NEW IN PAPERBACK

Fighting for black America

The day after winning the world heavyweight boxing championship in 1964, Cassius Clay announced his membership of the black Muslim group the Nation of Islam. A few weeks later he had a new name: Muhammad Ali. "And within a couple of years," writes David Remnick in *King of the World* (Picador £8.99/Vintage $13, 528 pages), "he would make out of himself a fast and funny kid from Louisville, Kentucky, one of the most compelling and electric American figures of the age."

This act of self-invention – "Who made me to me," Ali

covers the early stages of Ali's boxing career, focusing on his rivalry with the two men who preceded him as champion: Sonny Liston and Floyd Patterson. These three heavyweights, according to Remnick, represented competing visions of black America – "they seemed to mark the political and racial changes going on just as they were fighting one another for the title." Or as Ali tells the author: "I had to prove you could be a new kind of black man."

In the schema Remnick describes, Patterson stood for "the Good Negro" – "a deferential champion of civil rights, integration and

Liston, whose criminal record and Mafia connections led to his being easily caricatured as "the Bad Negro". But Ali, "this bizarre combination of loose-limbed athleticism and blatant showmanship", confounded the stereotypes. He recited poems, taunted opponents ("Chump! Big ugly bear!"), endorsed Malcolm X, and met the Beatles. Generational differences were underlined when Liston went to see the same band: "My dog plays drums better than that kid with the big nose," was his judgment.

So Ali was younger, whiter and more political than his immediate predecessors. Better at boxing too: he beat Liston in two controversial bouts, and Patterson once, between 1964 and 1965. These encounters form the spine of the book's narrative, alongside well-drawn

Sample Project 1: Black America and the Law in the Mid-20th Century

Overview

The mid-twentieth century in the United States was a time of immense transformation for people of color, particularly African Americans. Often referred to as the civil rights era, the 1960s and early 1970s saw protests, riots, violence, and eventually legislative responses to racialized injustice and discrimination. Numerous men and women fought to change how American society treated and understood the place of people of color, whether on buses, in streets, at home, in schools, or at work. Segregation was one of the main points of contention, and the focus of considerable legal effort. At the same time, cases involving African Americans proceeded through the legal system. An examination of the U.S. Supreme Court Records and Briefs reveals the effects of the pressures of the era on the legal system as well as those involved in civil rights cases.

1. Ideate

What are the core research questions?

How does the language about Black Americans shift in legal documents and records between 1950 and 1980?

What are other more precise, relevant questions?

What are the most common phrases or collocates used in documents mentioning Negro, Black, or African Americans?

What topics appear most frequently in documents mentioning Negro, Black, or African Americans?

Do these reflect themes that dominated Civil Rights-era conflicts?

What are the main concerns of legal records mentioning Negro, Black, or African Americans during this period?

Are there any specific states, statutes, or other entities that stand out amidst the analyses?

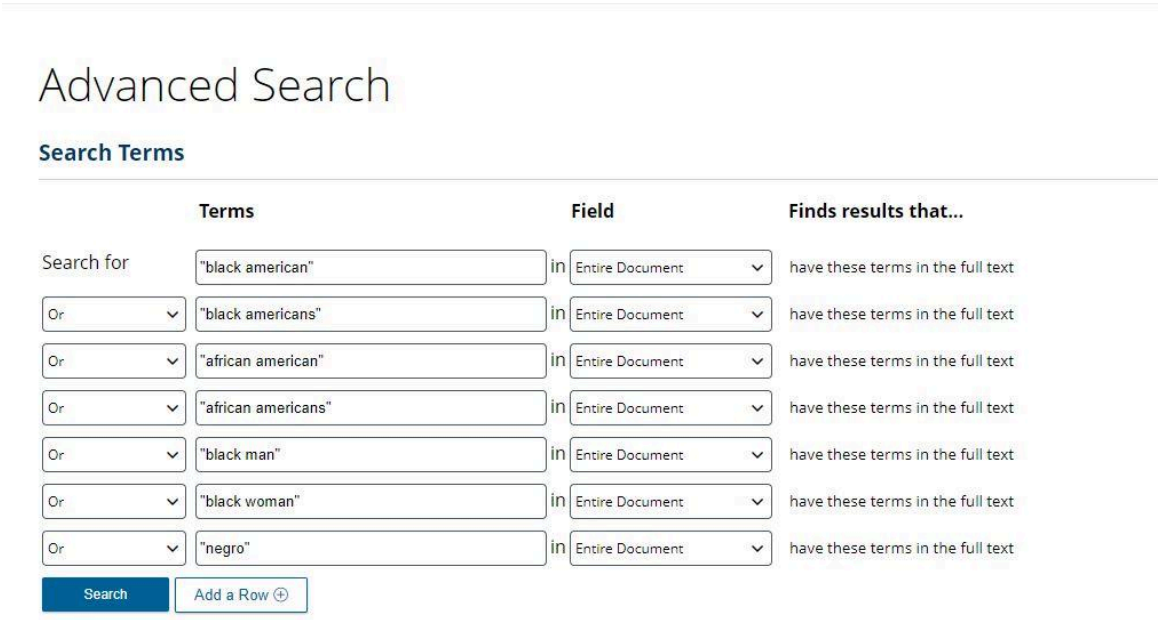
Are there any differences among Document Types in the Archive (U.S. Supreme Court Records and Briefs)?

2. Build

Steps

2.1 Searching

The search for this content set was limited to one Archive and a set Publication Date.



- Search Terms: black american, black man, black woman, negro, african american, african americans, black americans (Search for each keyword within the Entire Document field using the OR operator.)
- Selected Database to Search: U.S. Supreme Court Records and Briefs, 1832-1978
- Search Limiters - by publication year(s): Between 1950 - 1980
- Search Limiters - by content type: Monographs

2.2.Specific Tools

None of the tools required specific content sets.

2.3 Specific Questions

Question 6 could not be answered using the main content set, and so it required creating additional sub-content sets divided by Document Types: Briefs & Petitions; Statements, Memoranda, Appendices, etc.

3. Clean

Several attempts, or iterations, were necessary to get the cleaning right for each analysis. In the end, the project required several different Cleaning Configurations, as there were different stop words needed for different tools, as well as different approaches to punctuation. This required additional stop words that were unique to Topic Modeling and were not the same as those used in Ngrams; “state” and “court”, for instance, needed to be retained for Ngrams (for “United States”), but removed for Topic Modeling, which treats individual tokens. (For example, “United States” can never occur in a Topic Model because it includes two words. The software does not operate on phrases). The main stop word list required adding single letters (for each letter, in case they appeared as abbreviations), as well as additional stop words.

4. Analyze

Selecting particular views for each tool was extremely straightforward. The size of the content set suggested a search that (1) looked for more things; and (2) raised the bar for what made the cut for the results. Both were for very simple reasons:

Topic Modeling as a tool statistically discerns what words are more likely to appear near to one another. More topics, and more words, lowers the threshold of what is “significant”, presenting a finer grained picture of what the statistical analysis could suggest. In very similar documents, such as court records, the chance is that there will be similar phrases as questions and answers are posed, and rulings and arguments recorded.

Selecting more words and more topics is a good way to sift through some of these “known” similarities, and can work in tandem with stop word lists to help “drill down” into a large content set.

For Ngrams, a similar approach to thinking about potential “noise” helps to balance between the number of results with results that are meaningful to the user. There's a balance between number and noise.

Tools Used

4.1 Topic Modeling

It seemed best to cast a wider net in part to see what kinds of words appeared in the models created by the MALLET software that powers the tool. Requesting more words than the default, and double the topics, produces finer grained topics, reflecting the size of the content set. This example used 15 word topics and 20 topics.

4.2 Ngrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of the content set. Therefore, the threshold for the number of times an Ngram had to appear to be considered useful was raised to 4. Similarly, a desire to find collocates rather than just single words prompted a setting of the minimum Ngram size to 2 (biGram) and the maximum size to 5. These settings translate into a search for "Ngram of between 2 to 5 words that appear in documents at least 4 or more times."

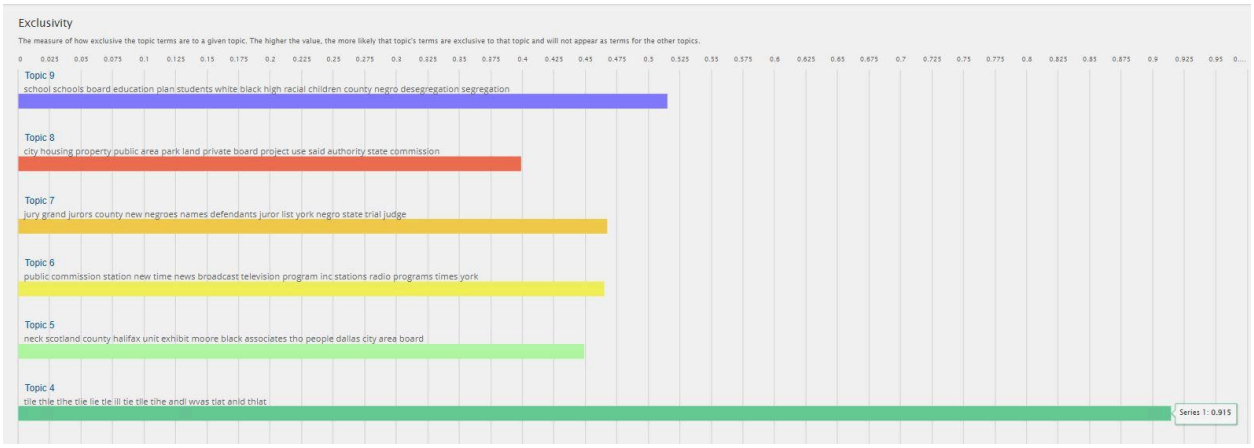
Understanding Results

Determining meaningfulness or significance is a critical part of scholarly inquiry. An essential element of this in a computational text analysis environment like Gale Digital Scholar Lab is understanding that raw counts or "more hits" does not always mean something important; it could be noise, meaning it is simply there because the content set was not cleaned enough, or the right stop words were not used, or perhaps it is something expected and known. In the end, understanding results really requires understanding what is being asked of the tools in the configurations and settings, and how the results relate to the selected variables and the algorithms that power the analysis.

Results by Tool

Topic Modeling

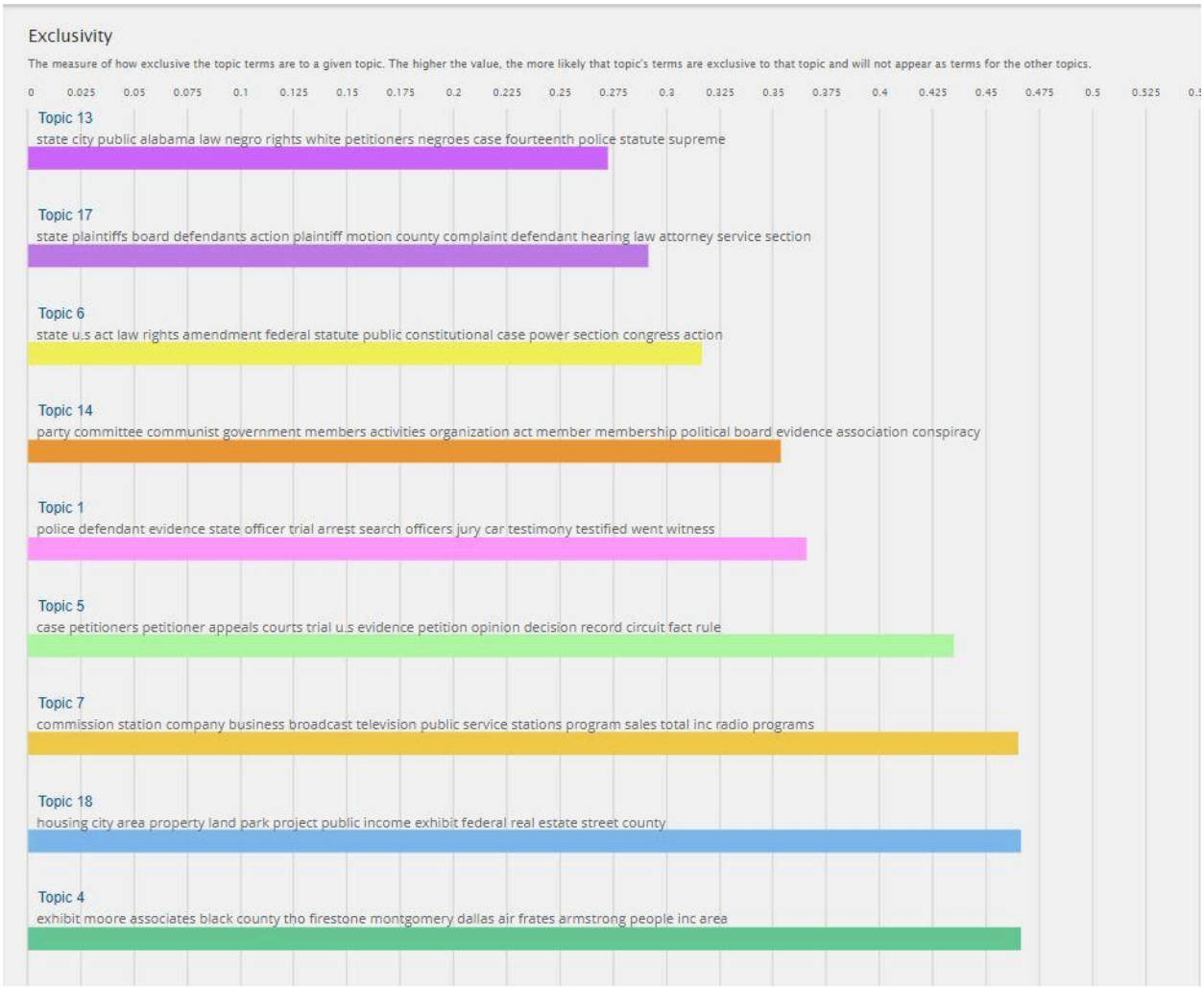
Refining the stop word list for topic modeling took some time, as the normal stop word list did not include Optical Character Recognition (OCR) error words. This is an example of the kinds of problems that can occur—notice Topic 4:



Revision of the Cleaning Configuration to add “tihe”, “andl”, “anld”, “that”, “tlat”, “tliat”, “thie”, “tlhat”, “inl”, removed this topic upon re-run; yet it produced more topics with unuseful words.

Topic Modeling reveals a series of possible themes within the U.S. Supreme Court Records:

- city, state, public, police, white, petitioners, peace, alabama, law, people
- party, communist, committee, government, bridges, testimony, member, union, activities, members
- county, vote, election, voting, state, districts, city, population, voters, political
- state, plaintiffs, defendants, action, plaintiff, defendant, motion, complaint, county, law
- school, schools, board, education, plan, students, black, white, high, racial
- jury, grand, jurors, county, negroes, defendants, state, juror, names, trial
- state, u.s, act, rights, law, amendment, case, federal, public, statute



Deciding what is the most meaningful or significant measure in these results can be tricky, depending on what the question might be, of course. The results from Topic Modeling could be meaningful simply by being unexpected or new. At the same time, they might confirm something already known, thereby acting as a touchstone to confirm that the user is on the right course with reading or analysis, or both. As can be seen in the topics returned above, some are clearly relevant to the civil rights era. Some may not be. There are other ways of understanding these results in relation to the content set, however.

The software powering the Topic Modeling tool, MALLET, is particularly well known and refined. It offers very rich results, which can be investigated in a number of ways by selecting different filters in the tool's legend and by examining results through the tool's

inspect panel. A list of measures describing how the topics relate to the content set and the analysis are included in the tool.

Tokens: This metric measures the number of words from the content set assigned to this topic.

Document Entropy: This metric measures the probability any given document will be in the topic. Low entropy topics will come from a small set of documents while higher entropy topics will come from a wider set of documents.

Average Word Length: This metric measures the average number of characters in the top terms. Longer words are assumed to be more meaningful, so higher word lengths indicate more specific topics.

Coherence: This metric measures how often words in the topic appear next to each other. The closer to 0, the more likely it is that terms occur next to each other.

Uniform Distance: This metric measures the distance between a uniform distribution and that of the topic's distribution over the words assigned to it. The larger the distance, the more specific the topic.

Corpus Distance: This metric measures the distance between the frequency of words in the content set to frequency of the words assigned to the topic. The larger the distance, the more distinct it is from the content set as a whole.

Exclusivity: This metric measures how exclusive the top terms for each topic are to that topic. The higher the value, the more likely that a topic's top terms do not appear as top terms for other topics.

Document Count: This measure measures the number of documents that make up a given topic. Smaller counts compared to the total number of documents may indicate an outlying topic whereas larger counts may be more representative of the Content Set as a whole.

These measures allow the user to explore how the topics created by the software relate to each other, and to the content set from which they are drawn. The user can look at raw counts, but also the possible kinds of interrelation the words have with each other and with the content set as a whole.

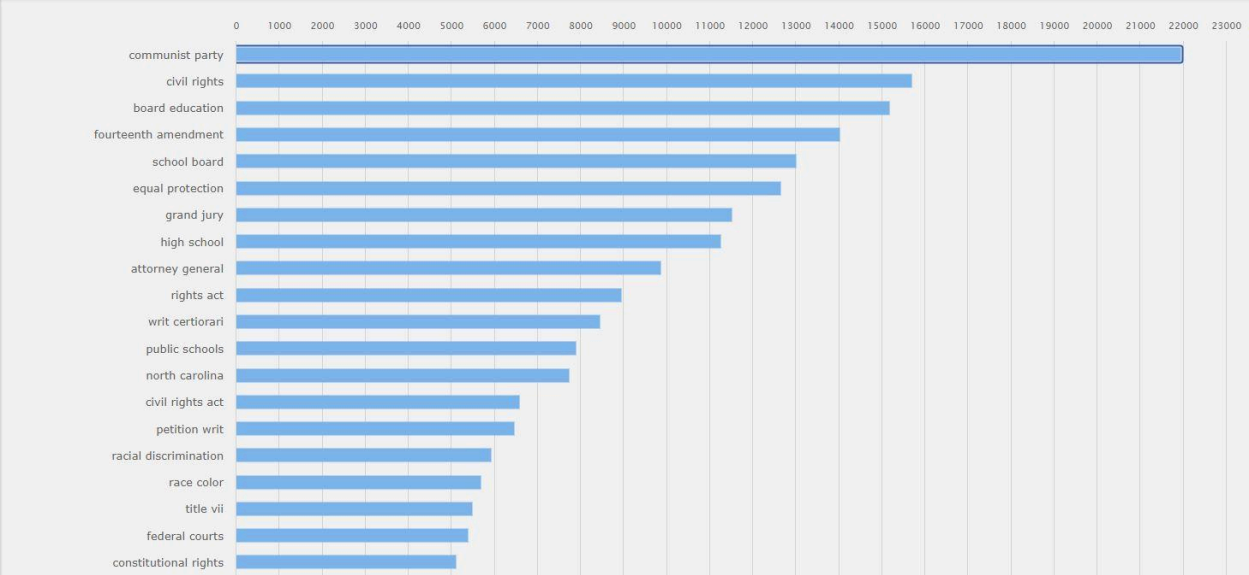
- Coherence overlaps conceptually with Ngrams. Can we compare the two tools in any meaningful way?
- Several measures point toward specificity or, to put it another way, the precision or clarity of a topic within the content set and documents: Document Entropy and Uniform Distance offer ways of examining how specific topics fit into the content set, as well as within documents.
- Another question concerns the uniqueness of a topic. Corpus Distance offers a means of thinking about the exceptionality of a topic.

Names for the topics created by the tool in this example are provided so they can be referenced later on. These names will appear in the Topic Proportion view, replacing “Topic [number]”, making it easier to navigate the results.

The Topic Modeling tool allows us to move through these measures and the topics themselves through the Topic Proportion view. We can select specific documents by title, and compare which Topics show up. This is particularly useful as a means of drilling into the content set itself.

In our results we see that the topics which have the greatest presence in the Proportion view in each case are those concerned with procedural or legal terms. This isn't surprising, but it also isn't particularly useful.

Ngrams



Configuration: Min 2 Max 5, Threshold 4

Cleaning: US Legal No Punctuation No Numbers

These are the top Ngrams, after rerunning the tool several times with different Cleaning Configurations. Strangely enough, the most frequent bigram (nGram with two tokens or words), is “communist party”. The next few, however, deal explicitly with themes one would readily expect: “civil rights”, “board education” (likely for “board of education”), “fourteenth amendment”, “school board”, “equal protection”, etc. All of these pertain to key legal battles surrounding the issues of race in the civil rights era of the 1960s and early 1970s.

It also suggests that the most frequent issues the U.S. Supreme Court handled in regards to the civil rights era had to do with schooling and access to it, and segregation. Importantly though, despite the fact that racial discrimination was the heart of the issue, it comes much lower on the list, following equal rights terminology. This suggests that while appellants were well aware of racial discrimination, they sought legal protection using equal rights arguments, rather than focusing on discrimination, as the basis for their legal filings. It is important to note that the original search did not specifically relate to “civil rights” or “discrimination”; these terms appeared as a matter of analysis.

Such outcomes confirm many of the things already known about the civil rights era and legal proceedings. The prominence of the “communist party” term, however, suggests a possible new direction for research.

1. Why does the term “communist party” appear so prominently in texts mentioning Black Americans in the U.S. Supreme Court Records between 1950 and 1980?
2. What connections can be made between the Cold War and race in the struggle for civil rights in mid-20th century America?

5. Interpret

Read more about ways you can expand this project with iteration, research questions, and analysis.

Iteration

Iteration for this project focused more on the Cleaning Configuration than working through the content set itself. This is common; the default Cleaning Configuration is only a base starting point. Each project will need to have at least one Cleaning Configuration of its own, if not more, which will require testing and rerunning until the results from the analysis are meaningful and uncluttered with “noisy” data.

Over the course of this project, it became increasingly clear that Ngrams and especially Topic Modeling required some tinkering to get the right Cleaning Configurations. This has as much to do with how the tools work as with problematic OCR. These two tools work with the actual words within a document, so if problematic OCR words have a significant enough presence, they will show up like any other words.

There were Ngrams that included correctly spelled words, but that had little usefulness to the research question. For example, the section numbers and abbreviations that are part of most legal documents showed up. Adding them to the stop word list removed them from being included in the Ngram analysis. Finding them all took several tries, but it was possible to get a fairly clean and meaningful output after a few iterations.

Topic Modeling, however, took much longer as the nature of the tool is to collate words that statistically appear often with one another. While problematic OCR was usually ignored by the Ngrams tool, it quickly became apparent with Topic Modeling because the tool returned results suggesting that problematic OCR words themselves constituted one or more topics. Rerunning the Topic Model analysis, allowed the outcomes to be used to revise the Cleaning Configuration; and then the analysis was run again.

Another problematic element with Topic Modeling is finding the right balance between getting results in line with the genre of the document, and eliminating words through the stop word list to ensure those results are meaningful. In the Topic Proportion view, the most prevalent topics were those concerned with procedural or legal terms, no matter how many attempts were made to clean out OCR or lesser words (such as prepositions, conjunctions, articles, etc.). But removing terms like “statute” or “federal” might alter themes important to the search even as they often appear as unified topics. This is to be expected, but it is another kind of “legitimate noise” —results that must be waded through in order to find the critical themes of the civil rights era.

Research Outcomes

It is clear from this brief test project that large-scale analysis of U.S. Supreme Court Records provides insight into broad themes and concerns that one would expect to find from civil rights-era legal proceedings that mention Negro, Black, or African Americans. At the same time the results also suggest new questions for consideration.

Original Questions

1. The U.S. Supreme Court Records and Briefs are clearly quite negative in tone. Even when divided by Document Type, the negative sentiment is striking. It fluctuates, and outside of Briefs and Petitions, tends toward a slightly more positive tone over the 1970s.

2. “Communist party”, “civil rights”, “board education”, “fourteenth amendment”, “school board”, “equal protection”, “grand jury”, “high school”, “attorney general”, “rights act”, “civil rights act”, “racial discrimination”, “race color”, “title vii”, “constitutional rights”

3. This requires a more precise Cleaning Configuration to remove problematic OCR. Some of the topics, however, do reflect civil rights themes—in particular, segregation and discrimination, as well as equal rights.

4. This is not easily discerned from our outcomes. Topic Modeling suggests some possibilities, but needs to be clearer. Ngrams also suggest dominant themes, but they are not as explicit as they could be.

5. In Ngrams the terms ``fourteenth amendment" and "title vii" both expressly forbid discrimination on the basis of race—the former as part of the Constitution, and the latter as a section in the 1964 Civil Rights Act.

6. It would seem so, especially when it comes to Ngrams. More exploration is required.

New Questions

1. Why does "communist party" appear in the Ngrams?

2. How does "Document Type" affect the outputs of the analysis tools?

Revising the Content Set

In light of the initial outcomes, it makes sense to revise or subdivide the original content set in order to pursue the new research questions. One of the possible means to do so is to create sub-content sets derived from the original content set using various metadata, such as the Document Types. For example, one of the Document Types in Gale Digital Scholar Lab is Legal Briefs and Petitions, which constitutes requests and filings made to the U.S. Supreme Court.

Perhaps the first question about "communist party" can be made clearer through the creation of sub-content sets consisting of different Document Types. Creating one content set that contains only briefs and petitions, and a second with the rest of the Document Types, might provide a different view of the research question by contextualizing all of the analyses that have been conducted with the issue of "genre". Genre allows for certain kinds of questions that can shape how one thinks about the outcomes of the tools:

1. What is a legal “brief” or a legal “petition”? Are they the same? Who writes them, and why?
2. What kinds of information does Briefs and Petitions contain? As a genre, does it have particular characteristics?
3. How might knowledge of a Document Type—a genre, or a form of writing—shape our research inquiries? What can we learn by understanding the form of a document, and what it might contain textually?

Sub-Content Set: Briefs and Petitions

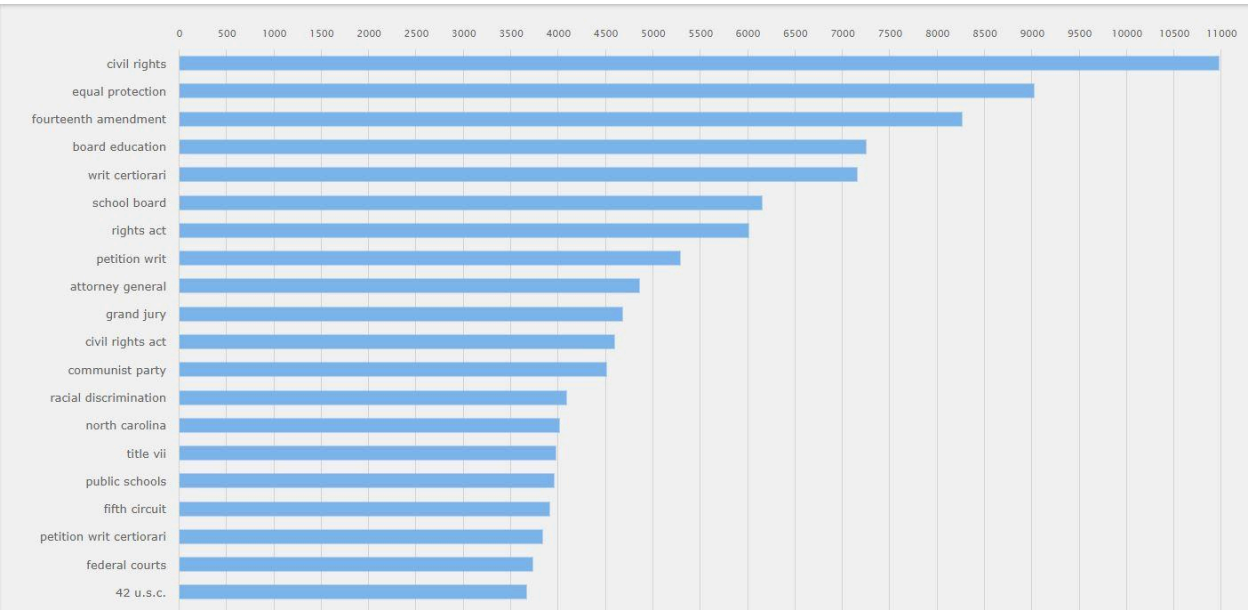
Topic Modeling

- jury state trial grand u.s county petitioner negroes death juror
- discrimination u.s title minority racial employment vii black cir program
- school negro race white schools state education public law equal
- school schools board plan education racial black students desegregation county
- city housing property public private state u.s park racial discrimination

This sub-content set seems to be more precisely concerned with the issues of the civil rights era than the main set.

Ngrams

We can see that the Communist Party no longer appears as the top nGram, and that the remainder of the top 10 or so Ngrams are focused solely on civil rights-era issues. The Communist Party does appear as the 12th highest nGram, in a dramatic drop from first place. Clearly it retains some relevance, but is not as prominent as in the main content set that included memoranda.



Sub-Content Set: Statements, Memoranda, etc.

Here is what the other Document Types look like with the same analyses.

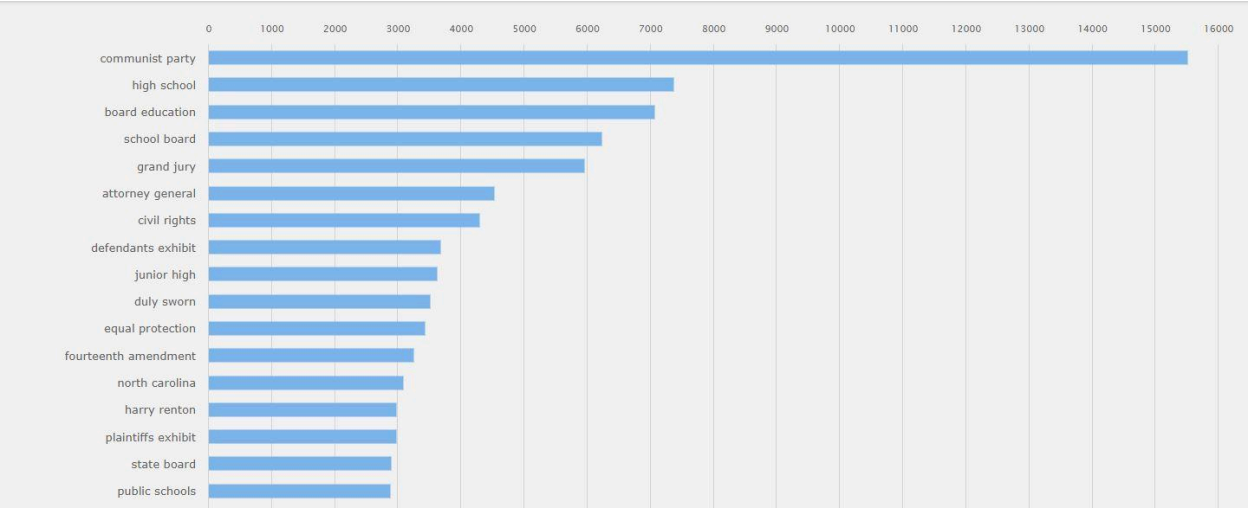
Topic Modeling

- school schools board plan education students high black white children
- party communist national exhibit mcgohey war sacher political objection class
- party communist bridges testimony harry committee union answer member donohue
- state defendants fol defendant plaintiffs alabama county plaintiff motion complaint

There are clearly some topics related to the civil rights era, but also those focused on the Communist Party topic, as well as other themes. This suggests a greater diversity of content in this sub-content set than the other.

Ngrams

Notice, “communist party” still appears dramatically in first place. It ranked 12th in Briefs and Petitions. Still many of the same Ngrams appear, suggesting that civil rights-era concerns remain dominant throughout the documents, regardless of their type.



Reflections on Method

This project provides insight on how to build content sets, and what iteration means when revising and cleaning them prior to analysis. The comparisons between the two sub-content sets—Briefs and Petitions and Statements, Memoranda, etc.—allows us to clearly see how the parameters or fields that are used to build a content set can affect or shape the kinds of results obtained from the analysis tools. Apparently, “communist party” was a matter of discussion in memoranda and not briefs or petitions to the U.S. Supreme Court between 1950 and 1980 in cases involving mentions of Negro, Black, or African Americans.

Thinking Critically About Research

Content Set Building

Building this method was fairly straightforward, as our research question was focused precisely on one Archive source, and had a clear and defined publication date window. The purpose was to use the DSL to explore and discover possible new research questions from

a large set of documents, rather than pursue precise questions around a specific author, genre, or perhaps another variable. That said, it's clear that as we worked with the content set, new questions emerged that required refining the Content Set. We built two new versions of the original Content Set, dividing it up using the Document Type values. We could've done this at the outset as well.

Another possible avenue for more precise Content Sets—allowing us to explore more precise research questions—would be to build additional Content Sets based on authorship, or by case. Such precision requires increasing in-depth knowledge and expertise with the subject matter itself. The DSL can build such Content Sets, but you, the researcher, need to have sufficient experience with the subject matter in order to define the parameters of your research question and how it relates to the Content Set you might build. Having a list of cases involving civil rights would offer a rich picture of views of African Americans during this era. However, it would also mean determining what cases involved civil rights issues. Were they just those grappling with civil rights statutes? Or can we learn anything about how the legal system treated or viewed African Americans in cases that didn't address civil rights statutes directly? These are two different kinds of questions and require different content sets.

Understanding Outcomes

What do these outcomes tell us about our research interests? We can understand outcomes in a variety of ways: how they answer the questions we originally posed and how they suggest new questions and new avenues for research.

It's clear that our outcomes corroborate many of the things we already know about the civil rights era:

- It coincided with the height of the Cold War and anxiety surrounding Communism.
- Cases focused on discrimination on the one hand and equality on the other.
- School segregation was a primary concern of cases shaping the experience of the legal system among Black, African, or Negro Americans.

What is unclear from our outcomes, however, is whether the sentiment of the documents in our Content Set is a matter of their genre and context (e.g., legal documents are always

"negative," perhaps because they involve contestation or argument) or actually related to racism and discrimination. It's likely a combination, but there's no assured method of teasing these two apart.

We can also examine the outcomes as a way of understanding and reflecting on our methodology. What kinds of fields or content-set-building techniques might we use to create more focused or precise collections of documents that could better answer our questions? How could we manage whether a document actually discusses what its metadata says it does: in other words, do the contents match what cataloguers or others, even perhaps the original authors, tell us? Closer reading of the documents before adding them to a document set will help us discern whether they should be included in a content set or not.

Revising Questions

Once we saw our initial outcomes in this project using the main Content Set, it was clear we could revise our research questions somewhat, both making them more precise, but also possibly exploring a new question around the presence of something unexpected—the "Communist Party" biGram. Where did this come from? Is there any way to isolate it to discern why it might appear so prominently in the Content Set? Why does it appear in documents that mention Black, African, or Negro Americans in the mid-20th century? Is there an overlap between the Cold War and fears of communism and the racial tensions of the civil rights era?

As discussed in the guide, it's not only normal to revise your research questions after running analysis tools on a Content Set, it is an integral part of the research process. Often, analysis will turn up new questions that could lay beyond the scope of your current project. This is how researchers develop new projects and lines of scholarship—by following clues and new questions that come up while pursuing other research.

Limitations of the Project

As useful as these results might be, there are limitations to what kinds of cleaning and analysis can be done with the *Gale Digital Scholar Lab*.

- Currently, there is no method within the Lab to compare all words against an English dictionary in order to identify problematic optical character recognition (OCR). Iterating through cleaning configurations using Topic Modeling is a sound method for finding problematic words, but it is a time-consuming process. Replacements can be made easily, allowing problematic OCR to be fixed, but there's no method of finding all instances of misspelled words.
- This project didn't build content sets using actual cases, nor were they built following a close reading of the documents included in the sets. A more precise content set could be built by determining, following examination of each document, whether or not it was appropriate to include in a Content Set focused on the specific parameters of the project.
- We haven't fully considered the difference between raw numbers or "counts" and statistical measures as distinct ways of thinking about significance. Although the Topic Modeling output allows us to examine counts, the Latent Dirichlet Allocation method used by MALLET is a kind of prediction of the likelihood of words appearing with each other. It's suggestive, in other words, of something significant. The nGrams, in contrast, are raw counts across the content set. Having more or longer documents—i.e., documents with more words—would increase those counts. Numerical presence, however, doesn't always translate into intellectual significance or meaningfulness.

Beyond the *Lab*

Presentations

All of the tool outputs can be downloaded as images to use in PowerPoints or embedded in web pages or other ways to present your work.

New Visualizations

It's also possible to download the data that power the visualizations as comma delimited (CSV) or JavaScript Object Notation (JSON) files, allowing you to create and format your own visualizations. If you have the skills, it's possible to collate or create new visualizations that may combine outputs from similar visualizations into one, allowing you to compare and

contrast in new ways that the Lab tool doesn't. The Topic Modeling tool downloads are especially rich with possibilities for new visualization. The Topic view download is large and contains results for each document and measure for the tool—much more data than the Topic Model visualizations can currently display. If you're a programmer, this is the ideal place to start to explore the data created by the DSL using other tools and visualization designs.

Refining the Content Sets

Understanding the limitations of the Lab allows us to consider what can be done to both build content sets and to use the results produced by its tools. Building content sets using U.S. Supreme Court case records and documents would provide a completely different method of considering the themes of civil rights and racial discrimination. At the same time, it would also isolate analysis to documents that are explicitly tied to such legal questions. We know, however, that these themes can't be, and were not isolated to explicit cases.

Similar Projects

Appreciating how we can structure a project or line of inquiry can be shaped as much by the tools and content we have as by modeling similar projects that explore similar kinds of documents. The Old Bailey Online project (oldbaileyonline.org/) involved large-scale analysis of court proceedings from the main municipal court in the city of London. Although its content has been encoded and cleaned and its platform doesn't contain the same kinds of tools, as a project it offers a way of considering how examination of the U.S. Supreme Court records might be modeled. It provides approaches, questions, and methods that could be applied and tested using our current project's contents and tools.