COSC 757 Final Project Write-up

Daphne McWilliams
Department of Computer Science
Towson University
Towson, MD
dmcwil2@students.towson.edu

***Abstract*—This is the final write-up for the COSC 757 term project. In this project, I attempted to implement topic-modeling techniques known as Latent Dirichlet Allocation on a data set containing all U.S. Supreme Court opinions issued between 1970 and 2019.**

I. Introduction

A. Data Set

This data set contains almost every Supreme Court opinion written in the United States since the founding of the Court in 1789 up until 2017. As such, it includes some 35,000 opinions written by 96 different justices. Many databases of information relating to Supreme Court justices and decisions already exist—a notable example is Washington University in St. Louis' Supreme Court Database; however, there do not appear to be any databases that feature the full text of each opinion. Although all Supreme Court opinions are publicly available, they do not exist in a form that can be easily downloaded or that naturally lends itself to analysis using data mining techniques, creating a need for a data set of this type. The data set I used in this project was produced using Courtlistener.com, which has an API and allows bulk downloads of opinion JSONs. It includes the full text of each opinion and associated metadata such as author, date, case name, type (majority opinion, dissenting opinion, concurring opinion), etc.

For the purposes of this project, I worked with a smaller subset of opinions written between 1970 and 2019 (about 7,500 opinions written by 25 different justices). This range was chosen for purposes of practicality and feasibility, and also because it seemed like a potentially interesting range to work with, since many of the issues that came before the Court in the 1970s (e.g., abortion, criminal justice, rights of the press, separation of church and state, etc.) have been continuously revisited over the intervening decades; some are even be in the process of relitigated as we speak. An example torn straight from recent headlines is the leak of a draft opinion written by Justice Alito, which would overturn the 1973 *Roe v. Wade* ruling, a landmark case that protects a woman's right choose to have abortion without excessive government restriction based on her constitutional right to privacy guaranteed by the Fourteenth Amendment.

B. Background

The Supreme Court is the highest court in the United States; its best-known power is that of judicial review, or the ability to declare a legislative or executive act in violation of or in compliance with the constitution. After hearing oral arguments and deciding a case, the Court issues an opinion, which announces a decision articulates the legal rationale that the Justices relied upon in making the decision.

Depending on how the Justices decide certain cases, the opinion may take one of several forms. When the decision is unanimous and all the Justices offer one rationale for their decision, the Court issues one *unanimous opinion*. When more than half (at least five out of nine) Justices agree, the Court issues a *majority opinion*. When there is no majority, the Court may issue a *plurality opinion.* Often, the Justices are not in complete agreement with the main opinion. Justices who agree with the result of the main opinion but base their decision on a different legal rationale may issue a *concurring opinion*. Likewise, justices who disagree with the main opinion in both result and legal rationale may issue a *dissenting opinion.*

The majority vote of the nine Justices is the sole determinant of the outcome of a Supreme Court case; however, the written majority opinion determines the scope of and the justification for the precedent that the immediate ruling establishes. So, while the case-deciding power of the Court rests with the rulings, the broader precedent-setting power of the Court lies in the written opinions. Precedent refers to a court decision that is considered authoritative in subsequent cases involving identical or similar facts, and so is very important. Precedents can shape the direction of litigation, legislation, and future rulings by courts at all levels. An example is the legal decision in *Brown v. the Board of Education*, which guided future laws about desegregation.

C. Project Overview and Goals

The overall goal is to use a topic modeling technique known as Latent Dirichlet Allocation on a data set of Supreme Court Opinions from 1970–2019 to group opinions by their subject matter. This could enable opinions about the same issue can readily be identified and accessed, potentially setting the groundwork for further comparative study of these documents and the Justices who wrote them.

II. Data Mining Model

Topic modeling is one of many techniques associated with natural language processing (NLP), a field of artificial intelligence that is concerned with the interactions between computers and human language, in particular the ability of computers to process and analyze the contents of documents, including the contextual nuances of the language within them. Topic modeling is an unsupervised learning technique that uses Bayesian statistical models to find and observe clusters of related or co-occurring words (i.e., topics) in a group of texts. All topic modeling techniques are based on the assumption that each document consists of a mixture of topics and each topic consists of a collection of words. The semantics of documents are governed by hidden or "latent" topics, and it is the task of topic modeling to uncover the

topics that shape the meaning of the document and the larger corpus to which it belongs.[1]

The topic modeling approach is useful for document clustering, organizing large blocks of textual data, information retrieval from unstructured text, and feature selection. As I have discovered over the course of this project, topic modeling is also capable of discovering hidden semantic structures in a text that might not immediately occur to human observer.

Finally, there are many approaches for obtaining topics from a body of text. A very popular approach is Latent Dirichlet Allocation (LDA), which I used on this data set to generate topics.

A. Latent Dirichlet Allocation

LDA was first developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002. It has since become a very popular topic modeling technique due to its ease of use and high rate of success in dealing with textual data from various fields including social media, medical science, and political science, among others. LDA is described as a probabilistic generative model that associates each document with a probability distribution over topics, where topics are probability distributions over words. It is worth noting here that the number of topics must be specified by the human user ahead of time.

LDA uses Dirichlet distributions for document-topic and word-topic distributions. In brief, a Dirichlet distribution can be thought of as a distribution over distributions. In essence, it answers the question: "Given this type of distribution, what probability distributions am I likely to see?" A probability distribution simply links each possible outcome to its probability of occurrence. In topic modeling, a document's probability distribution over topics (i.e., the mixture of topics most likely to be discussed in that document) might look like the example below.

Topic 1 = .73

Topic 2 = .10

Topic 3 = .17

A topic's distribution over words (i.e., the words most likely to be used in a topic might look like the example below.

Topic 1:

Cat = .39

Meow = .32

Kitten = .29

The basic idea behind LDA is that documents are represented as random mixtures of latent topics, where each topic is characterized as a discrete probability distribution that defines how likely each word is to appear in a given topic.[2] In other words, documents include some number of topics, and topics use a specific set of words. It assumes that a document looks something like Fig. 1 below.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Fig. 1. LDA topic distribution over a document (from Blei et al., 2002, Fig. 8 on pg. 1009)

LDA will tend to discover set of terms or topics based upon the co-occurrence of individual words. It is based upon the assumption that the semantics (i.e., meaning) of words can be grasped by looking at the contexts the words appear in. It treats each document as a bag of words with no consideration for structure (i.e., syntax, word order) beyond topic and word statistics. The topics "discovered" through LDA can then be used to classify any individual document within a collection in terms of how relevant it is to each of the discovered topics, or be he subject of further analysis themselves.

III. Data Preprocessing and EDA

Before cleaning, this data set had 10,991 entries. The variables in this data set include author_name (name of justice who wrote opinion unless it is a per curiam opinion), category (dissenting, majority, concurring, per curiam, second dissenting), per_curiam (true or false), case_name (Party 1 v. Party 2 name of case), date_filed (date on which the decision was issued), federal_cite_one (federal case citation), absolute_url (url of case's full page on courtlistener.com), cluster (url of associated cluster object on courtlistener.com), year_filed (year in which decision was filed), scdb_id (case ID in the Supreme Court database), scdb_decision_direction (decision direction), scdb_votes_majority (number of justices voting in majority), scdb_votes_minority (number of justices voting in the minority), and text (full text of opinion). For data cleaning and preliminary data analysis, I used Jupyter Notebooks and the following Python libraries: pandas, numpy, matplotlib, and seaborn.

A. Data Cleaning:

Shown in Fig. 2 is the number of opinions per justice before data cleaning.

[1] https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

[2] https://thesai.org/Downloads/Volume6No1/Paper_21-A_Survey_of_Topic_Modeling_in_Text_Mining.pdf

Justice Stevens	1181
per_curiam	755
Justice Rehnquist	752
Justice Brennan	732
Justice Scalia	718
Justice White	655
Justice Blackmun	654
Justice Marshall	611
Justice Thomas	551
Justice O'Connor	518
Justice Kennedy	461
Justice Powell	455
Justice Burger	417
Justice Breyer	416
Justice Ginsburg	407
Justice Stewart	328
Justice Souter	291
Justice Douglas	287
Justice Alito	239
Justice Sotomayor	175
Justice Kagan	110
Justice Roberts	94
Justice Black	59
Justice Harlan	50
Justice Gorsuch	47
Justice Kavanaugh	20
Justice 02122	2
Justice Holmes	1
Justice McReynolds	1
Justice Woods	1
Justice Connor	1
Justice Fuller	1
Justice Waite	1

Fig. 2. Number of opinions per justice (before cleaning)

One notable oddity in this data set is the 755 opinions attributed to "per_curiam." A per curiam opinion represents a unanimous agreement, in which case the opinion is written by the court and no individual author is cited. Per curiam opinions tend to be somewhat short and low on detail since they represent opinions that the Court considered fairly straightforward and uncontroversial. For this reason, I've chosen to drop per curiam opinions from this data set.

As is clear from the figure above, there were also a handful of justices that had only one or two opinions attributed to them (Justice 02122, Holmes, McReynolds, Woods, Connor, Fuller, and Waite), which warranted investigation. These entries all turned out to have made their way into the data set through a series of errors, which are detailed below.

Of the seven justices with no more than two opinions to their name, five predate 1970 by several decades. For example, Justice James McReynolds is the most recent of this group, and he retired from the Supreme Court in 1941. These Justices were somehow included in this data set because the "year-filed" variable for the cases associated with them was incorrect. For example, Justice McReynolds appears in this data set because he is associated with a majority opinion allegedly filed in 2005 on *Hamburg American Co. v. United State*s. In reality, this case originally came before the court in 1928. I looked through the records on the Supreme Court website, and couldn't find any reason for these earlier cases to be associated with an additional later filing date or any evidence that they were revisited at this later date, so I can only assume that their inclusion in this data set is due to some type of clerical or record-keeping error.

Additionally, no Supreme Court Justice with the last name "Connor" ever existed. The "Justice Connor" that appears towards the bottom of Fig. 1 is actually a misspelling of Justice O'Connor (Sandra Day O'Connor) and is associated with a dissenting opinion that Justice O'Connor gave in 1998 on the *Swidler & Berlin v. United States* Supreme Court case. Finally, two opinions were attributed to an incorrectly named "Justice 02122." For these reasons, the seven justices with fewer than 20 opinions were dropped from the data set.

Fig. 3 shows the breakdown of opinions by type before cleaning: majority, dissenting, second dissenting, concurring, and per curiam.

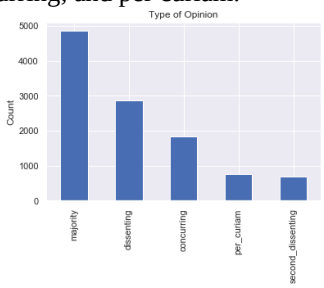

Fig. 3. Opinions by category

I ended up rolling 'second-dissenting' opinions into the same category as 'dissenting' opinions just for clarity and readability. The distinction also did not seem important for the purposes of this project, since a second dissenting opinion is merely an additional opinion that breaks from the majority opinion, but on different grounds than the first dissenting opinion.

B. Exploratory Data Analysis

After cleaning, the data set contains a total of 10,228 opinions. Fig. 4 below shows the distribution of opinions by justice after cleaning.

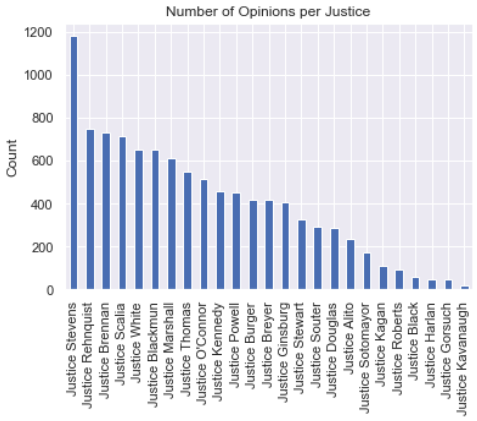

Fig. 4. Number of opinions per justice (after cleaning)

The most prolific writer by far would seem to be Justice Stephens, with 1181 opinions. Newer justices, such as Kavanaugh and Gorsuch, have the fewest opinions to their name, with 20 and 47 opinions, respectively. A handful of justices were also reaching the ends of their careers in the early 1970s, and have fewer opinions in this data set as a result. For example, Justices Harlan and Black both died in 1971, and each has fewer than 60 opinions in this dataset. Although this figure gives a sense of overall output of each justice over the course of their careers on the Supreme Court, it doesn't say much about their relative productivity in terms of the amount of time served.

Fig. 5 below shows the breakdown of number of opinions by category after cleaning. After cleaning, there are

4848 majority, 3554 dissenting, and 1826 concurring opinions in this data set.

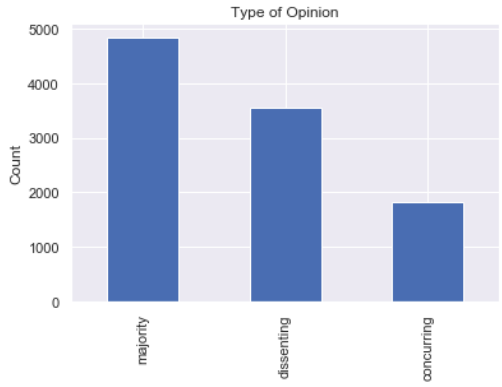

Fig. 5. Opinions by category (after cleaning)

Fig. 6 shows the distribution for opinion length (in number of characters). On the whole, it would seem that most opinions are shorter than 20,000 characters.

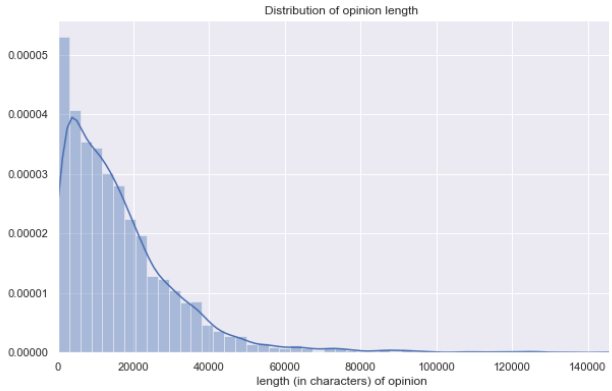

Fig. 6. Distribution of opinion length

Figs. 7 and 8 below show two different views of the number of opinions by category for each justice.

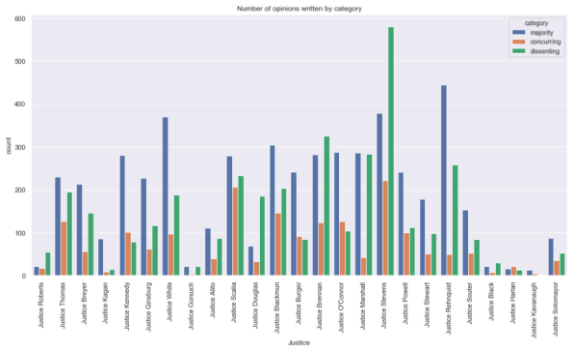

Fig. 7. Number of opinions written by category (relative to other justices)

Fig. 7 shows how justices stack up against one another in terms of type and number of opinions written. The graph has been sized down considerably and is unfortunately difficult to read in the context of this report, but does show some interesting trends. Majority opinions are indicated in blue, concurring in orange, and dissenting in green. Overall, it seems most justices write more majority opinions than any other category of opinion. A handful of justices have produced more dissenting opinions than any other category. These are Justices Stevens, Brennan, Douglas, and Roberts.

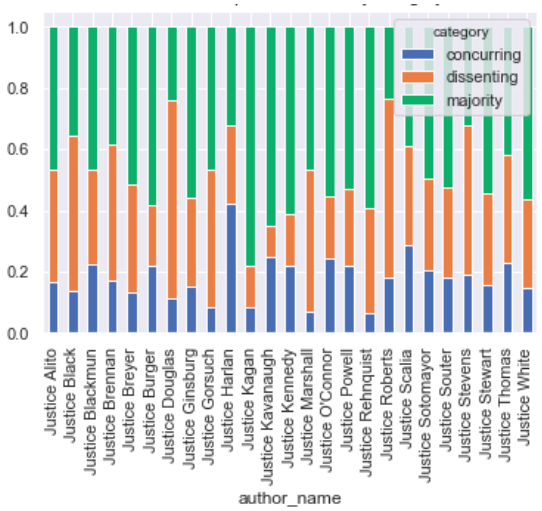

Fig. 8. Number of opinoins written by category (normalized)

Fig. 8 gives a normalized view of the same information. It gives a clearer picture of the proportion of opinion categories for each justice. It's important to note that the color-coding is different in this figure, with blue representing concurring opinions, orange representing dissenting opinions, and green representing majority opinions.

Fig. 9 below simply shows the average 'year_filed' for the opinions written by each justice in the data set. This information was mainly useful for imposing some sense of chronology on the results of this preliminary data analysis.

```
Justice Black        1970
Justice Harlan       1970
Justice Douglas      1972
Justice Stewart      1975
Justice Burger       1978
Justice Powell       1979
Justice Brennan      1980
Justice White        1981
Justice Marshall     1981
Justice Blackmun     1982
Justice Rehnquist    1985
Justice Stevens      1990
Justice O'Connor     1991
Justice Souter       1999
Justice Scalia       2000
Justice Kennedy      2001
Justice Ginsburg     2006
Justice Breyer       2007
Justice Thomas       2007
Justice Roberts      2011
Justice Alito        2013
Justice Kagan        2015
Justice Sotomayor    2015
Justice Gorsuch      2018
Justice Kavanaugh    2019
```

Fig. 9. Average year of opinions written by each justice

Fig. 10 below shows the average number of opinions each justice wrote per year and gives some idea of each justice's overall productivity relative to the amount of time they spent on the Supreme Court.

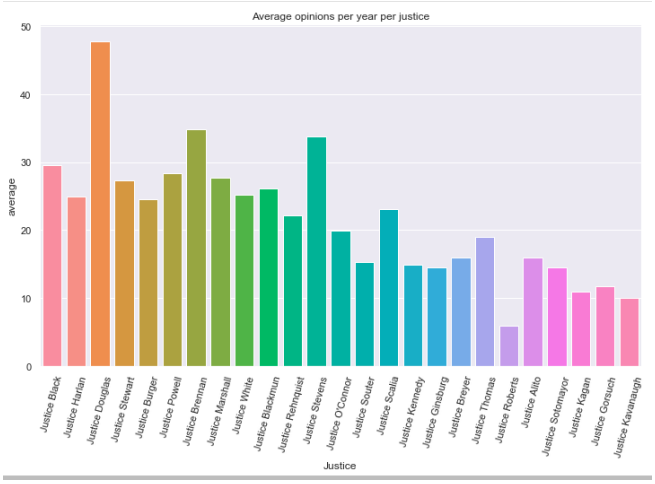

Fig. 10. Average opinions per year per justice

The justices in this table are ordered chronologically by average year of opinions written. Overall, the table seems to show a decreasing trend in average number of opinions written by each justice in recent years. Additionally, the justices with the highest average yearly output in Fig. 8 are not necessarily the same as the justices who wrote the most overall opinions, although there is some relation.

Fig. 11 below shows the average word count per opinion of each justice.

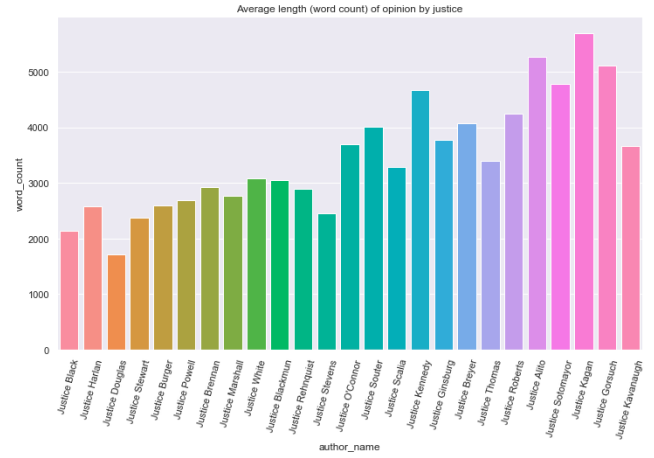

Fig. 11. Average word count of opinions per justice

Like Fig. 10, the justices in Fig. 11 are also ordered chronologically by average year of opinions filed. In contrast to Fig. 10, it appears that the average length (measured in word count) of opinions has been increasing over time. Additionally, the justices who wrote the most opinions did not necessary write very long opinions. Justice Douglas, who had the highest average in Fig. 10, seems to have written the shortest opinions on average of all the justices included in this data set, while Justices Kagan, Gorsuch, and Alito have the highest average word counts.

C. Data Preprocessing for Topic Modeling

There are a number of steps that need to be followed in preprocessing data for many natural language processing tasks, including topic modeling, which are detailed below. These steps are necessary to reduce noise and convert raw text into a form that is readable by the model and the computer.

- Convert words into lowercase
- Remove stop words (shorter words like articles, conjunctions, prepositions, etc.)
- Remove punctuation, symbols, and special characters
- Lemmatize words (truncates words down to their uninflected stems; for example, walks, walked, walking → walk-)

For this step, I used Python's nlkt library in Jupyter Lab to clean the contents of the "text" column (lowercased, removed stop words and punctuation, lemmatized) in the original data frame. For some reason, this ended up being one of the more time consuming aspects of the project.

There were a few additional steps I took in order to further prepare the data set for topic modeling. I ended up removing especially common (appearing in 10% of all documents) and especially rare (appearing fewer than 15 times) words from the data set. I realized this step was necessary the first few times I attempted to create topics. I found that there were a handful of words such as "court," "congress," "appeal," etc., that were appearing in the majority of topics and were making it difficult to interpret them. Given more time, I would also have experimented with removing all proper nouns from the data set for similar reasons. Occasionally, a name associated with a landmark case like "Miranda" or "Sherman" would appear in a topic, making it easier to interpret the relationship between the remaining words in the grouping. However, there were a few instances where the names of certain justices appeared in place of words that might have made it easier to interpret the essence of the topic.

After preprocessing and lemmatizing the textual data, the next step for LDA was to vectorize the text by creating a bag-of-words representation of the text. In the simplest terms, a bag of words is representation of the text that describes the occurrence or frequency of words within a document. This step in the process involved first creating a "dictionary" from the data set of lemmatized words that comprised every word appearing in the data set. The fir st10 words in the dictionary are shown in he figure below.

```
0 abid
1 abil
2 abl
3 abridg
4 absenc
5 absent
6 absolut
7 abus
8 accept
9 access
10 accomplish
```

Fig. 12. Dictionary created from training data

The bag of words was then created from this dictionary. The first 15 words in the bag of words for the first row in the data set are shown below. Each row includes a number referencing a word in the dictionary along with its frequency in the text of that specific opinion.

```
[(0, 1),
 (1, 4),
 (2, 4),
 (3, 1),
 (4, 1),
 (5, 3),
 (6, 4),
 (7, 1),
 (8, 1),
 (9, 74),
 (10, 1),
 (11, 1),
 (12, 1),
 (13, 1),
 (14, 3),
 (15, 2),
```

Fig. 13. Bag of words on its own

```
Word 0 ("abid") appears 1 time.
Word 1 ("abridg") appears 4 time.
Word 2 ("actor") appears 4 time.
Word 3 ("addi") appears 1 time.
Word 4 ("admonish") appears 1 time.
Word 5 ("advoc") appears 3 time.
Word 6 ("affili") appears 4 time.
Word 7 ("age") appears 1 time.
Word 8 ("agement") appears 1 time.
Word 9 ("aggreg") appears 74 time.
Word 10 ("aliti") appears 1 time.
Word 11 ("alli") appears 1 time.
Word 12 ("alvarez") appears 1 time.
Word 13 ("america") appears 1 time.
Word 14 ("amici") appears 3 time.
Word 15 ("analyz") appears 2 time.
```

Fig. 14. Dictionary words alongside bag of words frequencies

For ease of viewing, I looped through the bag of words and the dictionary to create an array that displayed their information side by side. The results for the first 15 words are shown in Fig. 14.

IV. Methods

I used the gensim python library to perform the actual LDA topic modeling. There were a number of parameters I had to consider when creating topics, and which I adjusted over the course of the project. They are detailed below

- **Num_topics:** the number of latent topics to be extracted from the corpus. Must be specified ahead of time with LDA
- **Alpha and Eta:** hyperparameters that affect the sparseness of the document-topic and topic-word distributions.
    - High alpha value: every document has a mixture of all topics (documents appear similar to each other).
    - Low alpha value: Every document has a mixture of very few topics.
    - High eta value: each topic has a mixture of most words (topics appear similar to each other).
    - Low eta value: each topic has a mixture of very few words.
- **Passes:** the number of training passes through the corpus. More passes usually resulted in more coherent topics.
- **Random_state:** Setting this = 1 significantly mitigated the problems I was having with consistency between runs, but did not completely eliminate them.

I attempted LDA with 5, 10, 20, and 30 topics; I kept alpha and beta at the default value (1/num_topic); and experimented with 10, 20, 30, and 40 passes through the corpus.

V. Results

A. Creation of Topics

I experimented with a few different inputs for num_topics as detailed in the previous section and generally found that a higher number of topics yielded better results. Fig. 15 shows the results of setting the number of topics at five with 10 passes. Smaller numbers tended to produce more vague and random-seeming groupings of words under each topic, which were then very difficult to interpret even after doing some research into relevant Supreme Court cases.

```
Topic: 0
Words: 0.008*"water" + 0.006*"bargain" + 0.006*"competit" + 0.0
06*"carrier" + 0.005*"antitrust" + 0.004*"abort" + 0.004*"rat"
+ 0.004*"railroad" + 0.004*"seizur" + 0.004*"privaci"


Topic: 1
Words: 0.009*"juror" + 0.009*"racial" + 0.008*"race" + 0.006*"s
tudent" + 0.006*"copyright" + 0.005*"guidelin" + 0.005*"parol"
+ 0.004*"juvenil" + 0.004*"inmat" + 0.004*"ordin"


Topic: 2
Words: 0.012*"habea" + 0.012*"arbitr" + 0.011*"patent" + 0.006
*"militari" + 0.005*"feloni" + 0.005*"candid" + 0.005*"water" +
0.005*"corpus" + 0.005*"aggrav" + 0.004*"voter"


Topic: 3
Words: 0.010*"alien" + 0.010*"fee" + 0.006*"fraud" + 0.006*"dis
closur" + 0.006*"taxpay" + 0.005*"jeopardi" + 0.005*"asset" +
0.005*"plea" + 0.005*"doubl" + 0.005*"grand"


Topic: 4
Words: 0.021*"indian" + 0.017*"tribe" + 0.017*"religi" + 0.013
*"child" + 0.013*"parent" + 0.010*"bankruptci" + 0.009*"treati"
+ 0.007*"religion" + 0.007*"tribal" + 0.006*"tax"
```

Fig. 15. Results of topic modeling with 5 clusters

As can be clearly seen, it is somewhat difficult to interpret the meaning of these five topics in Fig. 15. Topic 0, for example, casts a very wide net and appears to lump cases concerning business antitrust and competition laws in together with so-called "right to privacy" cases, such as abortion, which I'm not able to account for. The inclusion of "railroad" and "seizur" in this topic may have something to do with the 1989 case, *Skinner v. Railway Labor Executive Association*, which ruled that a drug- and alcohol-testing program for railroad employees did not violate the employees' rights under the fourth amendment, forbidding unreasonable searches and seizures of individuals and property. This could be construed as a right to privacy case, explaining its inclusion with "abort" in this cluster.

Next, Topic 1 is somewhat clearer and appears to include cases having to do with sentencing and the criminal justice system. "Copyright" seems to be a bit of an outlier in this topic cluster. Topic 2 is the most difficult to parse; it seems to include voting issues, patens, and a number of other subjects. Topic 3 probably encompasses cases having to do with financial and tax issues. Topic 4 is perhaps the clearest of all five topics and appears to pertain to rulings on the rights of Native peoples.

For purposes of comparison, Fig. 16 below shows the results of implementing LDA with 30 topics and 40 passes. An interpretation for each topic is also provided.

Topic 0: 0.029*"ordin" + 0.017*"obscen" + 0.013*"sexual" + 0.010*"adult" + 0.010*"vagu" + 0.008*"messag" + 0.008*"street" + 0.008*"park" + 0.007*"solicit" + 0.007*"film"	Obscenity
Topic 1: 0.029*"beneficiari" + 0.021*"reimburs" + 0.021*"erisa" + 0.020*"spous" + 0.017*"pension" + 0.016*"recipi" + 0.015*"retir" + 0.013*"wast" + 0.011*"afdc" + 0.011*"medicaid"	Pension, retirement benefis
Topic 2: 0.028*"taxpay" + 0.023*"asset" + 0.018*"invest" + 0.018*"deduct" + 0.017*"stock" + 0.015*"tax" + 0.013*"loan" + 0.013*"profit" + 0.013*"earn" + 0.013*"deposit"	Tax laws
Topic 3: 0.036*"retroact" + 0.024*"eleventh" + 0.022*"toll" + 0.018*"waiver" + 0.015*"retir" + 0.014*"adea" + 0.009*"chevron" + 0.007*"davi" + 0.007*"abrog" + 0.007*"young"	Unclear—11th ammendment rulings? ADEA = Age Discrimination in Employment Act. Chevron refers to 1984 case, *Chevron v. National Resources Defense Council*
Topic 4 : 0.041*"patent" + 0.012*"alito" + 0.011*"sion" + 0.009*"breyer" + 0.009*"sotomayor" + 0.007*"tive" + 0.007*"tional" + 0.006*"invent" + 0.006*"ginsburg" + 0.004*"robert"	Patents? Also unclear
Topic 5: 0.041*"parol" + 0.036*"mitig" + 0.034*"eighth" + 0.033*"juvenil" + 0.032*"aggrav" + 0.017*"kill" + 0.015*"graham" + 0.012*"cruel" + 0.011*"adult" + 0.010*"penri"	Criminal justice; sentencing for violent crimes; treatment and sentencing of juviniles; Eigth Ammendment rights – prohibiting cruel and unusal punishment; excessive use of force by police officers (*Graham v. Connor*, 1989)
Topic 6: 0.036*"competit" + 0.031*"antitrust" + 0.023*"manufactur" + 0.017*"emption" + 0.016*"retail" + 0.016*"empt" + 0.015*"sherman" + 0.012*"liquor" + 0.011*"dealer" + 0.010*"competitor"	Antitrust laws
Topic 7: 0.070*"water" + 0.026*"river" + 0.017*"ferc" + 0.016*"master" + 0.015*"pollut" + 0.015*"electr" + 0.015*"project" + 0.014*"compact" + 0.012*"plant" + 0.012*"environment"	Environmental safety/pollution
Topic 8: 0.088*"alien" + 0.038*"immigr" + 0.032*"detent" + 0.030*"deport" + 0.020*"preclus" + 0.019*"detain" + 0.017*"subsect" + 0.014*"citizenship" + 0.013*"pretrial" + 0.011*"estoppel"	Immigration laws
Topic 9: 0.034*"vehicl" + 0.022*"mail" + 0.020*"travel" + 0.016*"driver" + 0.014*"passeng" + 0.013*"postal" + 0.013*"accid" + 0.012*"highway" + 0.012*"traffic" + 0.011*"motor"	Driving, the mail?
Topic 10: 0.046*"fee" + 0.022*"rico" + 0.021*"punit" + 0.013*"puerto" + 0.012*"tort" + 0.011*"municip" + 0.011*"territori" + 0.009*"eeoc" + 0.007*"biven" + 0.007*"color"	Puerto Rico?
Topic 11: 0.088*"arbitr" + 0.064*"bargain" + 0.030*"grand" + 0.019*"settlement" + 0.017*"grievanc" + 0.016*"negoti" + 0.015*"nlrb" + 0.014*"breach" + 0.010*"card" + 0.009*"nlra"	The workplace (NLRB = national labor relations board; NLRA = National Labor Relations Law)
Topic 12: 0.079*"child" + 0.061*"parent" + 0.022*"classif" + 0.022*"father" + 0.017*"mother" + 0.015*"custodi" + 0.009*"marriag" + 0.008*"women" + 0.008*"birth" + 0.007*"illegitim"	Laws about family, marriage, custody of children
Topic 13: 0.023*"appelle" + 0.013*"decre" + 0.011*"certif" + 0.010*"declaratori" + 0.009*"moot" + 0.009*"preliminari" + 0.007*"affidavit" + 0.007*"exhaust" + 0.006*"indig" + 0.006*"suspens"	Unclear; random legal terminology
Topic 14: 0.027*"copyright" + 0.021*"seizur" + 0.017*"privaci" + 0.017*"miranda" + 0.014*"forfeitur" + 0.013*"interrog" + 0.013*"incrimin" + 0.011*"seiz" + 0.010*"custodi" + 0.010*"suppress"	Miranda rights, police interrogations and seizures, arrests, right to privacy (copyright is a weird outlier)
Topic 15: 0.080*"religi" + 0.047*"student" + 0.037*"religion" + 0.023*"church" + 0.020*"teacher" + 0.012*"secular" + 0.011*"colleg" + 0.010*"teach" + 0.007*"sectarian" + 0.006*"accommod"	Religion in schools
Topic 16: 0.022*"inmat" + 0.018*"punit" + 0.014*"tort" + 0.008*"liquor" + 0.007*"neglig" + 0.005*"compensatori" + 0.005*"recoveri" + 0.005*"color" + 0.005*"venu" + 0.005*"alcohol"	Serving alcohol?
Topic 17: 0.047*"disabl" + 0.025*"claimant" + 0.019*"supervisor" + 0.016*"harass" + 0.016*"eeoc" + 0.013*"retali" + 0.012*"tort" + 0.011*"veteran" + 0.011*"restitut" + 0.010*"director"	Workplace discriminaion; EEOC = equal employment opportunity Commission; disability rights
Topic 18: 0.031*"candid" + 0.018*"voter" + 0.017*"broadcast" + 0.014*"deleg" + 0.013*"ballot" + 0.011*"campaign" + 0.010*"expenditur" + 0.009*"elector" + 0.009*"democrat" + 0.008*"convent"	Elections, campaigning, voting
Topic 19: 0.027*"feloni" + 0.026*"plea" + 0.021*"firearm" + 0.015*"conspiraci" + 0.011*"violenc" + 0.010*"robberi" + 0.009*"burglari" + 0.009*"assault" + 0.008*"misdemeanor" + 0.008*"arm"	Crime, sentencing
Topic 20: 0.043*"abort" + 0.042*"jeopardi" + 0.039*"doubl" + 0.025*"patient" + 0.022*"hospit" + 0.021*"physician" + 0.015*"acquitt" + 0.013*"woman" + 0.013*"pregnanc" + 0.012*"clinic"	Abortion, healthcare
Topic 21: 0.049*"tax" + 0.030*"registr" + 0.023*"taxat" + 0.018*"michigan" + 0.017*"maryland" + 0.016*"regist" + 0.014*"jersey" + 0.012*"export" + 0.011*"vermont" + 0.011*"domest"	Taxes on imports and exports from other states.
Topic 22: 0.034*"disclosur" + 0.033*"fraud" + 0.024*"advertis" + 0.018*"client" + 0.015*"solicit" + 0.013*"discoveri" + 0.013*"fiduciari" + 0.012*"profession" + 0.011*"confidenti" + 0.010*"fraudul"	False advertising, fraud, financial issus
Topic 23: 0.043*"habea" + 0.034*"juror" + 0.016*"corpus" + 0.011*"guilt" + 0.010*"collater" + 0.008*"harmless" + 0.008*"magistr" + 0.006*"default" + 0.006*"confess" + 0.006*"ineffect"	Habeas corpus – unclear
Topic 24: 0.049*"militari" + 0.027*"vessel" + 0.022*"contractor" + 0.022*"maritim" + 0.020*"ship" + 0.014*"admiralti" + 0.013*"martial" + 0.010*"picket" + 0.009*"neglig" + 0.008*"port"	Maritime law

Topic 25: 0.053*"racial" + 0.042*"race" + 0.017*"segreg" + 0.016*"popul" + 0.013*"voter" + 0.013*"student" + 0.012*"negro" + 0.010*"desegreg" + 0.009*"redistrict" + 0.007*"preclear"	Civil rights, segregation, racial jusice
Topic 26: 0.063*"guidelin" + 0.026*"kansa" + 0.024*"probat" + 0.021*"maximum" + 0.016*"supervis" + 0.012*"fine" + 0.012*"rehabilit" + 0.012*"departur" + 0.010*"cocain" + 0.009*"revoc"	Unclear—Drug related offenses?
Topic 27: 0.025*"student" + 0.023*"forfeitur" + 0.014*"messag" + 0.013*"solicit" + 0.012*"picket" + 0.011*"speaker" + 0.010*"street" + 0.010*"viewpoint" + 0.009*"park"	Unclear—first ammendment, right of free expression? Protest?
Topic 28: 0.057*"bankruptci" + 0.041*"carrier" + 0.035*"debtor" + 0.033*"railroad" + 0.030*"debt" + 0.027*"creditor" + 0.020*"chapter" + 0.018*"rat" + 0.018*"lien" + 0.016*"truste"	Bankruptcy, debt
Topic 29: 0.108*"indian" + 0.083*"tribe" + 0.044*"treati" + 0.036*"tribal" + 0.021*"convent" + 0.016*"fish" + 0.012*"oklahoma" + 0.009*"montana" + 0.009*"territori" + 0.009*"allot"	Native rights

Fig. 16. LDA with 30 topics and interpretations (40 runs)

All but three of the topics shown in Fig. 16 above were easily interpreted. It's also worth mentioning that when I was starting out with LDA, I noticed that the model was not producing stable results between runs. At first, different runs, even with the same number of topics, did produce somewhat different groupings and results. Nevertheless, some topics did persist across runs, or at least continued to appear in very similar groupings. I'm not able to account for why this was happening, but I was finally able to mitigate the problem by setting the random_state variable to 1. Afterwards, topics were much more consistent across runs, although there was slight variation in the ordering of topics. For example, Topic 14, which has to do with police interrogations and seizures, was labeled as Topic 5 in a later run. Generally, however, he content of the topics remained the same.

Although it would take up too much space to reproduce the results of each run in this paper, it's also worth noting that running LDA with different settings produced results that differed in interesting ways. For example, depending on the number of topics selected, runs grouped abortion either with medical issues or with fourth amendment rights and search legality. The latter grouping likely represents the fact that abortion has traditionally been treated in the courts as a right to privacy issue. This demonstrates that the parameters of LDA are very important—the groupings can be unstable across different runs if "passes" is set too low. However, the variability in groupings also represents legitimate conceptual ambiguities. A group of 20 humans might also come up with just as many different ways of grouping the topics represented by the 10,000 opinions in this data set. Is abortion a medical issue or a right to privacy issue, legally speaking? Should trade be grouped with maritime law, or with immigration? It was interesting that LDA seemed to differ on some of the same issues that human interpreters of these texts might reasonably differ on.

Additionally, while almost every run included a topic cluster for criminal cases or sentencing and some version of finance and bankruptcy cases, the set of topic clusters varied somewhat. Native American cases, religion, obscenity, segregation, immigration law, and maritime law all appeared in some runs and not others, further underscoring the importance of choosing appropriate values for the each parameter.

B. Classifying Topics of Unseen Documents

After creating stable topics, I experimented with feeding randomly selected opinions and a few well-known landmark opinions from the testing data set into model to see if it would then correctly categorize them. A few examples are shown below with the top two topic matches. In most cases, the model performed fairly well. Even in cases where the closest match was incorrect, it was usually easy to figure out why that topic was chosen. Some examples are shown in the figures below.

Score: 0.32833701372146606	**Topic 14:** 0.027*"copyright" + 0.021*"seizur" + 0.017*"privaci" + 0.017*"miranda" + 0.014*"forfeitur" + 0.013*"interrog" + 0.013*"incrimin" + 0.011*"seiz" + 0.010*"custodi" + 0.010*"suppress"
Score: 0.19986803829669952	**Topic 5:** 0.041*"parol" + 0.036*"mitig" + 0.034*"eighth" + 0.033*"juvenil" + 0.032*"aggrav" + 0.017*"kill" + 0.015*"graham" + 0.012*"cruel" + 0.011*"adult" + 0.010*"penri"

Fig. 17. South Carolina v. Demetrius Gathers

*South Carolina v. Demetrius Gathe*rs is a 1989 murder case which held that testimony in the form of a victim impact statement is admissible during the sentencing phase of the trial only if it directly relates to the circumstances of the crime. It was later overruled in *Payne v. Tennessee* (1991). The closest topic matches proposed by the model were 14 and 5, which both have to do with criminal justice. Topic 14 ostensibly relates to Miranda rights, police interrogations and seizures, arrests. In broad terms, it might be said that the topic encompasses cases having to do with the rights of the individual in the criminal justice system, so perhaps it can also be extended to accommodate cases concerning the admissibility of certain forms of evidence in the court of law. Topic 5 also relates to criminal justice and sentencing for various crimes, and its relevance here is obvious.

Score: 0.4639028310775757	**Topic 15**: 0.080*"religi" + 0.047*"student" + 0.037*"religion" + 0.023*"church" + 0.020*"teacher" + 0.012*"secular" + 0.011*"colleg" + 0.010*"teach" + 0.007*"sectarian" + 0.006*"accommod"
Score: 0.25734126567840576	**Topic 18:** 0.031*"candid" + 0.018*"voter" + 0.017*"broadcast" + 0.014*"deleg" + 0.013*"ballot" + 0.011*"campaign" + 0.010*"expenditur" + 0.009*"elector" + 0.009*"democrat" + 0.008*"convent"

Fig. 18. McCutcheon v. FEC

McCutcheon v. Federal Election Committee is a 2014 ruling that struck down the aggregate limits on the amount an individual may contribute during a two-year period to all federal candidates, parties, and political action committees combined. It is a landmark ruling on campaign finance. The court held that limitations on aggregate contributions to campaign finances represented a restriction on an individual's political expression and therefore violated the free speech clause of the First Amendment. The model proposed topics 15 and 18 as the closest matches for this case. At first glance, Topic 18 seems to be the clear best fit for this ruling, as it has to do with voting and elections. The relevance of Topic 15, which has to do with the intersection of religion and secular institutions like schools, is not immediately clear. However, many Supreme Court rulings involving religion invoke the First Amendment, which protects freedom of religion in addition to freedom of speech, so it is easy to see why there might be some overlap here, especially since the model did not generate a particularly strong category for freedom of speech on its own.

Score: 0.710432767868042	**Topic 20:** 0.043*"abort" + 0.042*"jeopardi" + 0.039*"doubl" + 0.025*"patient" + 0.022*"hospit" + 0.021*"physician" + 0.015*"acquitt" + 0.013*"woman" + 0.013*"pregnanc" + 0.012*"clinic"
Score: 0.08293242752552032	**Topic 12:** 0.079*"child" + 0.061*"parent" + 0.022*"classif" + 0.022*"father" + 0.017*"mother" + 0.015*"custodi" + 0.009*"marriag" + 0.008*"women" + 0.008*"birth" + 0.007*"illegitim"

Fig. 19. Roe v. Wade

In addition to feeding random cases into the model, I also tried out a few well-known ones. *Roe v. Wade* was the 1973 ruling that protects a pregnant woman's liberty to choose to have an abortion without excessive government restriction on the grounds that the Fourteenth Amendment promises an individual the fundamental right to privacy. Here, the model proposed Topics 20 and 12 as the two closest matches. Topic 20 has to do with healthcare and abortion, and Topic 12 seems to have something to do with family law. "Mother," "child", "women", and "birth" are all important words in this grouping, so it's easy to understand why it was suggested as a possible match for *Roe v. Wade*.

Score: 0.7574606537818909	**Topic 25**: 0.053*"racial" + 0.042*"race" + 0.017*"segreg" + 0.016*"popul" + 0.013*"voter" + 0.013*"student" + 0.010*"desegreg" + 0.009*"redistrict" + 0.007*"preclear"
Score: 0.10395928472280502	**Topic 18:** 0.031*"candid" + 0.018*"voter" + 0.017*"broadcast" + 0.014*"deleg" + 0.013*"ballot" + 0.011*"campaign" + 0.010*"expenditur" + 0.009*"elector" + 0.009*"democrat" + 0.008*"convent"

Fig. 20. Milliken v. Bradley

Milliken v. Bradley was a 1974 ruling that dealt with the planned desegregation busing of public school students across district lines in Detroit. The District Court of Detroit attempted to remedy the racial imbalance of local schools by redrawing lines of suburban school districts to achieve balance within the city's schools. The Supreme Court ultimately ruled that school district lines cannot be redrawn for the purposes of combating segregation unless the segregation was the product of discriminatory acts by school districts. The rulings effects have been long lasting, and it is cited as the main reason why U.S. schools remain so segregated nearly 70 years after *Brown v. the Board of Education*. The model correctly identified this as a case primarily about race and desegregation.

In general, the model struggled to categorize shorter opinions. It also seemed to have some difficulty categorizing freedom of speech cases, perhaps because it didn't generate a distinct enough topic corresponding to that issue. For example, cases that had to do with first amendment rights were sometimes categorized under Topic 18, which has to do with religion. As the First Amendment protects both freedom of speech and freedom of religion, it is easy to see how this happened. The model also seemed to have difficulty with cases involving various aspects of the criminal justice system, since there are three categories that deal with these kinds of cases and that have a certain amount of overlap. I also noticed that opinions on the same case (e.g., the majority opinion vs. a dissenting opinion) were sometimes categorized slightly differently, perhaps because the authors appealed to different legal precedents in justifying their respective positions. I am sure that given more time to experiment, other interesting patterns would have emerged, since the 30 topics I generated in no way encompass every topic covered in this body of work.

C. Visualization of Topic Modeling Results

After the topics were created and tested, I then experimented with using them as tools to analyze the data set. It worth noting that the results in this section are based on a different LDA run than the results presented in sections A and B. Although the content of topics themselves are the same across runs, they are numbered differently. Additionally, as the model does not label topics with 100% accuracy, the results in this section should be taken with the grain of salt. Nevertheless, they provide an example of additional uses for topic modeling. In order to analyze the opinions in the data set in terms of their potential topics, I added a column in the training data set called "main_topic," which contained the most likely topic match for each opinion. Fig. 21 below shows the overall frequency of each topic in the data set.

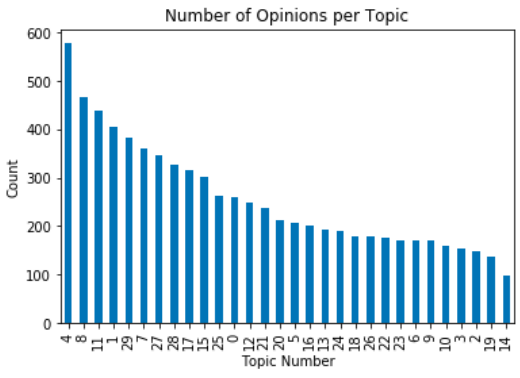

Fig. 21. Overal Frequency of Topics

Nearly 600 opinions were categorized as belonging to Topic 4, which in this instance encompassed fraud, bankruptcy, and financial issues. In second place was Topic 8, which encompassed cases having to do with religious issues. Topic 11 included workplace discrimination and disability cases. Finally, Topic 1 had to do with Native American tribal rights.

Figs. 22–27 give a sense of opinion topic frequencies by decade.

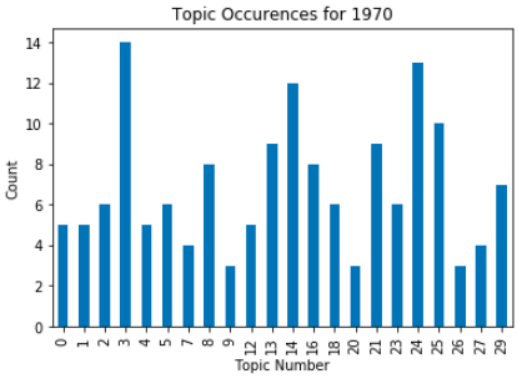

Fig. 22. Topic Frequencies for 1970

In 1970, the court seems to have seen a lot of cases about police seizures and arrests (Topic 3) and healthcare and abortion (Topic 14). Topic 24 is difficult to interpret and corresponds to Topic 23 in Section A. Important words for this topic include "habea" "corpus" "plea" "collater." It could possibly be interpreted as encompassing issues having to do with certain legal procedures.

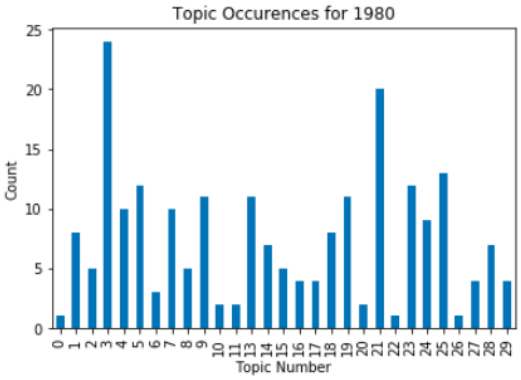

Fig. 23. Topic Frequencies for 1980

In 1980, Topic 3 was again dominant, followed by Topic 21 (liquor laws, Topic 16 in Section A) and 25 (obscenity, Topic 0 in Section A).

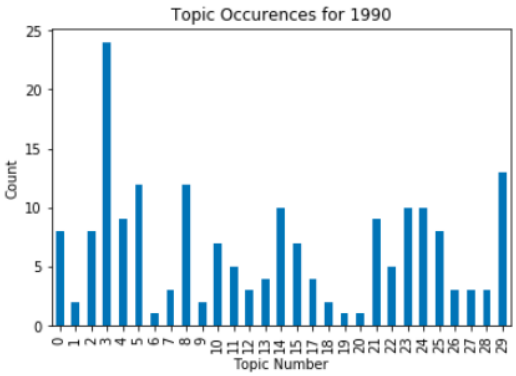

Fig. 24. Topic Frequencies for 1990

In 1990, Topic 3 was again the dominant category, followed distantly by Topics 5 (also related to police seizures, warrants, privacy), 6 (family), and 29 (railroads).

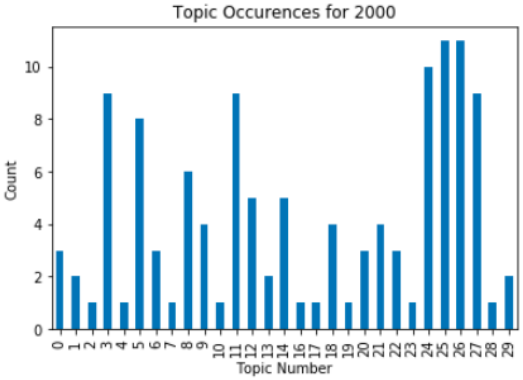

Fig. 25. Topic Frequencies for 2000

In 2000, Topic 3 is still prominent, but has been overtaken by 24 (vague, legal procedures?), 25 (obscenity), 26 (voting and elections), and 27 (criminal justice, sentencing).

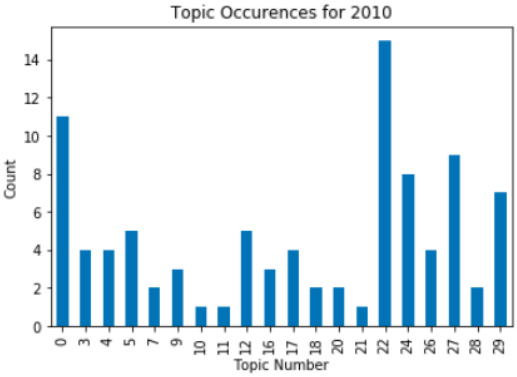

Fig. 26. Topic Frequencies for 2010

In 2010, the most common topics included 0 (bankruptcy), 22 (immigration), and 27 (criminal justice, sentencing). The prominence of bankruptcy cases possibly reflects the fall out of the 2008-2009 financial crisis.

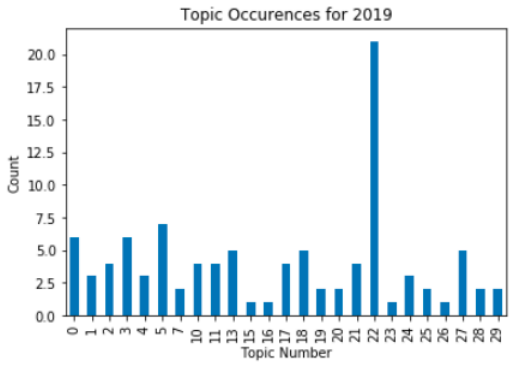

Fig. 27. Topic Frequencies for 2019

Finally for 2019, the last year in the data set, the most frequent topic by a long shot was 22 (immigration). Its prominence in this data set is mirrored in the news cycle of the past few years, as the issue has attracted increased attention during the Trump presidency.

Overall, these figures represent only a small sample of the various ways that the data set could be analyzed in terms of opinion topics. They give some sense of the country's shifting priorities over the past few decades, although a more accurate picture might be obtained by plotting the rise and fall in popularity of individual topics over the time period from 1970–2019.

VI. Conclusion

I really enjoyed this project and learned a lot about topic modeling, specifically LDA, and its capabilities and limitations. I found that it was a useful and interesting way to become acquainted with large body of text, specifically the possible topics contained in it and how they overlap and relate to one another. By applying topic modeling to a data set of Supreme Court opinions specifically, I learned a lot about the way that various issues are framed, the precedents that are often invoked in making certain kinds of arguments, and inner workings of the Supreme Court as an institution. It was interesting to examine the categories that the LDA model created to try to decipher the connections that the algorithm was making between various subjects covered in the opinions.

Given more time, there is a fair amount of work that could still be done with these topics. On the level of preprocessing, I would go back and spend more time trying to find a working method to drop named entities from the lemmatized text of the opinions. I would also spend more time experimenting with the settings for various parameters such as alpha and eta to see if there was any noticeable improvement in the quality of topics produced. I suspect that choosing slightly lower values for both parameters might produce more cohesive topic groupings.

Finally, additional analysis could be done on the topics themselves. I did some preliminary work in this direction, but some interesting next steps might be to try to plot the rise and fall of a topic's popularity over time and try to account for these fluctuations based on specific events in the news cycle. There is also much that could be done by looking at the breakdown of topics for each of the various justices in the data set. It could also be worthwhile to compare the performance of different topic modeling approaches such as Latent Semantic Analysis on this data set to see if one technique produced more stable or more interpretable results.

References

[1] Fiddler, G. (2017). *SCOTUS opinions* [Data set]. Kaggle. https://www.kaggle.com/gqfiddler/scotus-opinions

[2] American Bar Asssociation. (2018, November 27). *How to read a Supreme Court opinion.* https://www.americanbar.org/groups/public_education/publications/teaching-legal-docs/how-to-read-a-u-s--supreme-court-opinion/#:~:text=Main%20Opinion,-Following%20the%20syllabus&text=In%20legal%20terms%2C%20the%20opinion,the%20justices%20decide%20certain%20issues.

[3] United States Courts (n.d.). *About the Supreme Courts.* https://www.uscourts.gov/about-federal-courts/educational-resources/about-educational-outreach/activity-resources/about

[4] Supreme Court of the United States. (n.d.). *Opinions.* https://www.supremecourt.gov/opinions/opinions.aspx

[5] https://www.ideals.illinois.edu/bitstream/handle/2142/46405/ParallelTopicModels.pdf?sequence=2&isAllowed=y

[6] https://thesai.org/Downloads/Volume6No1/Paper_21-A_Survey_of_Topic_Modeling_in_Text_Mining.pdf

[7] towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

[8] www.toptal.com/python/topic-modeling-python

[9] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#8tokenizewordsandcleanuptextusingsimple_preprocess

[10] https://www.nytimes.com/2022/05/04/arts/roe-v-wade-abortion-history.html

[11] https://radimrehurek.com/gensim/models/ldamodel.html

[12] https://www.researchgate.net/publication/321804167_A_Guide_to_Text_Analysis_with_Latent_Semantic_Analysis_in_R_with_Annotated_Code_Studying_Online_Reviews_and_the_Stack_Exchange_Community

[13] https://ai.stanford.edu/~ang/papers/jair03-lda.pdf

[14] https://www.tdktech.com/tech-talks/topic-modeling-explained-lda-to-bayesian-inference/

[15] https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05

[16] https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/