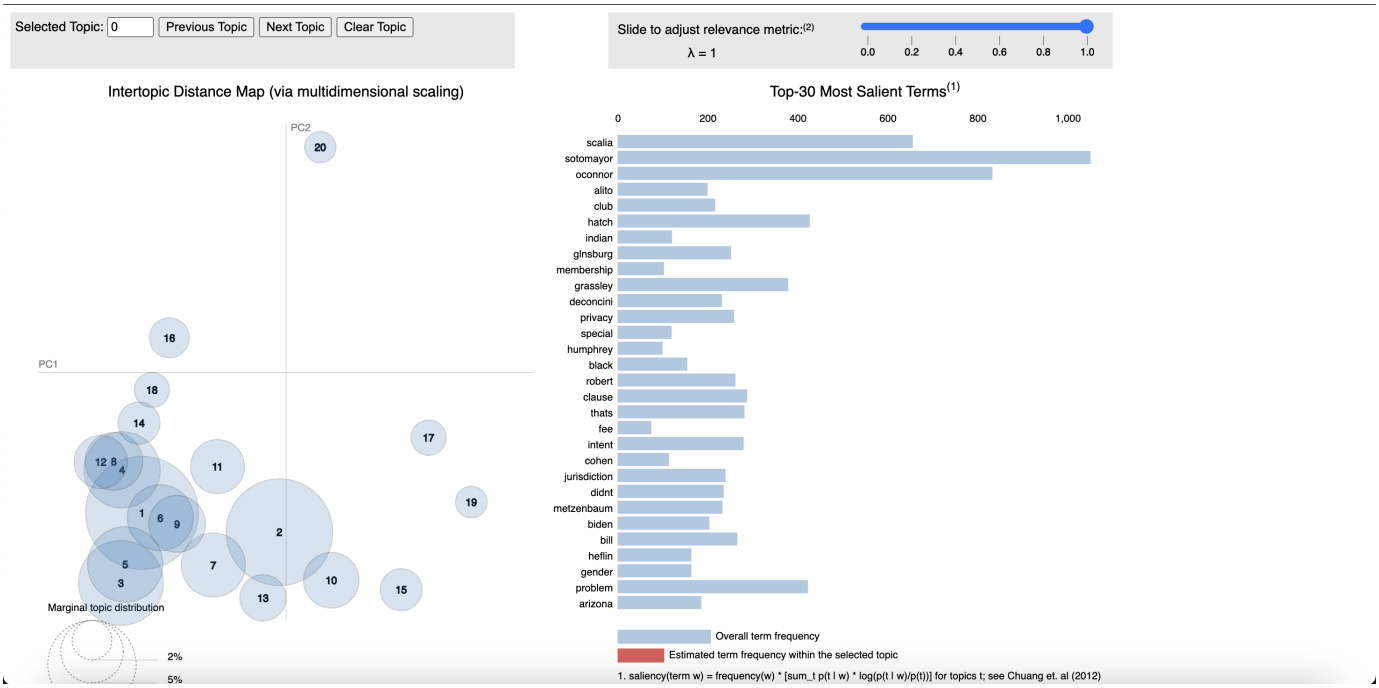

Figure 2: Results from LDA Topic Modelling

As shown in figure 3.2, the intertopic distance map visualizes the relationship between the different topics where each bubble represents a topic with the bubble's size indicating the topic's overall prevalence in the dataset. Topics 12, 8, 4, 1, 6, 9, 5, 3 and more are overlapping and clustered together suggesting similarity.

The 30 most important words are either names, identity-related terms (gender, black) or general legal jargon.

3.3 Experimentation

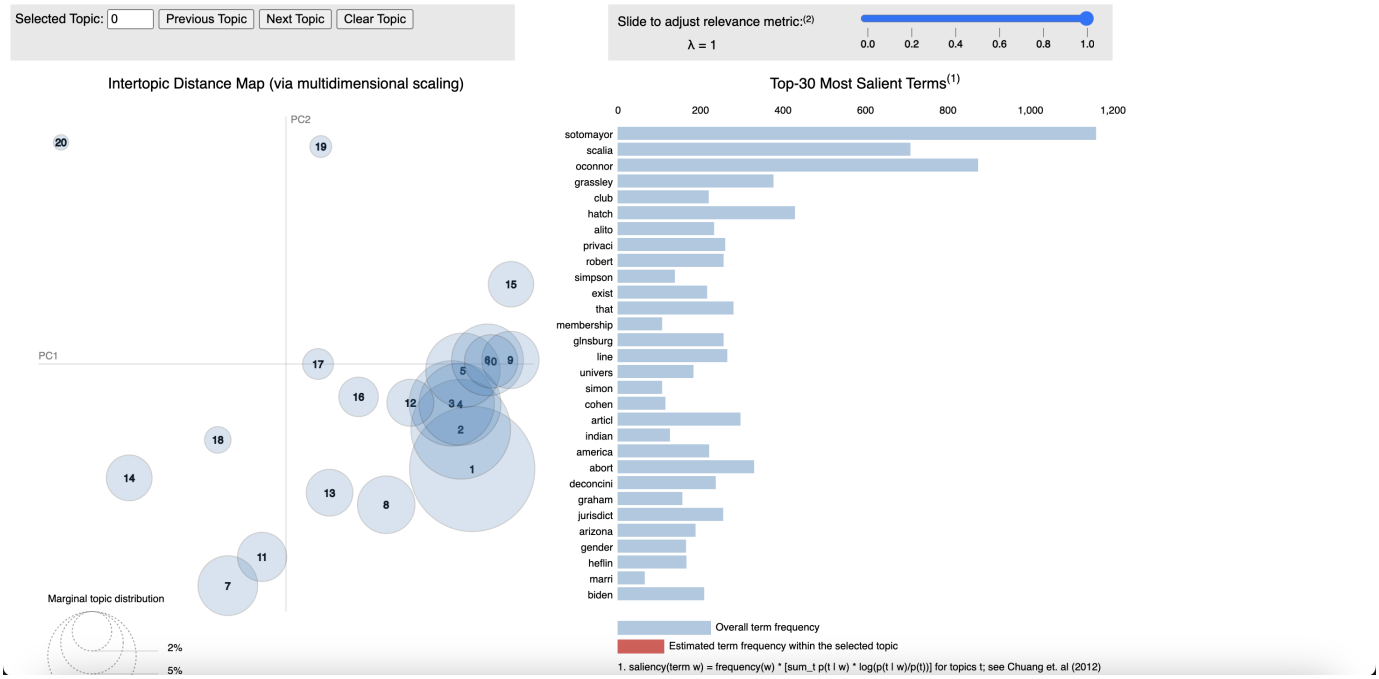

Figure 3: Results from LDA Topic Modelling Post-Stemming and with TF-DIF

As shown in figure 3.3, the intertopic distance map visualizes the relationship between the different topics where each bubble represents a topic with the bubble's size indicating the topic's overall prevalence in the dataset. Topics 5, 6, 10, 9, 3, 4, 2, 1 and more are overlapping and clustered together suggesting similarity.

The list of the 30 most popular words seems to have a lot of incomplete words suggesting improper text preprocessing and analysis.

4 Discussion

4.1 Findings

4.2 Reflection

5 Appendix