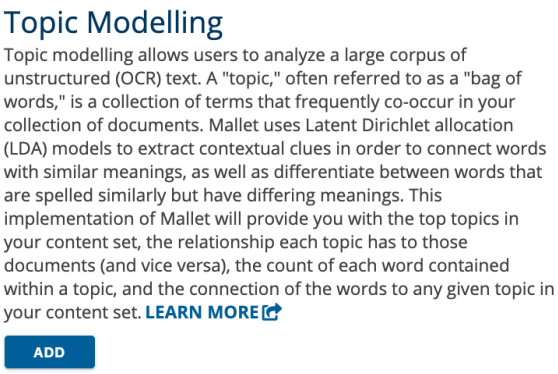
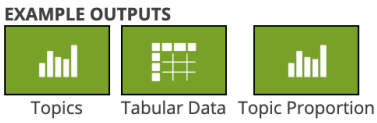
Digital Humanities in Practice

WEEK 8b Topic Modeling

Tool Overview





When you run topic modeling, the algorithm will sift through your content set and group together themes or topics it considers related in some way. Posner (below) notes that: 'it should be abundantly clear that no part of this process is “scientific”; it’s just one way of getting your head around a large body of text. So there’s no right or wrong topic name, just schemas that do and don’t help you find interesting features of the text you’re looking at.'

Digital Scholar Lab Implementation

The topic modeling tool in the DSL is built on MALLET, which initially was accessible only via the command line but now also has a GUI implementation. It's fully integrated in the DSL making it pretty streamlined to use.

Reading

Megan R. Brett, Topic Modeling: A Basic Introduction, *JDH* 2:1, 2012 http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/

Ted Underwood, 'Topic modeling made just simple enough', *The Stone and the Shell*, 2012. https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Ted Underwood, 'Visualizing Topic Models' *The Stone and the Shell*, 2012. https://tedunderwood.com/2012/11/11/visualizing-topic-models/

Shawn Graham, Scott Weingart, and Ian Milligan, Getting Started with Topic Modeling and MALLET https://programminghistorian.org/en/lessons/topic-modeling-and-mallet

Miriam Posner, 'Very Basic Strategies for Interpreting Results from the Topic Modeling Tool', http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/

Example Projects using Topic Modeling

Mining the Dispatch, https://dsl.richmond.edu/dispatch/pages/home

Cameron Blevins, 'Topic Modeling Martha Ballard's Diary', 2010 http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/

Quintus Van Galen & Bob Nicholson, 'In Search of America', *Digital Journalism*, 2018. DOI: 10.1080/21670811.2018.1512879

Configuration Options in the DSL

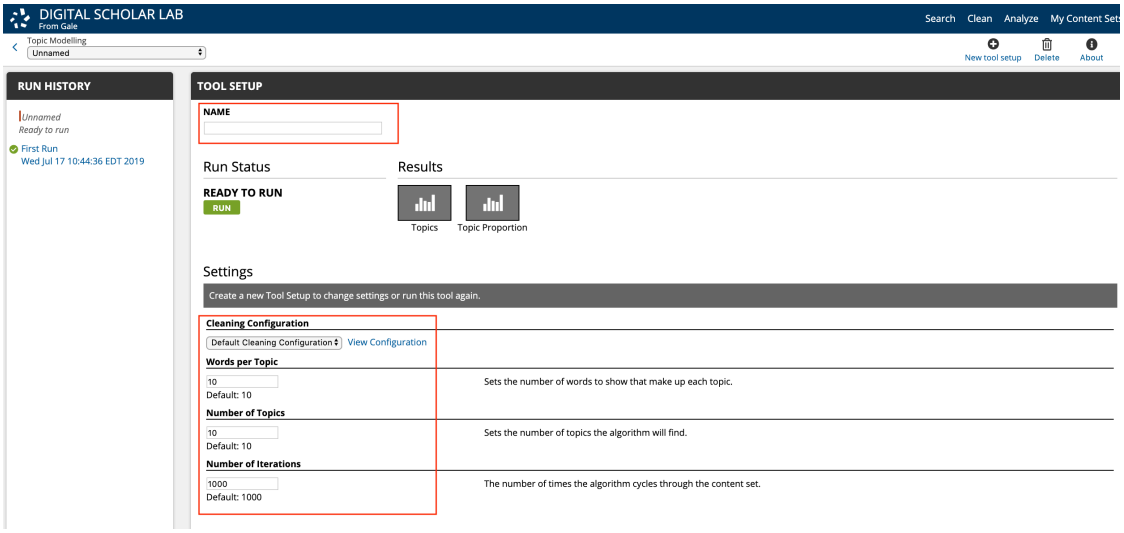
1. Name the run of your tool.

2. Apply cleaning configuration, as appropriate. Things to note: MALLET - the software powering the Topic Modeling Tool - is case sensitive. If you decide to make everything lower case, it won't distinguish between Smith (perhaps someone's last name), and smith (an occupation, like a blacksmith). MALLET also handles possessive apostrophes in a slightly awkward manner, turning them into their own words - you can add 's to the Stop Word list to prevent this from happening.

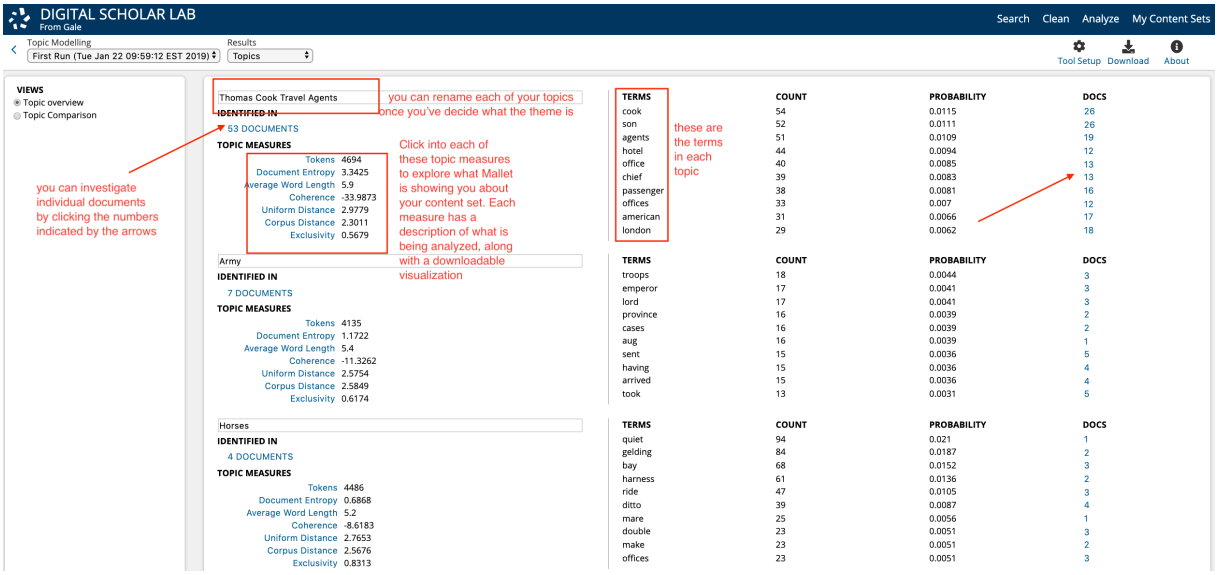
3. Choose the number of words you'd like to appear in each topic. 10 is a reasonable default.

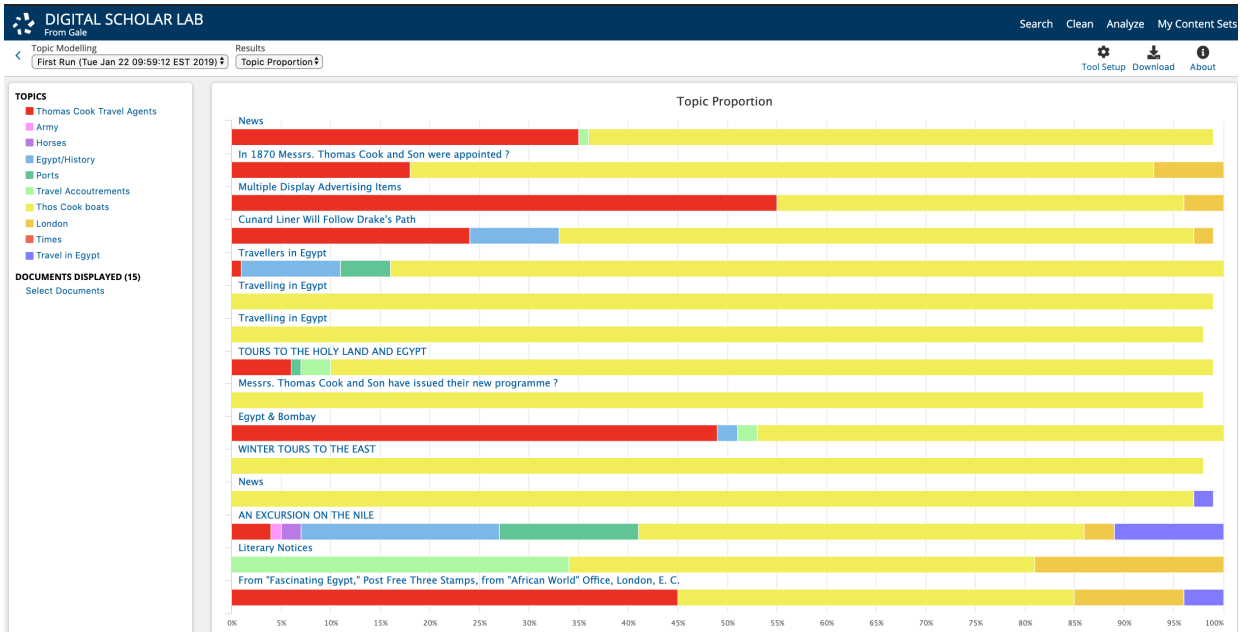
4. Choose the number of topics - you could start with 10, then increase the number to see what else you can discover.

5. The number of times the algorithm will iterate through the content set before returning a result is set at 1000 as a default. It's fine to leave this as it is.



Here's an overview of what is shown on the 'Topics' page. You can also download a CSV of this data in it raw form, in the same format that a user of MALLET might expect to see.

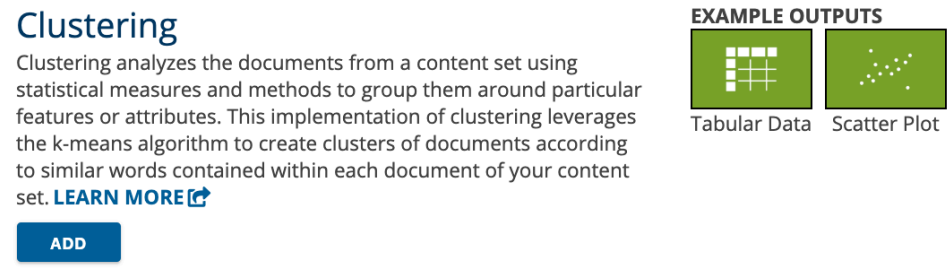




This is the topic proportion by document, displaying the tabular data in a graphic format. The visualization is interactive, so click through to explore each aspect of the page. You can also download this visualization.

Clustering

Tool Overview



K-means clustering is arguably the most challenging analysis tool to understand in the DSL.

- **Q:** What variables are being plotted on the scatter plot that is produced by the tool? The axes don't have explicit labels, and we haven't been able to find information on what features of the text are actually being plotted here.
- **A:** The x/y axis do not have labels because the scatter plot represents flattened multi-dimensional vectors. The points within the scatter plot represent documents within a Content Set. Their similarity to one another is based on their distance within these multi-dimensional vectors which are then flattened in 2-D space.

Digital Scholar Lab Implementation

The clustering tool in the DSL is built leveraging open source software, in this case scikit-learn's K-means Clustering algorithm, which is described in detail here.

Reading/Viewing

Ben Schmidt, 'Machine Learning at Sea', *Sapping Attention*, 2012 https://sappingattention.blogspot.com/2012/11/machine-learning-on-high-seas.html

Descriptive lectures/videos:

- https://www.coursera.org/lecture/machine-learning/k-means-algorithm-93VPG Watch the first 3mins 41 seconds of the Coursera video, unless you want to dig deep into the K-Means algorithm in which case keep watching!
- Lexos overview of K-means clustering: https://youtu.be/B7_cJBeofn4. There is a good explanation of what K-means is, starting around 30 seconds. Keep watching to see the visualization in 3D, vs. the flattened 2D representation currently available in the DSL. This may give you a better sense of how to interpret the clustering visualization.

Example projects using Clustering

See Ben Schmidt's Whaling project, above.

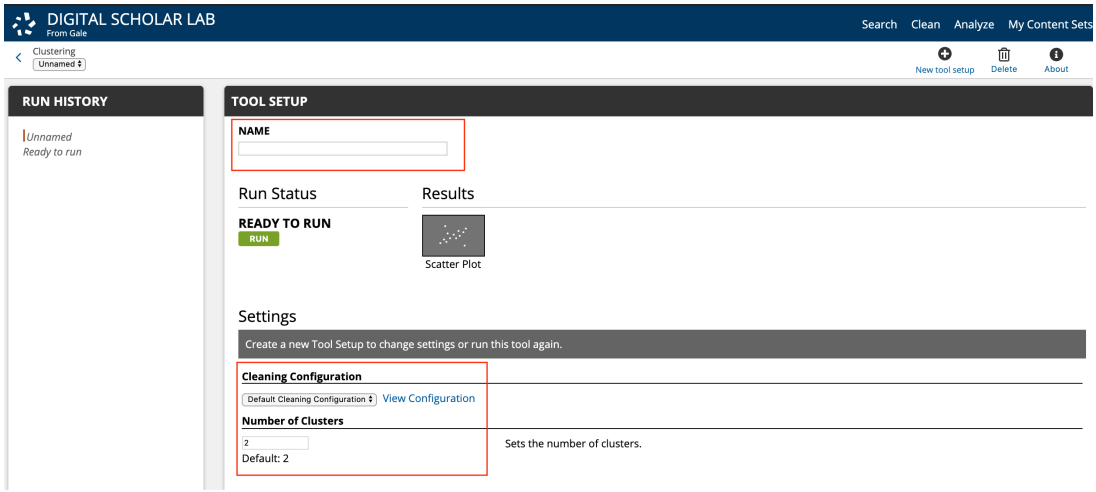
Configuration Options in the DSL

You have two configuration options for clustering.

1. Choose the cleaning configuration you want to apply.

2. Choose how many clusters you want the algorithm to group your content in.

Once you have done this, name your tool setup and click 'Run'.



Sample Project #2

Black America & The Law in the mid-20th Century

Synopsis

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 American men and women proceeded through the legal system. How that legal system, and those involved in it responded, implicitly or explicitly, to the pressures of the era in its trials and cases, can be seen by examining the US Supreme Court Records and Briefs.

Core Research Question:

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

More Precise Questions:

1. Are the Supreme Court Records and Briefs mentioning negro, black, or african americans more positive or negative in sentiment?

    1. Does this change over the course of the Civil Rights Era?

2. What are the most common phrases or collocates used in documents mentioning negro, black, or african americans?

3. What topics appear most frequently in documents mentioning negro, black, or african americans?

    1. Do these reflect themes that dominated Civil Rights Era conflicts?

4. What are the main concerns of legal records mentioning negro, black, or african americans during this period?

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

6. Are there any differences between Document Types in the Archive (US Supreme Court Records and Briefs)?

Thinking about Methodology & Specific Tools

- Topic Modeling - we can use this tool to see if there are any themes or topics which cut across a collection of texts

- nGram - we can use this tool to track different kinds of phrases or terms which might occur together, and the number of times a phrase appears. In some cases names with several words - like United States or North Carolina - might appear

- Sentiment Analysis - we can use this tool to examine whether the contents of the documents were overall positive or negative according to the AFINN dictionary

Building the Content Set

Searching

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

Limits & Parameters

Content Type ("Monograph")

Archive (U.S. Supreme Court Records and Briefs, 1832-1978)

Publication Date (1950 - 1980)

Keywords (in individual rows)

Entire Document: black, black american, black man, black woman, negro, african american, african americans, black americans

Statistics & Info

Content Set Name: 1950-1980 Black America & The Law - no subject terms

Content Set ID: 1579633475592

Number of Documents: 4289

Specific Tools

None of the tools required specific content sets.

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.

Cleaning the Content Set

It took several attempts, or iterations [Link on Iteration], to get the cleaning right for each of these analyses. In the end, it requires several different cleaning configurations, as there were different stop words needed for different tools, as well as different approaches to punctuation. The main stop word list required adding single letters (for each letter, in case they appeared as abbreviations), as well as additional stop words. Also, some replacements were obvious from the test cleaning configurations.

Specific research questions:

1. No additional cleaning configurations required for Sentiment Analysis.
2. This required considerable iteration, in order to remove abbreviations and single letters, and additional stop words.
3. This required additional stop words that were unique to Topic Modeling, and were not the same as those used in nGrams - ‘state’ ‘court’, for instance needed to be retained for nGrams (for ‘United States’), but removed for Topic Modeling which treats individual tokens (ie. ‘United States’ can never occur in a Topic Model because it includes two words. The software doesn’t operate on phrases).

Running Tools

Selecting particular views for each tool was extremely straightforward. We selected an approach that reflected the size of our content set: we looked for more things and 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’, meaning we get a finer grained picture of what the statistical analysis could suggest. In very similar documents, like court records, the chance is that there will be similar phases as questions and answers are posed, and rulings and arguments recorded. Selecting more words and more topics is a good way of sifting 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, we took a similar approach to thinking about potential ‘noise’ - we want to see what turns up. But the highest count in a result doesn’t always translate into the most meaningful or interesting. There’s a balance between number and noise.

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, in reflection of the size of the content set. We opted for 15 word topics, and 20 topics.

Sentiment Analysis

This tool has no settings other than selection of the cleaning configuration.

nGrams

Like Topic Modeling, it seemed worthwhile to go beyond the default settings given the size of our content set. We raised the threshold for the number of times an nGram had to appear to be considered useful, and set it at 4. Equally, we wanted to find collocates rather than just single words, so we set the minimum nGram size to 2 (biGram), and the maximum size to 5. These settings translate into a search for "nGrams 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 set up in scholarly inquiry and how we pursue research questions. An essential element of this in a computational text analysis environment like the DSL is understanding that raw counts or 'more hits' doesn't always mean something important - it could be noise, meaning its simply there because the content set wasn't cleaned enough, or we didn't use the right stop words, or perhaps it's something expected and known. In the end, understanding results really requires understanding what we've asked of the tools in our configurations and settings, and how the results relate to the variables we've selected, and the algorithms which power the analysis.

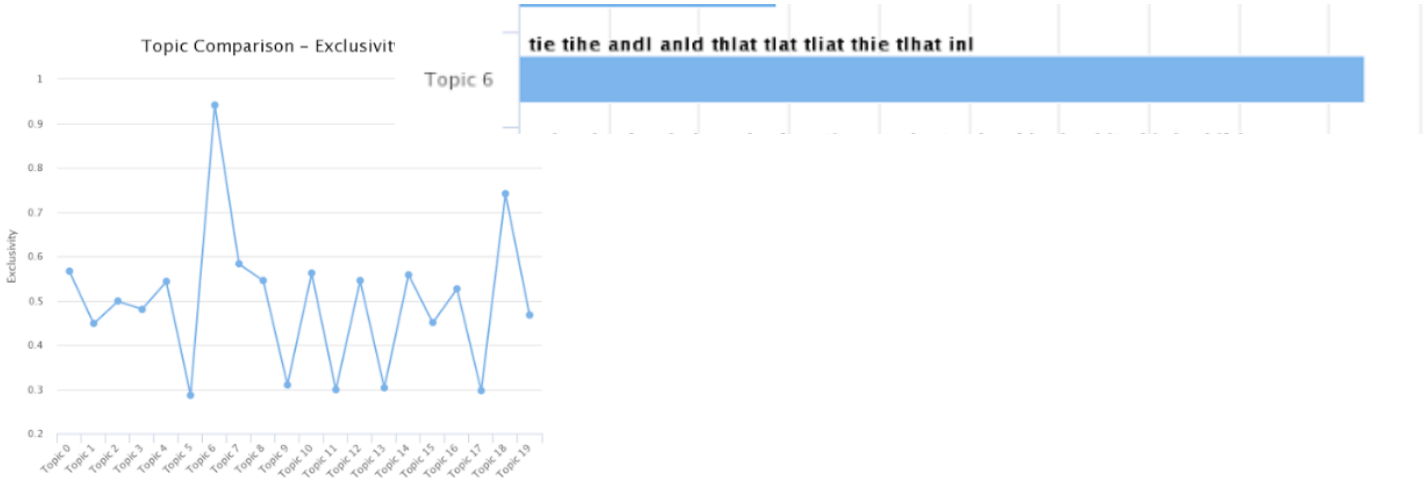
Topic Modeling

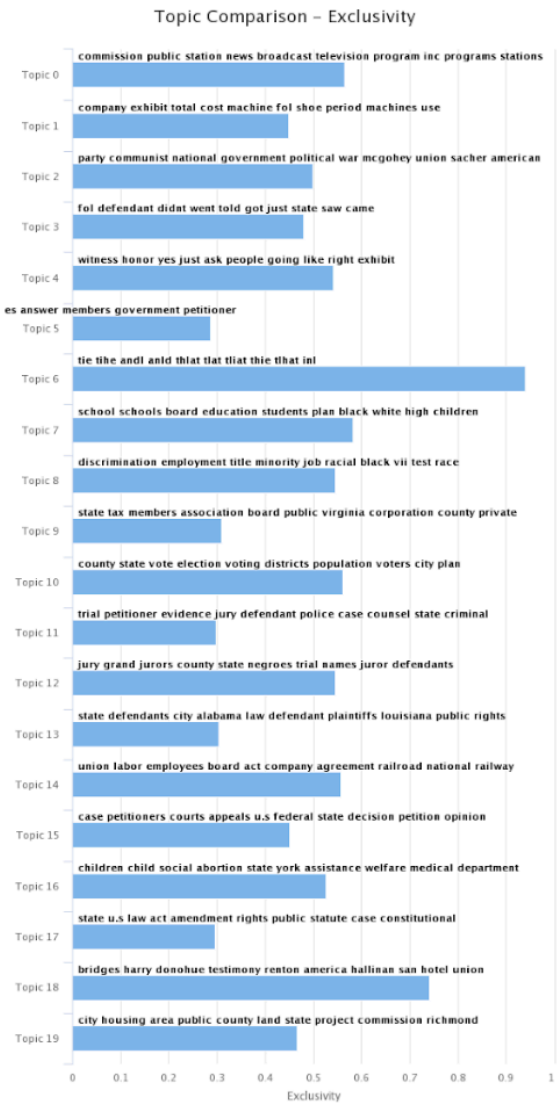
Refining the stop word list for topic modeling took some time, as the normal stop word list didn't include OCR error words. This is an example of the kinds of problems that can occur - notice Topic 6:

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 less than useful words.

What we can see from the Topic Modeling are a series of possible themes within the US 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

How we decide what is the most meaningful or significant measure in these results can be tricky - it depends on what our question might be, of course. The results from Topic Modeling could be meaningful simply by being unexpected or new, suggesting something that we didn't already know, or perhaps were unaware of. At the same time, they might confirm something we already know, and can act as a touchstone to confirm that we're on the right course with reading or analysis, or both. We see in the topics returned above, that 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 looking at the Topic results section in the tool view, and by selecting 'Topic Comparison'. Here we see a list of measures describing how the topics relate to the content set, and the analysis.

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. Because longer words are assumed to be more meaningful, 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.

These measures allow us to explore how the topics created by the software relate to each other, and to the content set from which they're drawn. We 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.

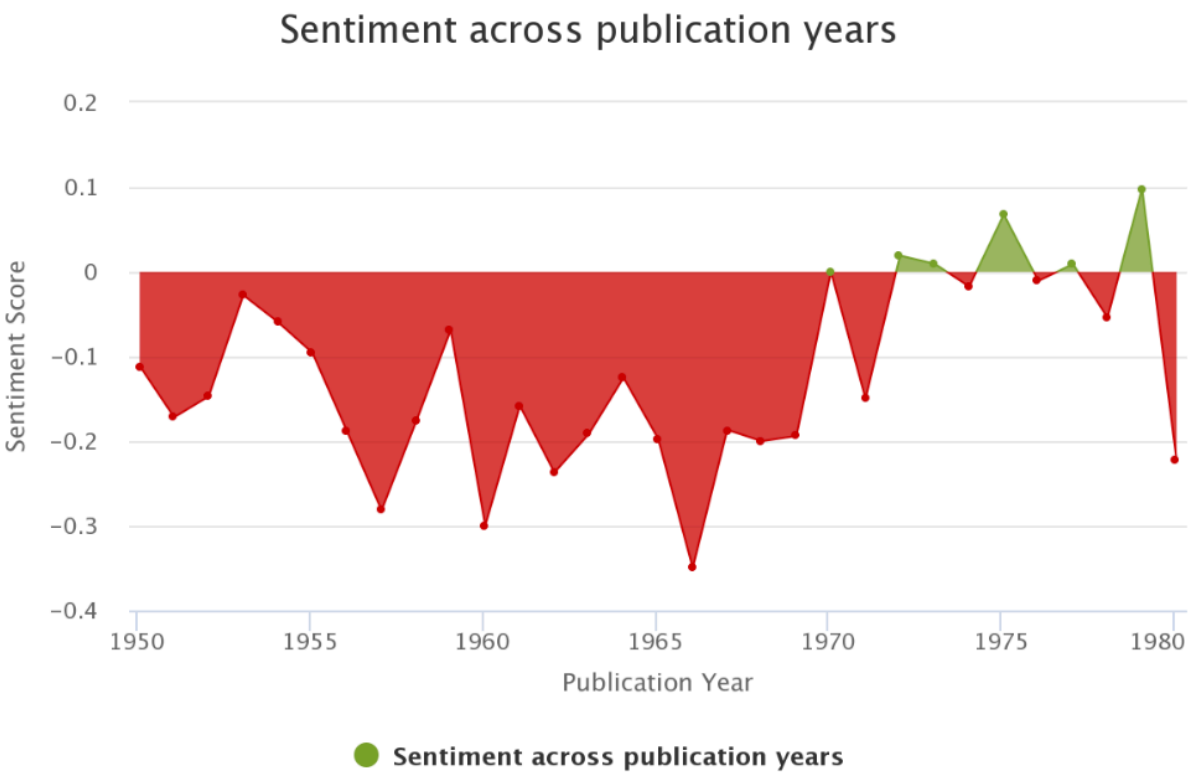
- Coherence overlaps conceptually with nGrams - can we compare the two tools in any meaningful way?
- Several measures point towards specificity, or to put it another way, the precision or clarity of a topic within the content set and documents: Document Entropy and Uniform Distance offer ways of examining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

We can provide names for the topics created by the tool, so they can be referenced later on. These names will appear in the Topic Proportion view, replacing "Topic [number]", making it easier to navigate what the results are.

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.

Sentiment Analysis



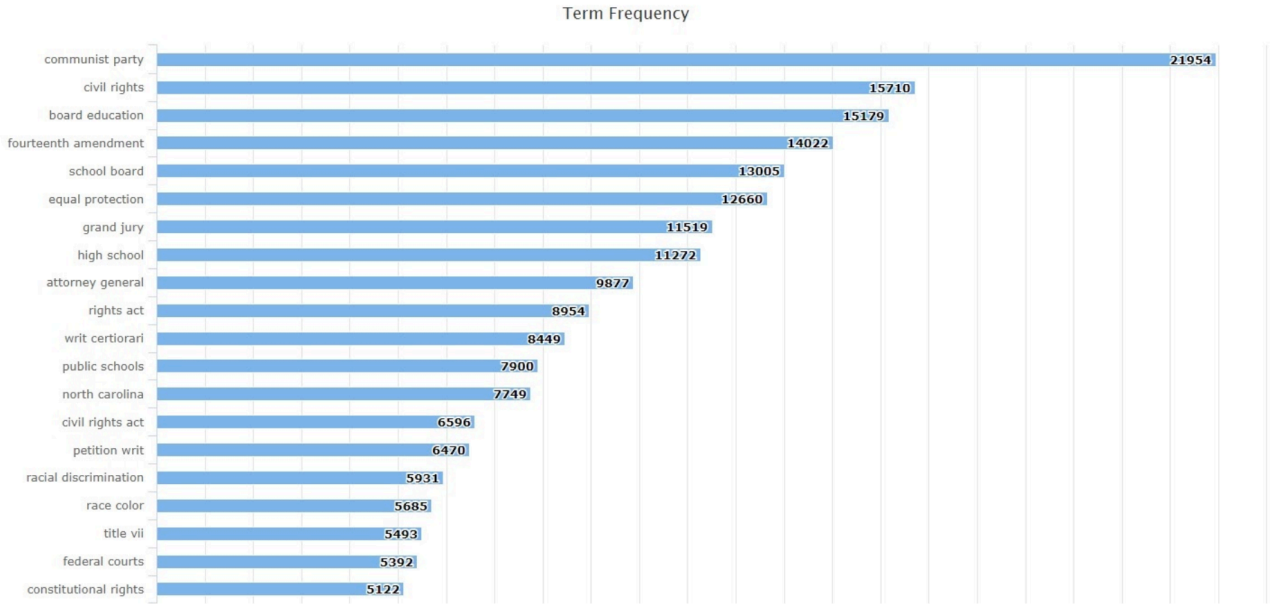
This seems unsurprising, given the fraught nature of the Civil Right Era, with its protests, violence, and focus on systemic racial discrimination. It raises several NEW research questions, however:

1. Can we associate or link the years with the lowest sentiment scores with particular events or moments in the Civil Rights Era?
2. Why might there be years with positive scores throughout the 1970s?
3. To what degree can we attribute these scores of sentiment to the issues that shaped the Civil Rights Era versus the general nature of legal cases as contestations or conflicts, or matters of friction? IE, will most of the US Supreme Court Records have a negative sentiment because they deal with legal matters, or is there something specific about the records dealing with Black America that makes them different or stand out?

nGrams

Configuration: Min 2 Max 5, Threshold 5

Cleaning: US Legal No Punctuation No Numbers



Here we have 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 we’d 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 US 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 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. Remember, we didn’t search for anything related to ‘Civil Rights’ or ‘Discrimination’ - these appeared as a matter of analysis.

Such outcomes confirm many of the things we know about the Civil Rights Era, and legal proceedings. But the Communist Party prominence suggests something that could be followed up - at least two NEW Research Questions:

1. Why does the 'Communist Party' appear so prominently in texts mentioning Black Americans in the US Supreme Court Records between 1950 and 1980?
2. What connections were perceived between the fear and conflict of the Cold War, and race, in the struggle for Civil Rights in mid-20th century America?

Research Outcomes

It's clear from this brief test project that large scale analysis of US Supreme Court Records provides insight into broad themes and concerns that we expect to find from Civil Rights era legal proceedings which mention Black, African, or Negro Americans. At the same time the results also suggest new questions we could consider.

Original Questions

1. The US 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 towards 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 we do see, 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 suggests dominant themes, but it is not as explicit as they could be.
5. In the nGrams we see the presence of the Fourteenth Amendment and Title VII - both of which 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 our initial outcomes, we can now think about revising or subdividing our original content set in order, to pursue our new research questions. One of the possible means of getting at the concerns of those within the legal establishment and the Civil Rights Era is to create sub-content sets derived from the original content set using various metadata. Legal Briefs and Petitions were requests and filings made to the US Supreme Court - they are specific genres, noted as Document Types in the DSL.

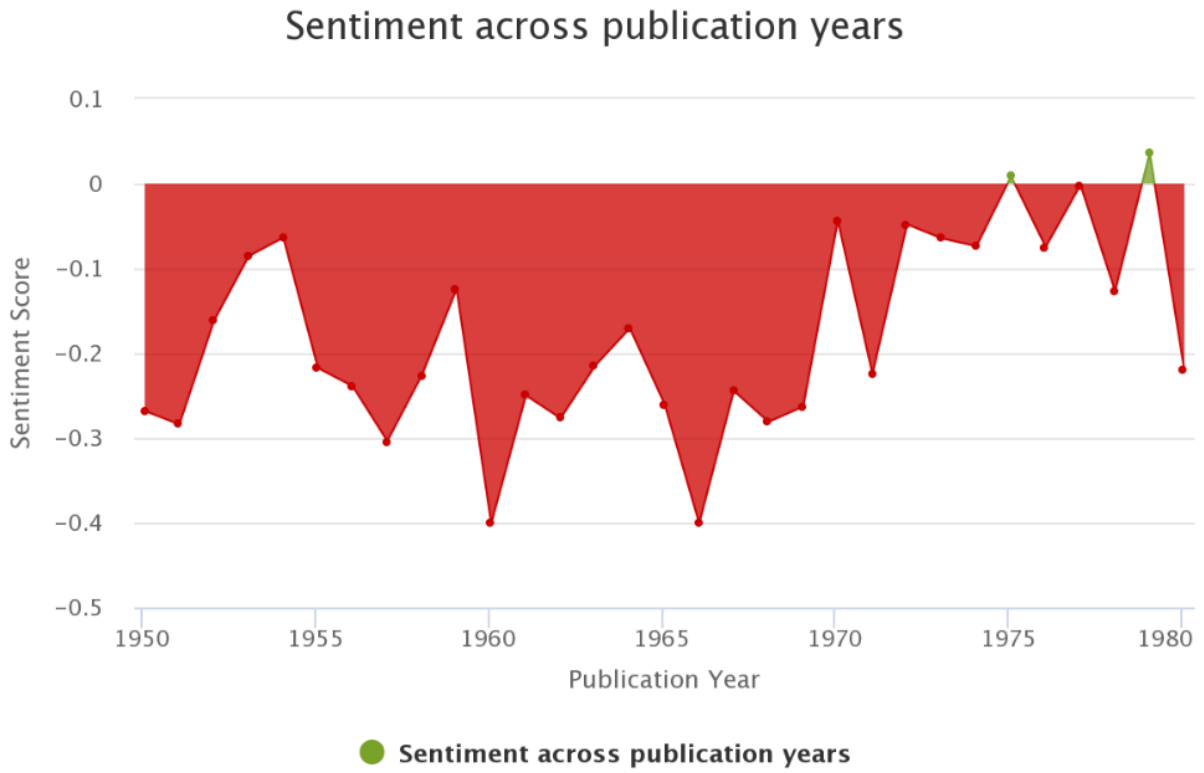
Perhaps the first question about ‘Communist Party’ will be clearer when we create sub-content sets consisting of different Document Types. Creating content sets which contains only briefs and petitions, on the one hand, and another with the rest of the Document Types, might allow us to gain a different view of our research question by contextualizing all of the analyses we’ve conducted with the issue of ‘genre’. Genre brings certain kinds of questions that can shape how we think 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 do Briefs and Petitions contain? As a genre, do they 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?

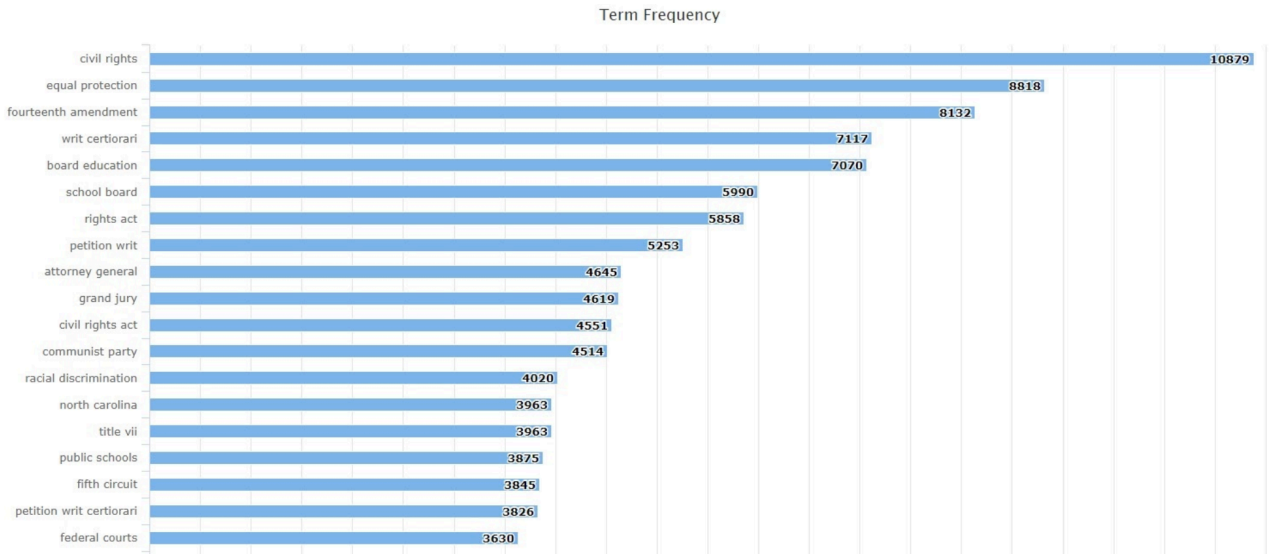
Sub-Content Set - Briefs and Petitions

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.

Sentiment Analysis

It would seem that Briefs and Petitions remain almost consistently negative when it comes to sentiment analysis, except for some small positive years in the mid to late 1970s. This could suggest that as genres, Briefs and Petitions are normally negative in tone or sentiment, and that perhaps the negativity isn't associated with Civil Rights Era issues. Or, it could suggest entirely the opposite. A good check would be running the same Sentiment Analysis on a new content set containing documents that aren't Briefs or Petitions in the main content 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 which included memoranda.

Sub-Content Set - Statements, Memoranda, etc.

Let's see what the other Document Types look like with the same analyses.

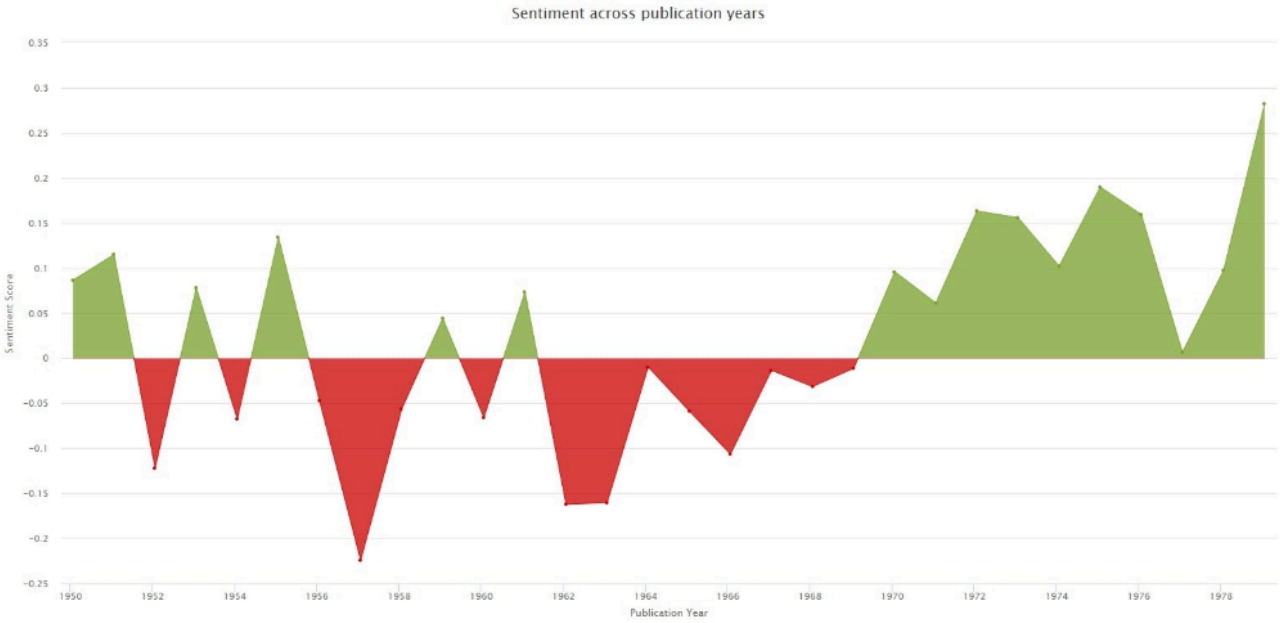
Topic Modeling

- 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.

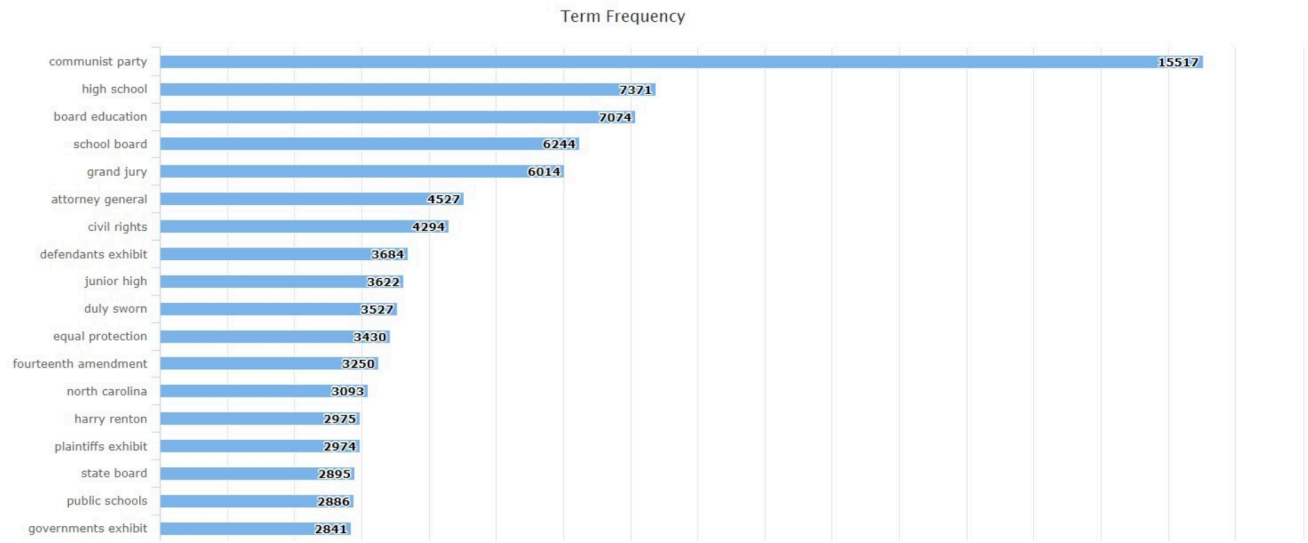
Sentiment Analysis

Sentiment Analysis is dramatically different. Clearly Statements and Memoranda had very different tones when it came to race earlier and after the 1960s. But the core Civil Rights Era of the 1960s was still overwhelmingly negative in sentiment.



nGrams

Notice, 'Communist Party' still appears dramatically in first place. It ranked 12th in Briefs & 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 allows us to reflect on how we build content sets, and what iteration means when revising and cleaning them prior to analysis. The comparisons between our two sub-content sets - Briefs & Petitions and Statements, Memoranda, etc. - allows us to clearly see how the parameters or fields we use to build a content set can affect or shape the kinds of results we obtain from the analysis tools. Apparently, 'Communist Party' was a matter of discussion in memoranda and not briefs or petitions to the US Supreme Court between 1950-1980 in cases involving mention of Black, African, or Negro Americans.

Content Set Building

Building this method was fairly straightforward as our research question was focused precisely on one Archive source, and 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 which 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 have 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.

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. And these require testing and rerunning until the results you receive from the Analysis are

meaningful and uncluttered with ‘noisy’ data. Over the course of this project, it became increasingly clear that while the Content Set was usable for Sentiment Analysis without much tweaking of the Cleaning Configuration, nGrams and especially Topic Modeling required some tinkering to get the right Cleaning Configurations. This has much to do with how the tools work as with problematic OCR. Sentiment Analysis matches words it already knows, meaning if something is misspelled, it is ignored. Cleaning, in this context, only adds words into the mix; problematic OCR isn’t analyzed. The other two tools work with the actual words within a document, and so if problematic OCR words have a significant enough presence, they’ll show up like any other words. When it came to nGrams, there were nGrams which included correctly spelled words, but which had little usefulness to our research question: often legal documents contain many sections, and those section numbers showed up, along with abbreviations as they’re in every document. 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 became quickly apparent with Topic Modeling because the tool returned results suggesting that problematic OCR words were 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 results we expect due to the genre of the document, and eliminating words through the stop word list so we might find things that could be meaningful. In the Topic Proportion view, the most prevalent topics were those concerned with procedural or legal terms, no matter how much we cleaned out OCR or lesser words (like prepositions, conjunctions, articles, etc.). But removing terms like ‘statute’ or ‘federal’ might alter the themes we’re interested in investigating at the same time they often appear as unified topics. This is to be expected, but it’s another kind of ‘legitimate noise’ - results we need to wade through in order to find the themes of the Civil Rights era we’re looking for.

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 not the sentiment of the documents in our Content Set is a matter of their genre and context (ie 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 prior to 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 which 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's an integral part of the research process. Often, analysis will turn up new questions which 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

As useful as these results might be, there are limitations to what kinds of cleaning and analysis that can be done with the DSL.

- Currently, there is no method within the DSL to compare all words against an English dictionary in order to identify problematic OCR. Iterating through Cleaning Configurations using Topic Modeling is a sound method for finding problematic words, 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 did not 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 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 documents, or longer documents - ie 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 webpages or other ways to present your work.

New Visualizations

It's also possible to download the data which 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 DSL tool does not. 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 DSL allows us to consider what can be done to both to build content sets, and to use the results produced by its tools. Building content sets using US 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 cannot 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 (https://www.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 US Supreme Court records might be modeled. It provides approaches, questions, and methods which could be applied and tested using our current project's contents and tools.