STEP 1: Define Project Scope and Objectives

Objectives

The primary goals of this project include understanding banning patterns over time, building a digital archive and searchable database of banned works, and creating visualizations such as interactive maps to convey findings.

Time Periods and Focus

The project spans the 19th century to 2024, focusing on significant periods of censorship:

- 19th century: Abolitionist literature, such as *Uncle Tom's Cabin* by Harriet Beecher Stowe.
- Early 20th century: The Comstock Act and the "Banned in Boston" era during the 1920s.
- 1950–2024: Civil Rights era, LGBTQIA+, and BIPOC authors.

Corpus Description

The dataset comprises a curated selection of books that have faced outright bans at different times in the history of the United States. Spanning the 19th century to the present, this corpus reflects the evolving nature of prohibition and offers a comprehensive look at how literature has been suppressed for a variety of cultural, political, and social reasons. It includes works that address both subject matter and content, representing a wide range of literary forms, genres, and historical significance.

This collection highlights books that challenged prevailing norms, addressed taboo subjects, or explored issues that were considered controversial in their time. The inclusion of diverse authors, from authors who were banned to authors who were banned, ensures that the corpus is representative of the broad range of issues that have prompted banning over the years.

Included Books

19th Century

- *Memoirs of a Woman of Pleasure* by John Cleland
- *The Sorrows of Young Werther* by Johann Wolfgang von Goethe
- *The Age of Reason* by Thomas Paine
- *Narrative of the Life of Frederick Douglass* by Frederick Douglass
- *Uncle Tom's Cabin* by Harriet Beecher Stowe
- *Leaves of Grass* by Walt Whitman
- *Incidents in the Life of a Slave Girl* by Harriet Jacobs
- *The Kama Sutra* (English translation by Sir Richard Francis Burton)
- *The Canterbury Tales* by Geoffrey Chaucer
- *The Adventures of Huckleberry Finn* by Mark Twain
- *The Picture of Dorian Gray* by Oscar Wilde
- *Jude the Obscure* by Thomas Hardy
- *On the Origin of Species* by Charles Darwin
- *The Decameron* by Giovanni Boccaccio

20th Century

- *Alice's Adventures in Wonderland* by Lewis Carroll
- *An American Tragedy* by Theodore Dreiser
- *Elmer Gantry* by Sinclair Lewis
- *Lady Chatterley's Lover* by D.H. Lawrence
- *Oil!* by Upton Sinclair
- *The Sun Also Rises* by Ernest Hemingway
- *Ulysses* by James Joyce
- *Antic Hay* by Aldous Huxley

- *Desire Under the Elms* by Eugene O'Neill

21st Century

- *1984* by George Orwell
- *The Catcher in the Rye* by J.D. Salinger
- *The Scarlet Letter* by Nathaniel Hawthorne
- *To Kill a Mockingbird* by Harper Lee
- *The Grapes of Wrath* by John Steinbeck
- *Slaughterhouse-Five* by Kurt Vonnegut
- *The Great Gatsby* by F. Scott Fitzgerald
- *The Color Purple* by Alice Walker
- *Gender Queer: A Memoir* by Maia Kobabe

STEP 2: Data Collection

Corpus Selection

Books were selected based on the following criteria:

1. Status of Censorship or Ban: Only books that were officially banned, rather than simply challenged, were included.
2. Restrictions in Specific Environments: Focus was placed on books banned at the statewide or national level. For more recent works, books banned in specific counties or school districts were also included due to changing censorship mechanisms.
3. Accessibility for Extraction: Books were chosen based on their availability in public digital repositories such as Project Gutenberg and other open digital libraries.

Sources

Books banned between the 19th and 20th centuries were retrieved from Project Gutenberg (HTML format). Books banned between the 1950s and the present were retrieved from Project Gutenberg and the Internet Archive(non-structured HTML files). Texts from the Internet Archive required manual retrieval and were organized into a separate directory (*NON_SCRAPABLES_TXT*).

Adjustments for Recent Texts

For books published after 2020, a new approach was adopted:

- Books banned in schools during the 2022–2023 school year were identified using PEN America's list of banned books.
- Florida and Texas, with the highest instances of censorship, were cross-referenced to refine the list.
- Only books available in public repositories were included.

STEP 3: Jupyter Notebook

Project Overview

The goal is to process a dataset of books, mainly sourced from Project Gutenberg and the Internet Archive, for analysis, including metadata extraction, text cleaning, and preparation for further analysis. The steps are divided into data retrieval, preprocessing, and output generation.

Data Retrieval & Organization

Objective: Download and organize book texts.

1. HTML File Retrieval:

HTML versions of the books were downloaded from Project Gutenberg and stored in a specified directory (HTML_BOOK_LIST). This directory will contain HTML files of books, including banned books between the 19th and 20th centuries (including "The Scarlet Letter" and "The Great Gatsby" from the period 1950-2024).

2. Non-Scrapable Books:

Books from the Internet Archive, which lacked structured HTML files, were manually retrieved, grouped into a separate directory (*NON_SCRAPABLES_TXT*), processed, and combined with other texts.

For this part of the project, we used Jupyter Notebook as our development environment and Python as the programming language to process and analyze the dataset of banned books. The Python libraries utilized include BeautifulSoup for HTML parsing, pandas for data organization, SpaCy for tokenization and lemmatization, and NLTK for stopword removal. The goal of this project is to extract, clean, and preprocess the text from various sources like Project Gutenberg and the Internet Archive, and prepare it for further analysis.

HTML Parsing, Metadata Extraction & File Organization

Objective: Parse the HTML files, extract metadata, and organize the data.

1. Directory Navigation & File Retrieval:

Os and glob libraries were utilized to navigate the HTML_BOOK_LIST directory (for Project Gutenberg files). All HTML files were loaded for parsing and text extraction. For books that couldn't be scraped (from the Internet Archive), they're added manually to the NON_SCRAPABLES_TXT folder.

2. HTML Parsing:

BeautifulSoup was used to parse each HTML file and extract the book's metadata, including Title, Author, and FIleName. This is achieved by locating the preamble or metadata section of the HTML structure. Regular expressions (re library) were used to match HTML patterns and extract metadata more precisely, ensuring consistent results.

3. Metadata Storage:

The extracted metadata (title, author, year, and source filename) was stored in a pandas DataFrame. For books without HTML scraping capabilities, metadata were retrieved and stored manually in a secondary non_scrapables.csv file. Both data sources (scraped and manually retrieved) were merged to ensure a comprehensive record of all books. The combined metadata was saved to a final CSV file (Banned_Books_Metadata.csv) with UTF-8 encoding (UTF-8-sig) to prevent issues with special characters and ensure cross-platform compatibility.

Text Preprocessing (Cleaning & Tokenization)

Objective: Clean, tokenize, and preprocess text for analysis.

1. Text Loading:

The text files, stored in the HTML_BOOK_LIST and NON_SCRAPABLES_TXT directories, were programmatically loaded for preprocessing.

2. Lowercasing:

All text was converted to lowercase to standardize it and prevent issues with case-sensitive discrepancies during analysis.

3. Text Cleaning:

The text underwent several cleaning steps:

- Remove Punctuation and Numbers: Using the string module and custom filters, punctuation was stripped and numerical data from the text, ensuring we focused only on meaningful words.
- Remove Stopwords: Using the nltk.corpus.stopwords library common stopwords were removed that don't contribute to the analytical process. Custom stopwords were also added based on initial dataset analysis.
- Non-ASCII Characters: Any non-ASCII characters were removed to avoid encoding issues and ensure uniformity.

4. Tokenization:

   Using SpaCy, the cleaned text was tokenized into words, which are the basic units for analysis. Tokenization prepares the text for further manipulation and ensures that it's ready for more advanced techniques like frequency analysis.

5. Lemmatization:

   To standardize word variations, SpaCy's lemmatization feature was applied to reduce words to their base or root forms, ensuring consistency in the dataset.

6. Whitespace Normalization:

   Extra spaces and unnecessary whitespace were cleaned up, making the text more consistent and easier to process.

7. Saving Preprocessed Text:

   The cleaned and tokenized text was stored in a dictionary (preprocessed_texts), where filenames were used as keys and the cleaned text as values. The text was then saved as .txt files in the FINAL_TXT directory, ensuring each book's processed text was ready for analysis.

Output Generation & Final Dataset Preparation

Objective: Save the processed data and prepare it for further analysis.

1. Final Output - Cleaned Text Files: After preprocessing, the cleaned text for each book was saved in the FINAL_TXT directory as .txt files. This organized structure ensured that each book was in a consistent format, ready for detailed topic modeling analysis.
2. Final Metadata CSV: The final metadata, including information such as the book's title, author, and year, was saved in the Banned_Books_Metadata.csv file. This CSV contained all the relevant metadata for each book, ensuring it was easily accessible for future analysis or reporting.
3. Documenting the Process: The entire workflow was documented for reproducibility. By structuring the code and data in a clear manner, it was ensured that others could replicate the process or build upon it for additional research.

STEP 4: Analysis

Topic Modeling

Objective

The main goal of topic modeling in this project was to identify recurring themes, concerns, and societal issues present in banned books. By analyzing these texts, we aim to uncover how different time periods and themes contribute to the banning patterns, revealing insights into changing social norms, political climates, and cultural concerns.

To achieve this, we utilized Voyant Tools and Zeta Analysis (using R project, a software environment for statistical computing and graphics), two powerful tools for text analysis. These tools helped us process and analyze the data, uncovering meaningful patterns related to censorship and societal attitudes.

1. Voyant Tools: Voyant is an interactive web-based tool designed for text mining and visualization. It was used for:
    - Exploratory Text Analysis: Visualizing word frequencies, trends, and associations across texts.
    - Topic Modeling: Generating topics that emerge from the texts by analyzing word co-occurrence patterns and associations, allowing us to explore common themes.
    - Visualization: Generating word clouds, frequency graphs, and trends to better understand the relationships between different terms and topics across time periods.
2. Zeta Analysis: The R project package, (using the oppose function to do Zeta analysis) provided a more in-depth analysis of the stylistic and thematic markers in the books. With Zeta analysis, we were able to:
    - Cluster Texts: Group books based on similar stylistic features and thematic content.
    - Identify Censorship Themes: Categorize texts by their thematic concerns and explore stylistic differences across time periods.

- Stylistic Features: Identify recurring stylistic features within clusters, contributing to the understanding of banning patterns and their connection to different themes.

Cluster Texts by Time Period and Themes

To ensure a more structured analysis, we divided the texts into two primary groupings:

1. Time Period Division:

The first division was based on the century in which books were banned. This created three main categories:

- 19th century
- 1900-1949
- 1950-2024

This separation allowed for the comparison of banning trends across different historical contexts.

2. Censorship Themes:

In addition to the time division, we also categorized the texts based on six distinct themes related to the theme they have. Each theme represents a different aspect of society that has often been challenged by censorship. The six themes include:

- *Human Desire and Relationships*: Books that explore themes of passion, intimacy, and sexual relationships, which have often faced censorship due to their controversial content. Examples include:
  - "Memoirs of a Woman of Pleasure" by John Cleland
  - "The Kama Sutra" - English translation by Sir Richard Francis Burton
  - "Lady Chatterley's Lover" by D.H. Lawrence

        - "The Da Vinci Code" by Dan Brown
        - "Call Me by Your Name" by Andre Aciman
    - *Religion and Morality*: Texts that challenge traditional religious beliefs or present controversial views on morality. Examples include:
        - "The Age of Reason" by Thomas Paine
        - "The Picture of Dorian Gray" by Oscar Wilde
        - "The Handmaid's Tale" by Margaret Atwood
        - "To Kill a Mockingbird" by Harper Lee
    - *Slavery, Racism, and Power*: Books that address issues of racial injustice, slavery, and power dynamics. Examples include:
        - "Narrative of the Life of Frederick Douglass" by Frederick Douglass
        - "Uncle Tom's Cabin" by Harriet Beecher Stowe
        - "The Color Purple" by Alice Walker
        - "The Grapes of Wrath" by John Steinbeck
    - *Nature, Science, and Philosophy:* Books that challenge scientific norms or present unconventional philosophical ideas. Examples include:
        - "Leaves of Grass" by Walt Whitman
        - "On the Origin of Species" by Charles Darwin
        - "Slaughterhouse Five" by Kurt Vonnegut
    - *Mental Health and Existential Struggles*: Books that delve into the complexities of mental health and personal existential crises. Examples include:
        - "The Sorrows of Young Werther" by Johann Wolfgang von Goethe
        - "The Catcher in the Rye" by J.D. Salinger

        - "Thirteen Reasons Why" by Jay Asher
    - *Identity and Growth*: Books that explore themes of personal identity, growth, and coming-of-age stories. Examples include:
        - "Alice's Adventures in Wonderland" by Lewis Carroll
        - "Ulysses" by James Joyce
        - "Eleanor and Park" by Rainbow Rowell
        - "Gender Queer: A Memoir" by Maia Kobabe
    - *Social Critique and Class Dynamics:* Books that critique social structures, class dynamics, and the impact of wealth and power. Examples include:
        - "The Canterbury Tales" by Geoffrey Chaucer
        - "1984" by George Orwell
        - "The Great Gatsby" by F. Scott Fitzgerald
        - "Oil!" by Upton Sinclair

By categorizing the texts into these six themes, we could more effectively identify trends, patterns, and reasons for banning across various genres, subjects, and societal concerns. This theme-based approach was used to group books with similar concerns, allowing for a more focused analysis of how censorship manifests in different contexts.

Steps Followed for Voyant Tools

1. Data Preparation:

The processed texts from the Jupyter Notebook (FINAL_TXT) were organized into three separate corpora: one for the whole corpus (all texts), one for the 19th-century texts, and one for the 20th century and beyond.

2. Uploading Data to Voyant:

The cleaned and categorized text files were uploaded into Voyant Tools. The three corpora were stored in separate folders within Voyant to facilitate comparison between time periods (19th vs. 20th century).

3. Exploratory Data Analysis:

After uploading, Voyant's default exploratory tools were used. The Word Cloud displayed the most frequent terms across the corpus, while the Corpus Summary provided essential stats like the total word count, number of documents, and distinct terms.

4. Word Frequency & Trend Analysis:

Using the Term Frequency tool, the frequency of key terms was visualized across the texts. This helped identify recurring words associated with major themes like freedom, religion, and sex.

Trend Graphs were used to track how the frequency of these terms fluctuated over time, which allowed us to see how specific themes were emphasized or censored in different periods.

5. Topic Modeling:

Voyant's Topic Modeling feature helped us identify dominant themes and clusters of related terms across the entire corpus.

The topic modeling provided insights into how specific subjects were consistently linked to books that were banned, revealing underlying cultural and social concerns.

6. Comparative Analysis:

One of the most valuable features of Voyant is its ability to compare multiple corpora. We used this to compare texts from different time periods (19th century vs. 20th century. This comparison assists in observing shifts in the language used and how different themes were treated by authors over time, providing insights into evolving patterns of banning.

7. Visualization & Reporting:

Voyant's visualization tools, including Trend Graphs, and Word Clouds, were used to present the data visually. These visualizations provided an intuitive way to interpret the results and communicate the findings.

The results were compiled into a report that summarized key insights, such as the most common themes, significant shifts in word frequency, and topic modeling results. The visualizations and graphs helped illustrate the analysis clearly.

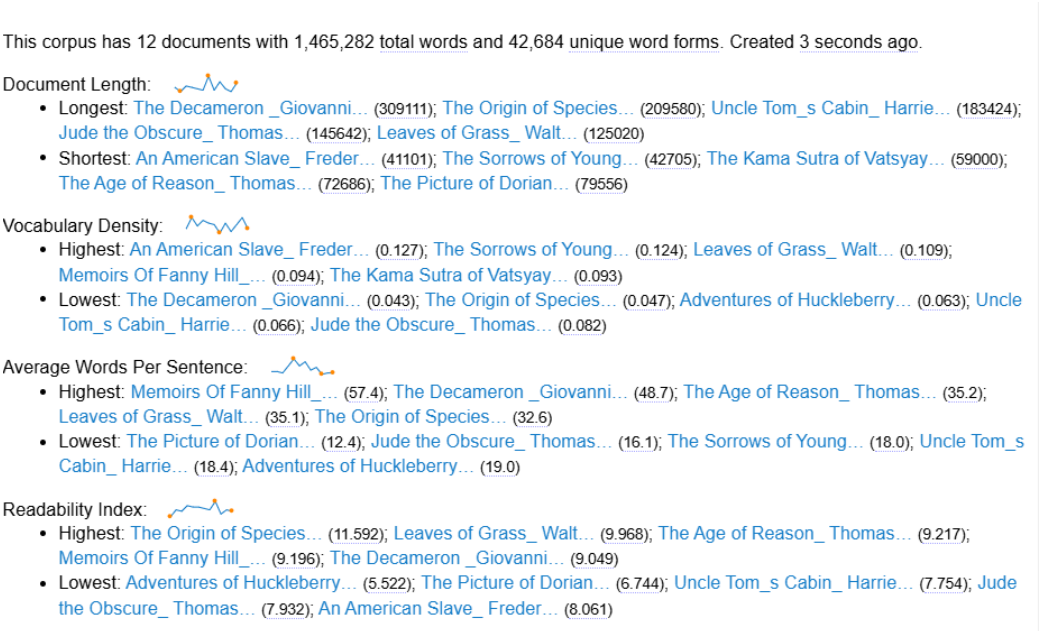
Voyant Report

Voyant is a corpus analysis tool being employed to study the themes of banned and censored books from the 19th century to 2025. The analysis will be divided into three periods. The first phase will focus on the 19th to 20th centuries, utilizing a corpus comprising 12 novels. The second goes from 1900 to 1949, and the third from 1950 to 2024.

First Period: 19th to 20th centuries

In the next visualisation, the most frequent words can be seen. The word most frequently used is "said" which is due to the narrative style that these works possess where characters interact and the story is recounted in the past tense. Additionally, many of the most common words in this analysis are adjectives and adverbs.

For the next visualisation, the corpus of the 19th century to the 20th century can be seen which have undergone some tests that are prominent such as the length of documents, readability, and most importantly, the most used words.



This corpus has 12 documents with 1,465,282 total words and 42,684 unique word forms. Created 3 seconds ago.

Document Length:

- Longest: The Decameron _Giovanni… (309111); The Origin of Species… (209580); Uncle Tom_s Cabin_ Harrie… (183424); Jude the Obscure_ Thomas… (145642); Leaves of Grass_ Walt… (125020)
- Shortest: An American Slave_ Freder… (41101); The Sorrows of Young… (42705); The Kama Sutra of Vatsyay… (59000); The Age of Reason_ Thomas… (72686); The Picture of Dorian… (79556)

Vocabulary Density:

- Highest: An American Slave_ Freder… (0.127); The Sorrows of Young… (0.124); Leaves of Grass_ Walt… (0.109); Memoirs Of Fanny Hill_… (0.094); The Kama Sutra of Vatsyay… (0.093)
- Lowest: The Decameron _Giovanni… (0.043); The Origin of Species… (0.047); Adventures of Huckleberry… (0.063); Uncle Tom_s Cabin_ Harrie… (0.066); Jude the Obscure_ Thomas… (0.082)

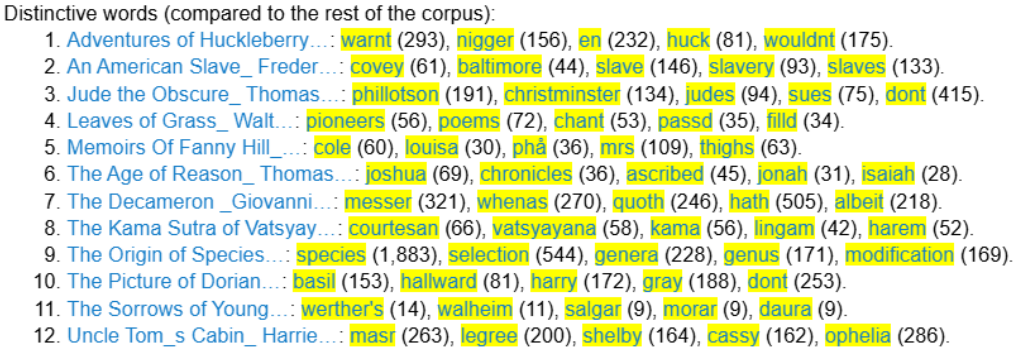
Average Words Per Sentence:

- Highest: Memoirs Of Fanny Hill_… (57.4); The Decameron _Giovanni… (48.7); The Age of Reason_ Thomas… (35.2); Leaves of Grass_ Walt… (35.1); The Origin of Species… (32.6)
- Lowest: The Picture of Dorian… (12.4); Jude the Obscure_ Thomas… (16.1); The Sorrows of Young… (18.0); Uncle Tom_s Cabin_ Harrie… (18.4); Adventures of Huckleberry… (19.0)

Readability Index:

- Highest: The Origin of Species… (11.592); Leaves of Grass_ Walt… (9.968); The Age of Reason_ Thomas… (9.217); Memoirs Of Fanny Hill_… (9.196); The Decameron _Giovanni… (9.049)
- Lowest: Adventures of Huckleberry… (5.522); The Picture of Dorian… (6.744); Uncle Tom_s Cabin_ Harrie… (7.754); Jude the Obscure_ Thomas… (7.932); An American Slave_ Freder… (8.061)

For a further study of the theme, the use of stop-words must be employed, as the appearance of words such as “said” “man” “come” and “like” does not prove theme alterations or theme indications. For the stop words, the characters’ names would be added for a more fruitful study.

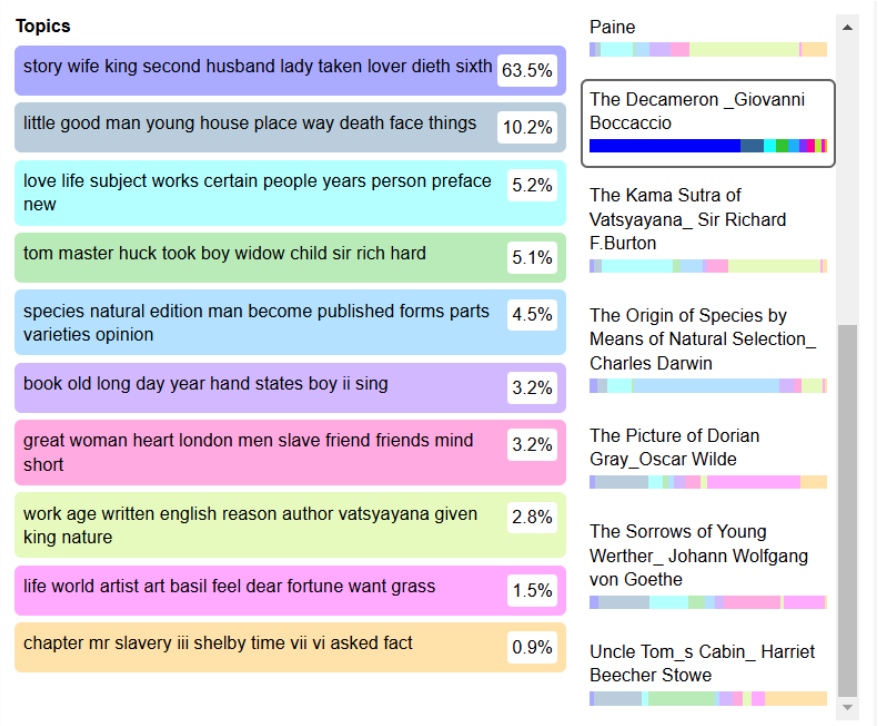


Distinctive words (compared to the rest of the corpus):

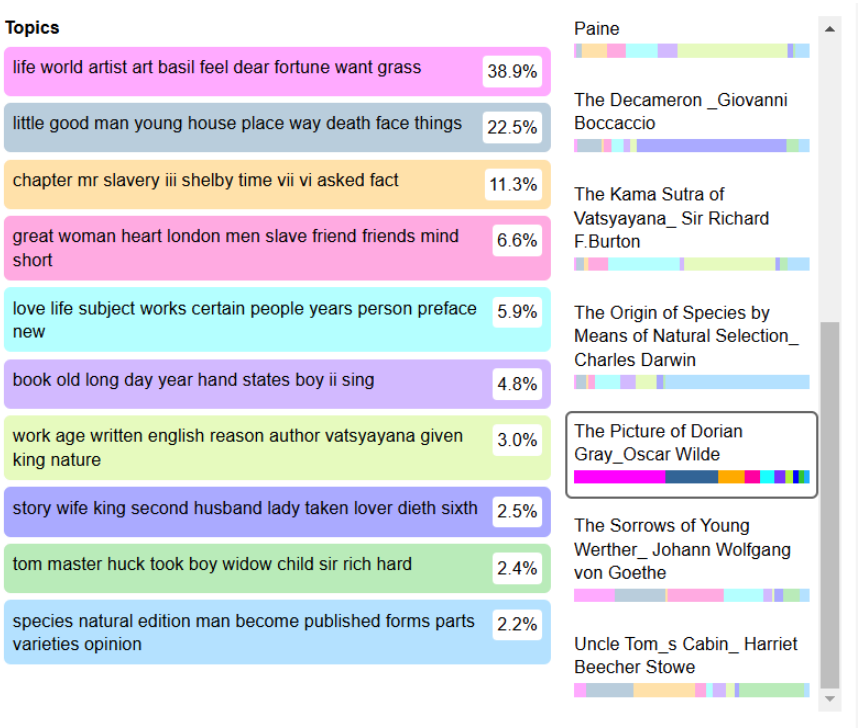
1. Adventures of Huckleberry…: warnt (293), nigger (156), en (232), huck (81), wouldnt (175).
2. An American Slave_ Freder…: covey (61), baltimore (44), slave (146), slavery (93), slaves (133).
3. Jude the Obscure_ Thomas…: phillotson (191), christminster (134), judes (94), sues (75), dont (415).
4. Leaves of Grass_ Walt…: pioneers (56), poems (72), chant (53), passd (35), filld (34).
5. Memoirs Of Fanny Hill_…: cole (60), louisa (30), phå (36), mrs (109), thighs (63).
6. The Age of Reason_ Thomas…: joshua (69), chronicles (36), ascribed (45), jonah (31), isaiah (28).
7. The Decameron _Giovanni…: messer (321), whenas (270), quoth (246), hath (505), albeit (218).
8. The Kama Sutra of Vatsyay…: courtesan (66), vatsyayana (58), kama (56), lingam (42), harem (52).
9. The Origin of Species…: species (1,883), selection (544), genera (228), genus (171), modification (169).
10. The Picture of Dorian…: basil (153), hallward (81), harry (172), gray (188), dont (253).
11. The Sorrows of Young…: werther's (14), walheim (11), salgar (9), morar (9), daura (9).
12. Uncle Tom_s Cabin_ Harrie…: masr (263), legree (200), shelby (164), cassy (162), ophelia (286).

Some problematic topics are being reflected in the results. In *The Adventures of Huckleberry Finn*, one of the most used words is the n-word. It should be remembered that it was either censored or banned mainly for racial slurs. Similarly, *Memoirs of Fanny Hill*, features the word “thighs”, among its most common themes, highlighting the explicit content that has contributed to its censoring for “explicit language”. To gain a deeper understanding of the themes presented in each novel, topic modeling will be utilized with Voyant. This tool is effective for identifying word

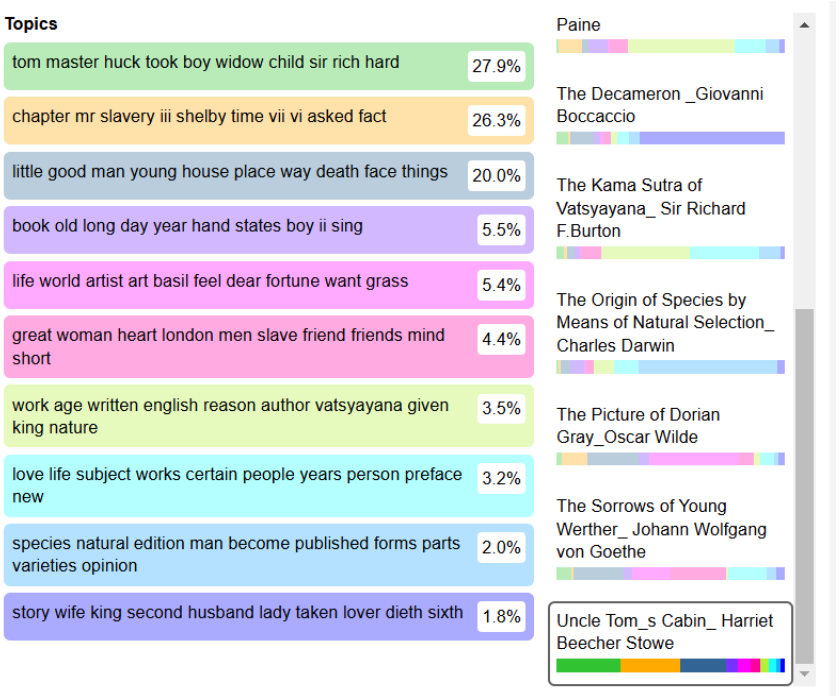
patterns, as it employs algorithms to group frequently co-occurring words into categories, ultimately creating intricate topics and their percentage of occurrence in each novel.

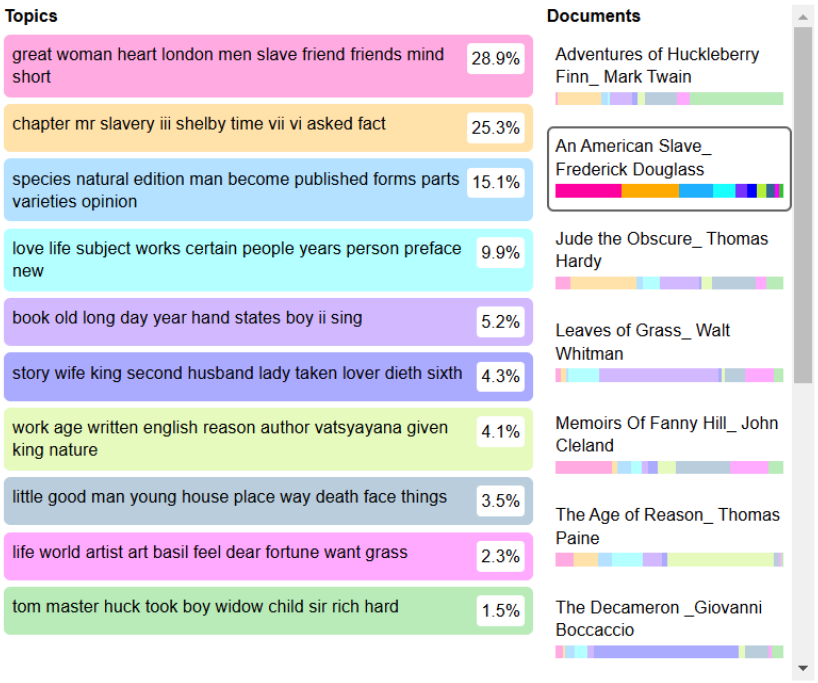


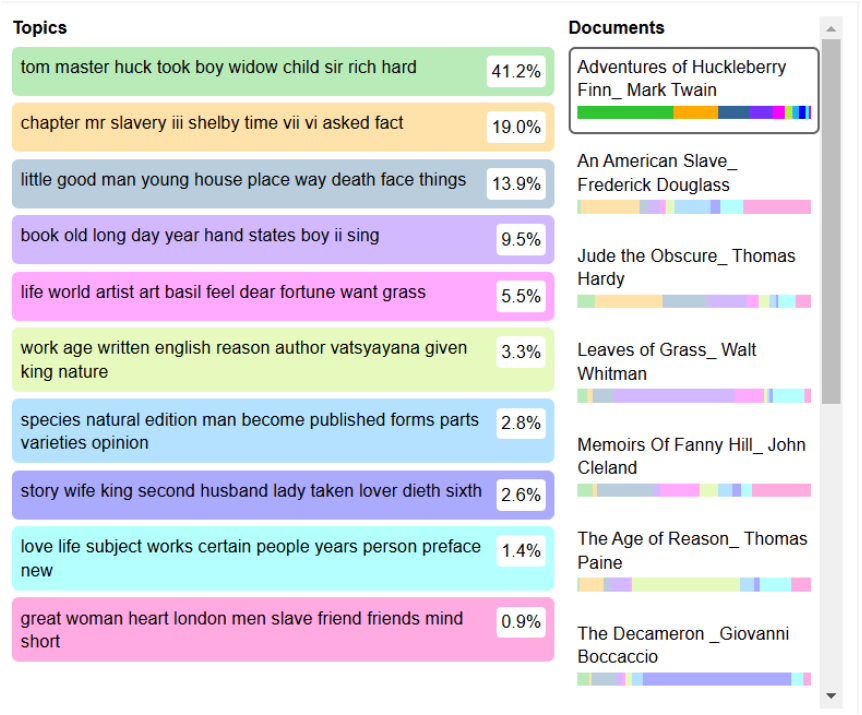
The Decameron, which was banned for its sexual encounters, in the topic modeling can be seen in its main theme when adultery is drafted, “second husband lady taken lover dieth”, being one of the most abundant topics with a 63.5%. Moreover, some examples are being discarded in this study as with *The Origin of Species* by Darwin, its themes are prevalent but they are not demonstrative as it is the main theme of evolution. In the means of Oscar Wilde, in the topic modeling it can be seen how the themes that appeared in the study, do not represent the reasoning behind its banning or censoring. This might be because it is sometimes about the words used, but the implication of these, and instead of distant reading, close reading would be needed to prove the theme correlation with banning or censoring.



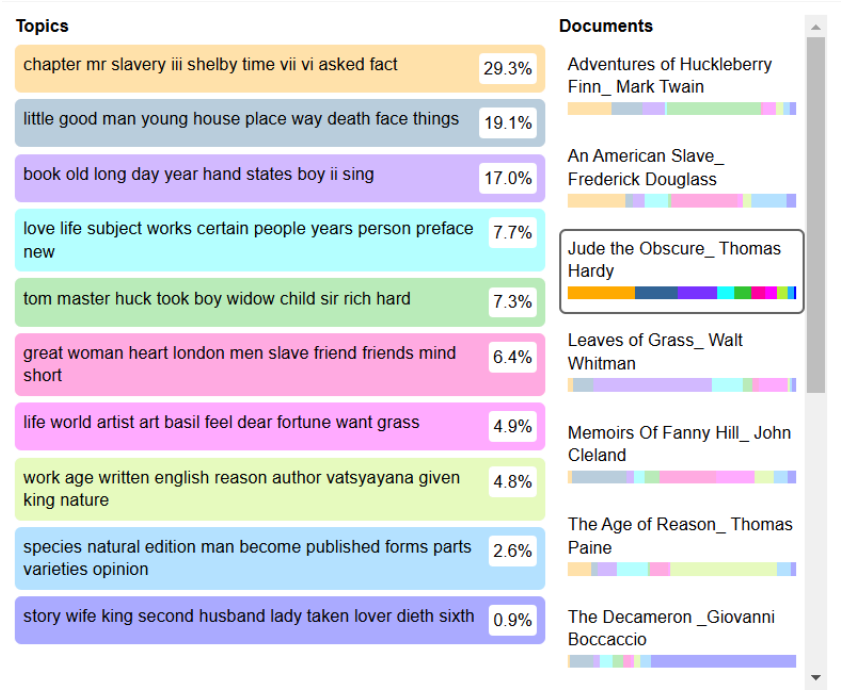
In the case of *Uncle Tom's Cabin*, the means of slavery can be seen described in the novel, with a high occurrence of master, and slavery. The same themes seem to repeat with an *American Slave* by Douglass and *The Adventures of Huckleberry Finn* by Twain.







It can be seen how a multitude of topics are involved in Thomas Hardy's novel *Jude the Obscure*.

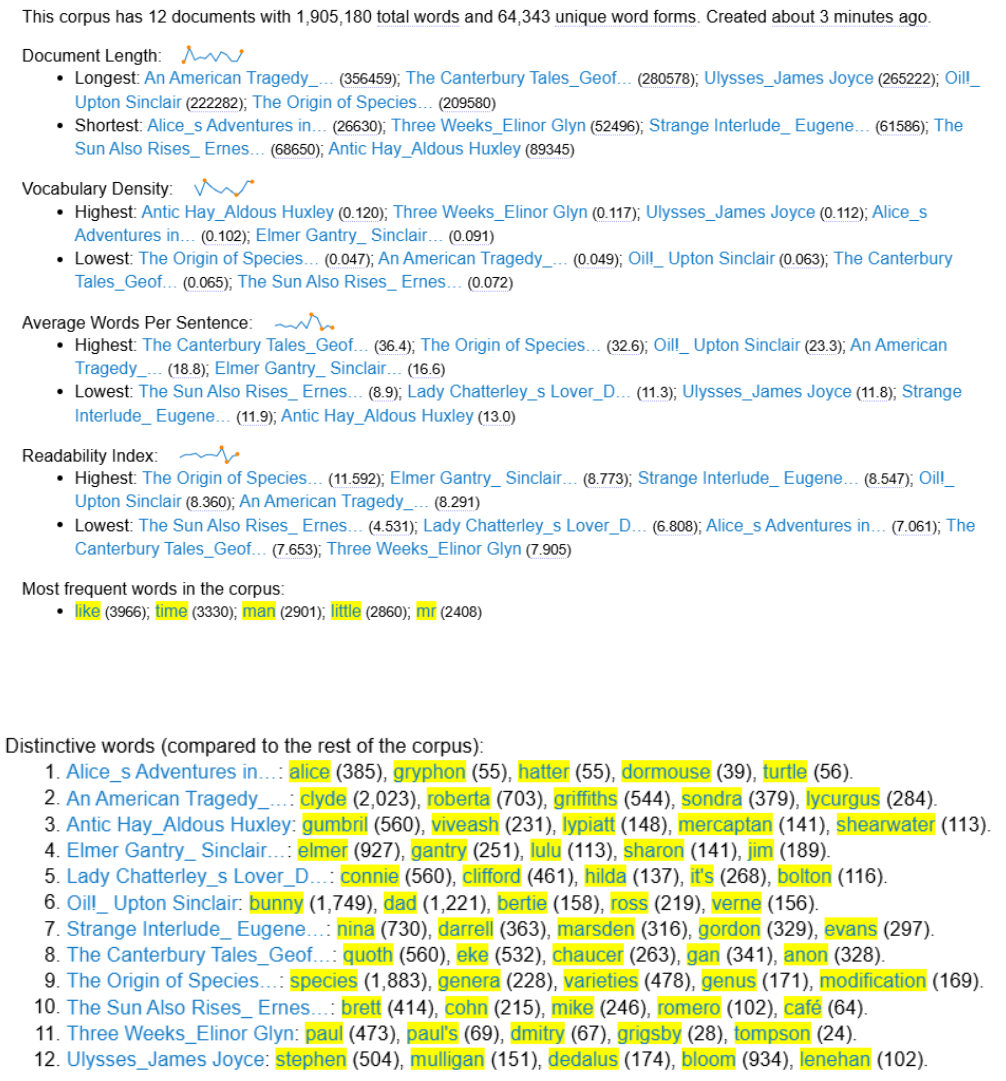


Second Period: 1900 to 1949

The period from 1900 to 1949 will be thoroughly explored in the following section. The corpus is composed of 12 documents, and the same techniques are applied. It must be taken into consideration the fact that there is a replicate book in the 19th century, which is Darwin's book *The Origin of Evolution*, as it was also banned or censored within this century.

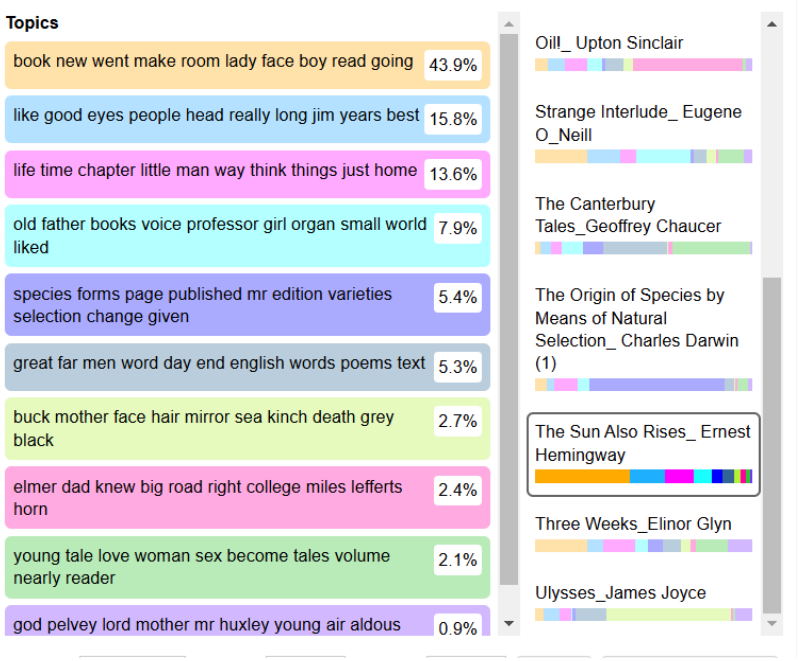
The visualisation portrays the most frequent words in a bigger letter size. Stop words were added including verbs for a further analysis of the themes. One of the

words that is frequently repeated is Clyde, which belongs to Alice's adventures, and it means "stupid".

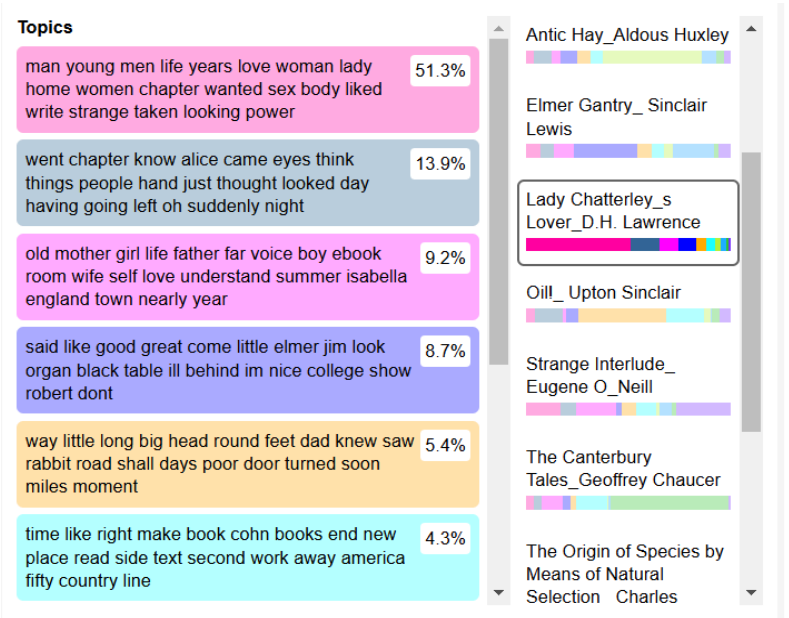


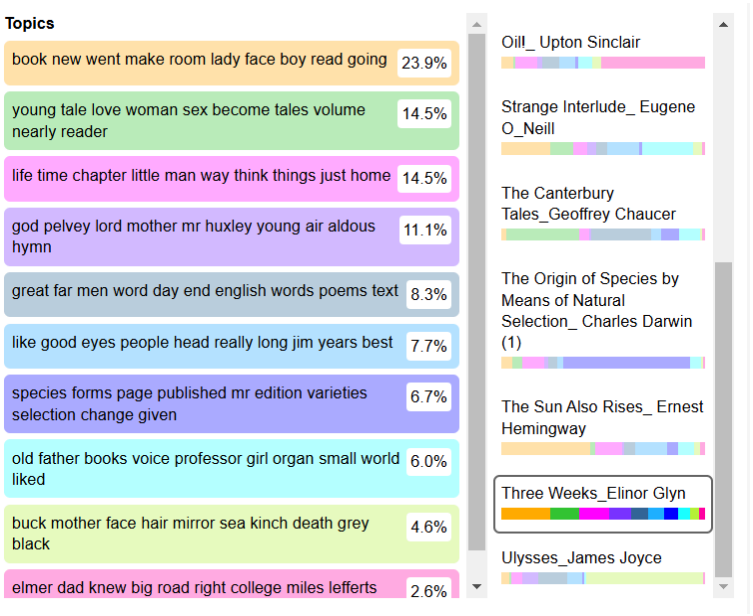
Frequent words in this study are populated by character's names, therefore, for further analysis topic modeling is also used.

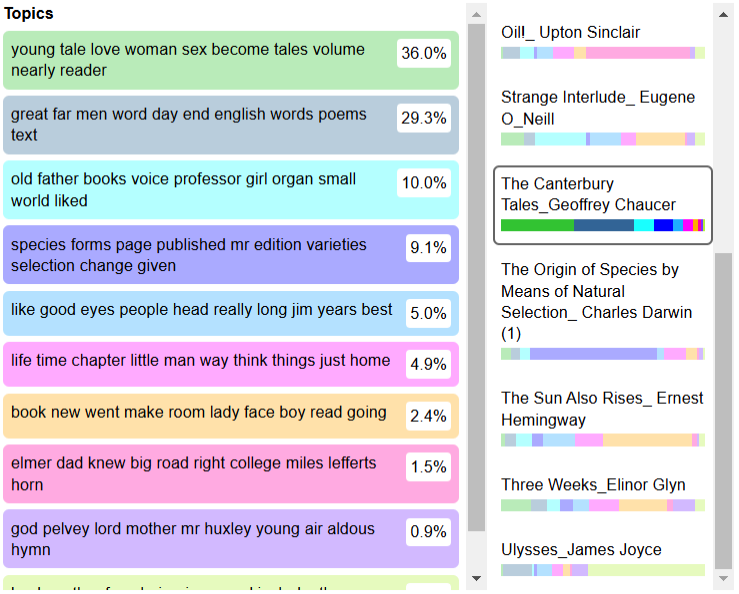
In the next visualisation, it can be seen how *The Sun Also Rises*, banned in Boston for its themes, can be seen that solely the 2.1% is elicited. And 43.9% could be also underlying those themes.



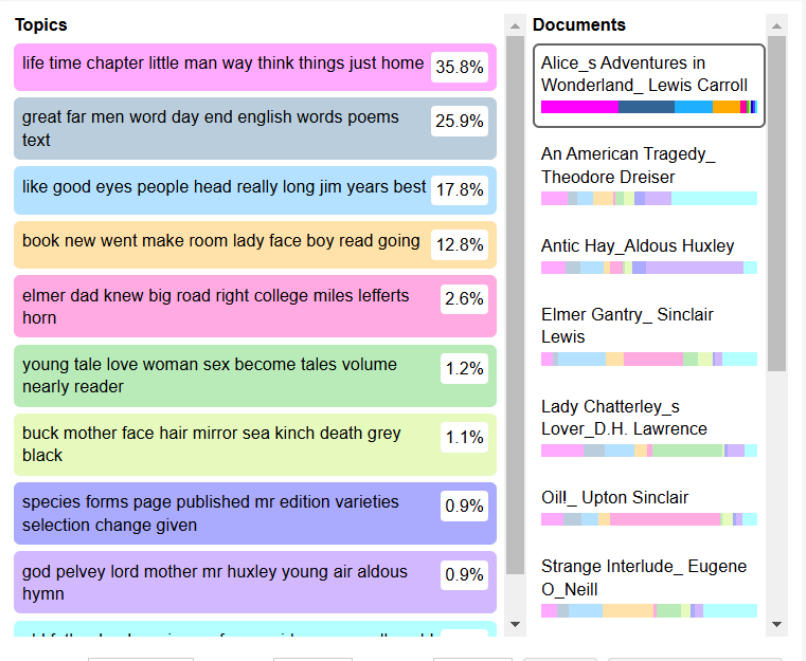
On the other side, in *Three Weeks*, and *Canterbury Tales* sex is a more prevalent theme in those books. *Lady Chatterley's* main themes of love, sex and female power are represented in the topic.





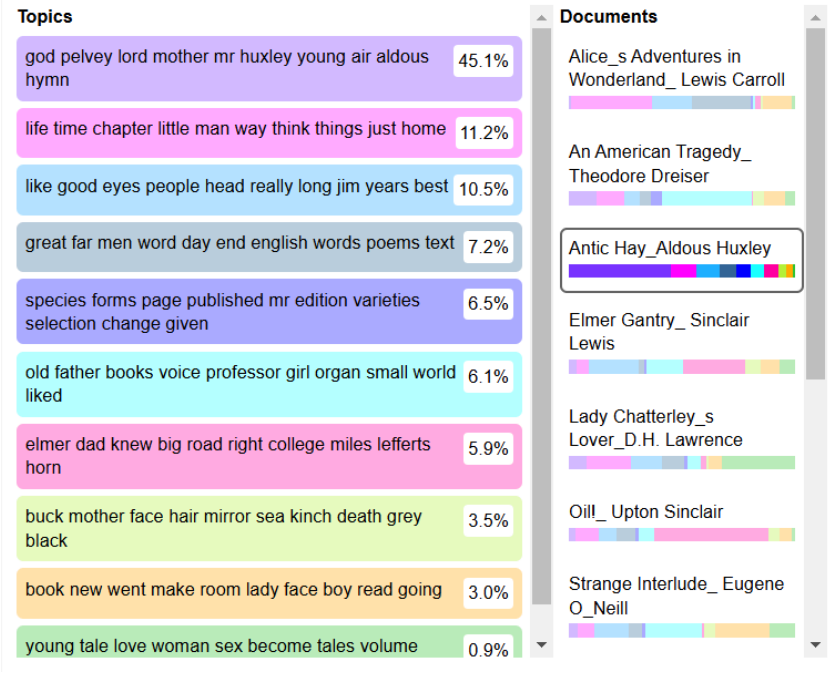


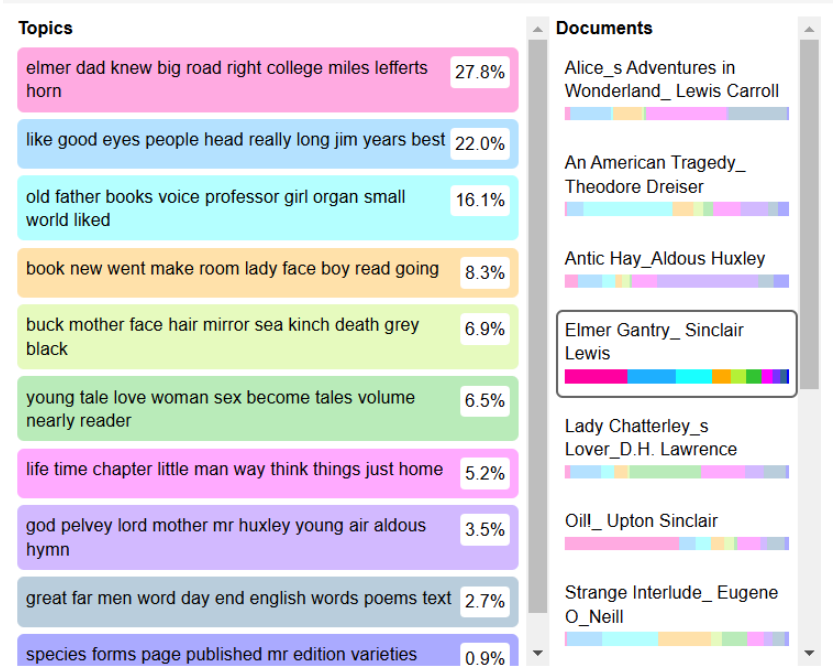
In the case of *Alice's Adventures in Wonderland,* the allusion of magical elements and creatures can be identified.



The keywords for each topic reveal thematic connections to the books in the dataset. Topic 1, with terms like “god”, “lord”, “mother”, and “mr” aligns with themes of identity and growth, reflecting narratives centred on personal journeys and self-discovery. Topic 2, including words such as “life”, “time”, “chapter”, “man”, and “home”, ties to themes of social critique and dynamics, often exploring societal structures and human interactions. Topic 3, highlighted by terms like “good”, “eyes”, “people”, and “head” suggests themes of mental health and existential struggles, delving into the human psyche and the search for meaning. These thematic groupings help contextualize the literary focus of the books analyzed.

Topic 4, with keywords such as “great”, “far”, “men”, “word”, “day”, and “end”, appears to focus on themes of broad exploration and the passage of time, possibly reflecting narratives that involve journeys, historical contexts, or philosophical musings. The presence of terms like “great” and “far” suggests an emphasis on expansive ideas or physical and metaphorical distances, while words like “day” and “end” hint at temporality and finality. This topic might be associated with stories or reflections that address the human experience in relation to time, legacy, or the vastness of the world. The moderate percentage (**7.2%**) suggests that this theme plays a meaningful, though not dominant, role in the dataset.





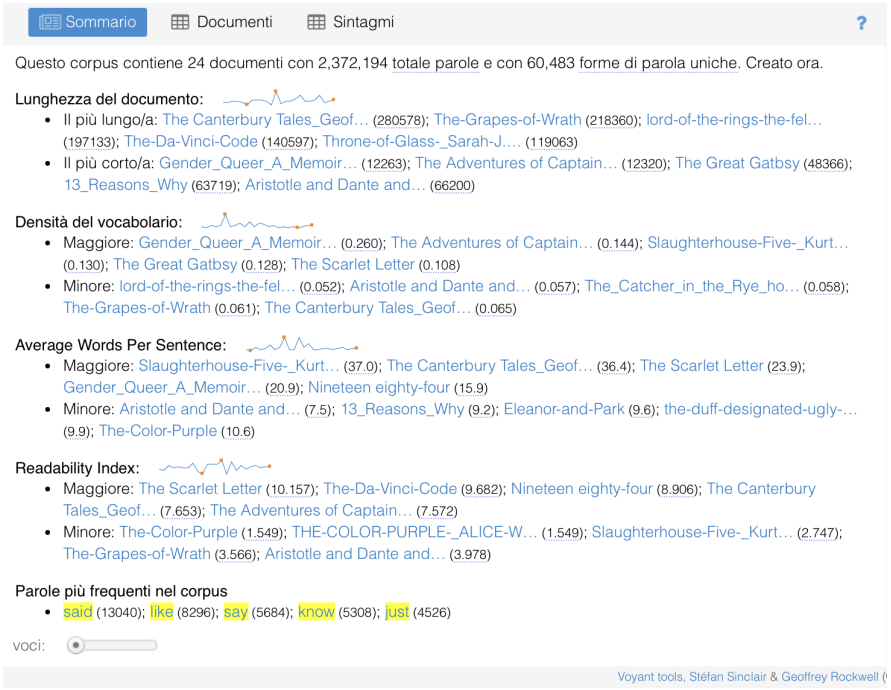
Furthermore, the use of the Voyant tool and especially topic modeling, can be fruitful in some instances where the theme is easily represented within the frequent words. However, there are instances where this is not evident. In such cases, Zeta, another topic modeling tool would be used.

Third Period: 1950 to 2025

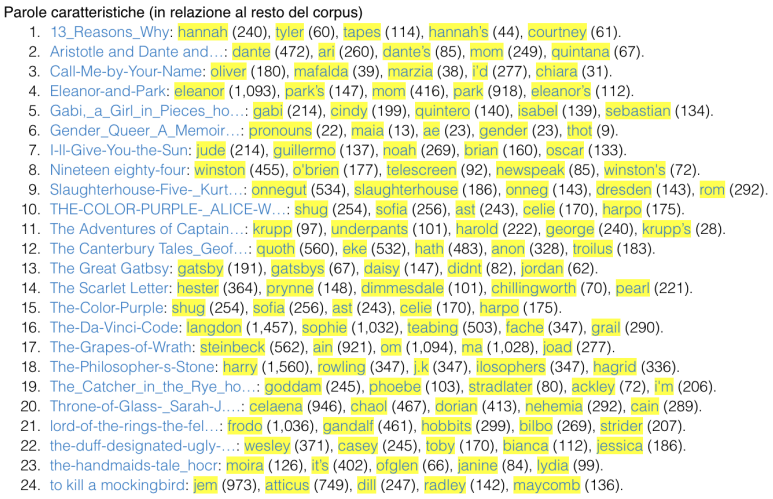
The analysis regarding books banned between the 1950s and present-day is performed on a corpus of 24 books banned between 1950 and 2023. The bans occurred mostly at the County level in different States, as a Federal Law regarding book censorship no longer exists in the United States. Regarding the present day, a decision has been made to consider books from a list obtained from PEN America, which includes all books banned in schools across every U.S. state during the 2022-2023 academic year. Specifically, a cross-reference was conducted to identify the same banned books in schools in Florida and Texas, two of the states with the most stringent censorship policies. Although, currently, the protections afforded by the First Amendment concerning freedom of expression and access to digital forms of texts make large-scale book censorship unfeasible, an exponential increase in book bans has been recorded across various school districts. Conservative states such as Texas and Florida lead the rankings, with thousands of books banned or challenged each year.

Information from the Summary panel

The parameters do not disclose relevant patterns chronologically, as all of them include older and more recent books. Also, no specific words are mentioned, so it is not possible to connect specific topics to them.



Most common words for each text without adding stop words



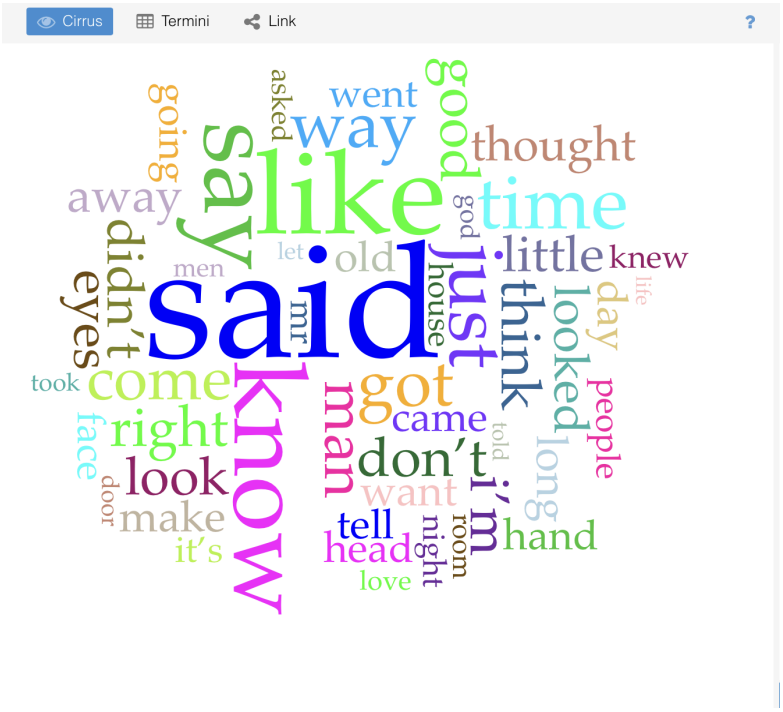
Immediate assumptions based on the most frequently occurring words reveal significant insights into the themes and controversies surrounding various texts. In many instances, the most common words identified are the names of characters, indicating a strong focus on character-driven narratives.

For example, *Gender Queer: A Memoir* clearly highlights the central themes of the book, as evidenced by the frequent appearance of terms such as “pronouns,” “gender,” and “thot.” These words reflect the book's exploration of LGBTQ+ themes and its use

of language that some may consider inappropriate, contributing to its ban. Similarly, *The Catcher in the Rye* has faced censorship due to its perceived promotion of rebellion and inclusion of foul language. This is underscored by the predominant presence of the word “goddam,” which aligns with the book's controversial reputation. Additionally, *The Adventures of Captain Underpants* was flagged for multiple reasons, one of which pertains to its portrayal of partial nudity. The inclusion of the word "underpants" in both the title and the list of most common words suggests that this element played a significant role in its classification as inappropriate for certain audiences.

These examples illustrate how analyzing frequently occurring words can provide valuable insights into the reasons behind book bans and the thematic content that raises concerns among censors.

Most common words altogether

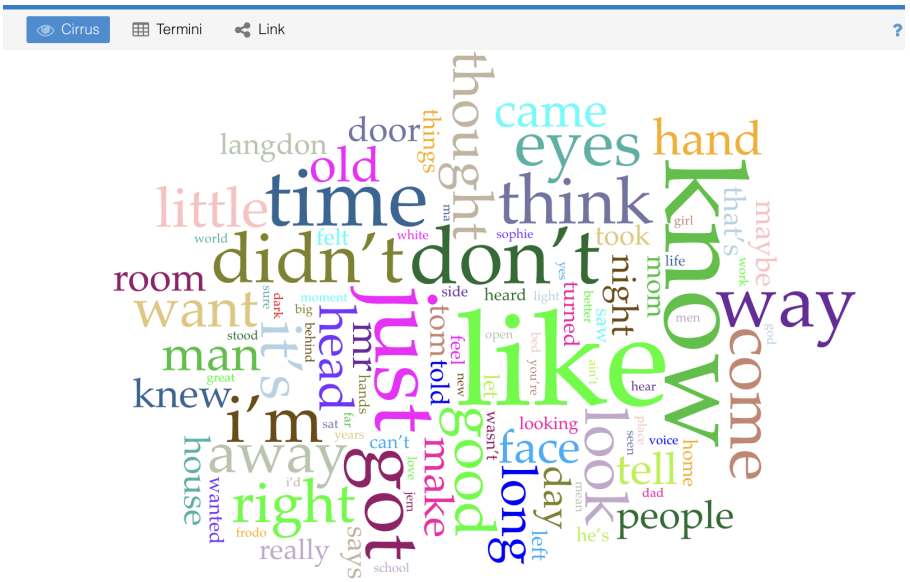


The predominance of the verb “said” suggests a preference for narrative styles across the analyzed texts, indicating dialogue-driven stories with many interactions among the characters.

Dialogues usually create a sense of dynamicity that helps readers engage with the story: this is particularly common within the genre of fiction, to which a few of

the books belong (especially the ones banned in the 2000s). Moreover, since the ban involves school districts and libraries, particularly in the present day, a lot of the books belong to the "coming-of-age" literary genre, targeting young adults and teenagers, hence the dialogue-ridden stories, reflected in the presence of other verbs such as "say", "asked". "Thought" and "know" are also present, which might suggest more introspective storylines, as some of the books focus on self-discovery and reflection (*Gabi: a girl in pieces, Gender Queer: a Memoir, Thirteen reasons why*).
The rest of the words are not very articulated or seem to carry particular meanings that might unveil a blatant topic as a target for banning.

100 most common words with stop words



For a more relevant analysis, some terms are excluded as stopwords, such as verbs (say, say, it's, I'm, looked) and characters' names, which are usually highly present and appear in the "most common words" lists for most books.

We can still observe a preponderance of action verbs, negative forms, character names, and anatomical references (e.g., "head," "hand," "eyes," "face"). The latter may be indicative of a heightened focus on physical descriptions and actions, which could potentially be perceived as obscene or inappropriate, thus heightening censorship efforts.

It is noteworthy that the majority of books subject to bans during this temporal period, particularly in the 2000s, were predominantly removed from school libraries. This suggests that the content was deemed unsuitable for juvenile readership, reflecting societal concerns about age-appropriate literature and the protection of young minds from allegedly unsuitable material. The prevalence of such bans in educational institutions underscores the ongoing debate regarding the balance between intellectual freedom and the perceived need to shield younger audiences from controversial or explicit content.

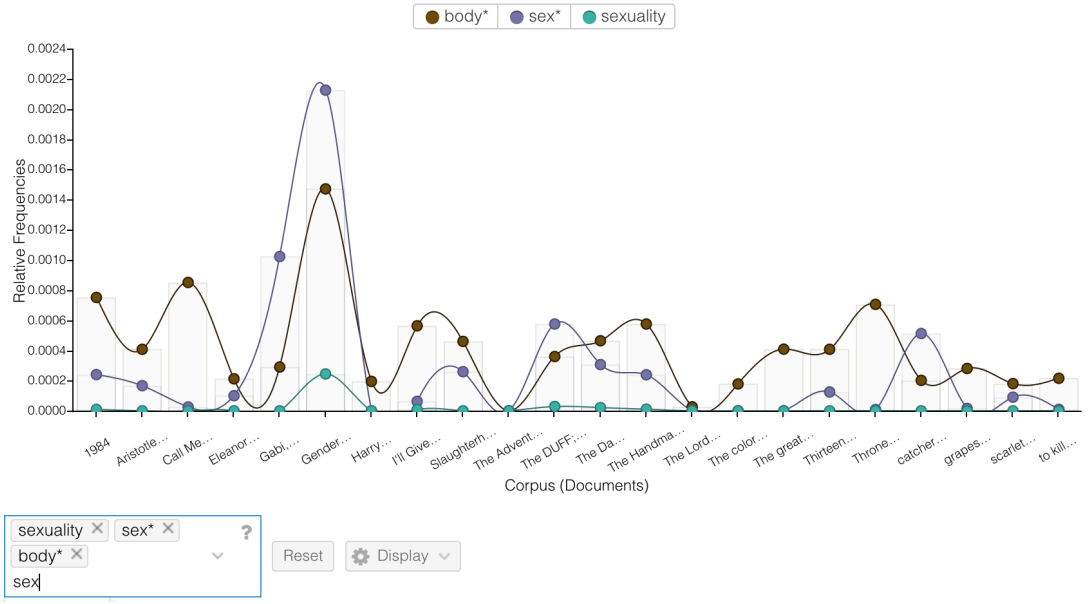
Parole caratteristiche (in relazione al resto del corpus)

1. 13_Reasons_Why: tapes (114), hannah's (44), tony (50), rosie's (34), it's (190).
2. Aristotle and Dante and…: dante's (85), mom (249), quintana (67), didn't (432), don't (396).
3. Call-Me-by-Your-Name: i'd (277), chiara (31), anchise (27), piazzetta (25), vimini (24).
4. Eleanor-and-Park: park's (147), mom (416), park (918), eleanor's (112), tina (135).
5. Gabi,_a_Girl_in_Pieces_ho…: gabi (214), mom (256), tia (101), agirlin (76), martin (151).
6. Gender_Queer_A_Memoir…: pronouns (22), maia (13), ae (23), gender (23), thot (9).
7. I-ll-Give-You-the-Sun: oscar (133), it's (487), he's (388), i'm (581), mom (220).
8. Nineteen eighty-four: telescreen (92), newspeak (85), oceania (62), proles (47), ministry (64).
9. Slaughterhouse-Five-_Kurt…: onnegut (534), slaughterhouse (186), onneg (143), dresden (143), rom (292).
10. THE-COLOR-PURPLE-_ALICE-W…: shug (254), ast (243), git (139), olivia (58), mama (68).
11. The Adventures of Captain…: underpants (101), krupp's (28), diaper (27), doo (28), hama (16).
12. The Canterbury Tales_Geof…: hath (483), troilus (183), ye (881), evry (160), unto (393).
13. The Great Gatbsy: didnt (82), wasnt (39), dont (96), daisys (35), wilson (62).
14. The Scarlet Letter: dimmesdale (101), chillingworth (70), hesters (44), roger (77), minister (156).
15. The-Color-Purple: shug (254), ast (243), git (139), olivia (58), mama (68).
16. The-Da-Vinci-Code: teabing (503), fache (347), grail (290), keystone (168), collet (167).
17. The-Grapes-of-Wrath: om (1,094), ma (1,028), rath (361), ruthie (224), jus (420).
18. The-Philosopher-s-Stone: ston (347), ph (347), harry's (113), mcgonagall (94), dudley (117).
19. The_Catcher_in_the_Rye_ho…: goddam (245), stradlater (80), didn't (325), you're (108), chrissake (32).
20. Throne-of-Glass-_Sarah-J…: assassin (165), endovier (95), eyllwe (84), she'd (311), champions (89).
21. lord-of-the-rings-the-fel…: hobbits (299), strider (207), shire (234), aragorn (202), boromir (146).
22. the-duff-designated-ugly-…: casey (245), didn't (299), wesley's (55), duffy (43), i'm (244).
23. the-handmaids-tale_hocr: it's (402), serena (63), commander (110), luke (84), rita (48).
24. to kill a mockingbird: jury (61), jem's (60), taylor (86), gilmer (54), stephanie (51).

A closer inspection of the most common words list, after attempting to remove as many irrelevant words and characters' names, reveals a number of interesting patterns. For instance, *The Da Vinci Code* appears to feature often the term "grail", a sacred relic of great significance for the Catholic Church, which is here, according to the contexts, treated lightheartedly as an object to be found. The connection is relevant, as the book was harshly criticized for being offensive to Christianity and its blasphemous contents. Moreover, *Throne of Glass* presents the word "assassin", disclosing in some way a violent nature that has resulted in its ban from Florida and Texas schools.

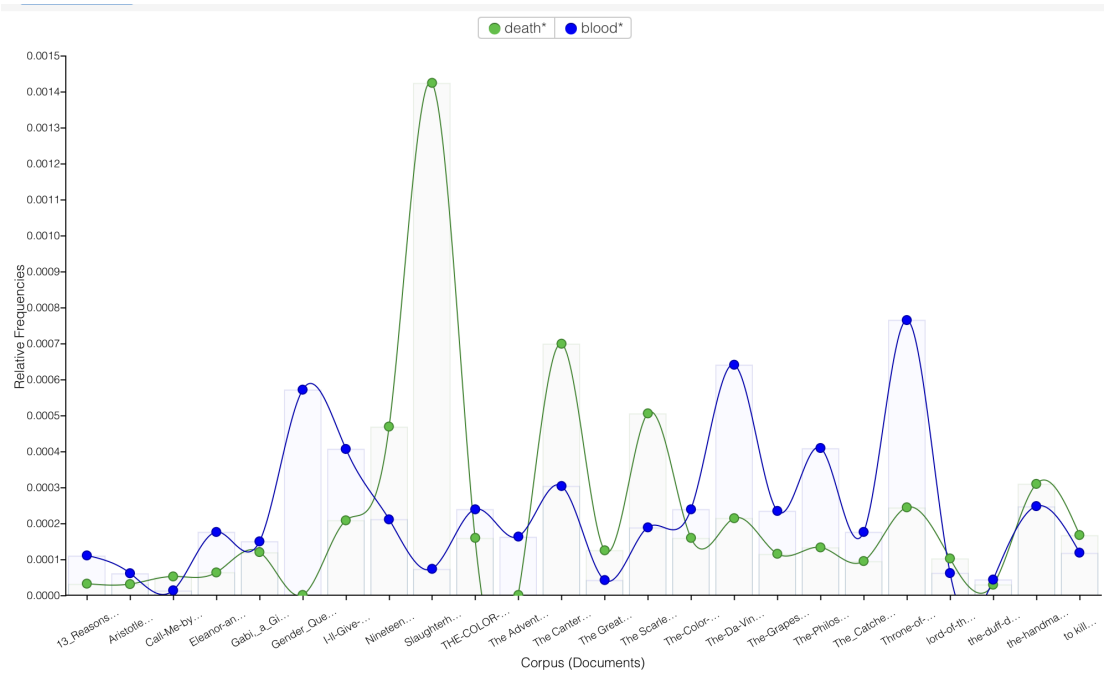
Finally, *1984* includes as frequent words “telescreen”, “newspeak”, and “ministry”, which highlight the novel's focus on oppressive government control and the manipulation of truth. The book was primarily banned due to its exploration of social and political themes, particularly its critique of totalitarianism and perceived communist propaganda. In reality, they stem from misunderstandings of its message; rather than promoting communism, it serves as a cautionary tale against authoritarianism.

Topic-focused research



Many of the books in the list were banned for obscenity or sexual references, which as the trends graph highlights are indeed quite frequent themes, appearing, among others, in *Throne of Glass, The DUFF: Designated Ugly Fat Friend, Gender Queer: a Memoir, Aristotle and Dante discover the Secrets of the Universe, Slaughterhouse Five, Gabi, a Girl in Pieces.* The words “body” and “sex” appear in a large number of books, often following the same trends. Many of the books subject to the 2022-2023 Florida and Texas bans, most of which of fairly recent publication (2010 onwards), were targeted for tackling LGBTQ+ themes. Those are relatively challenging to spot using a distant reading tool such as Voyant, as they might be embedded into conversations about identity, love or romance which could apply to

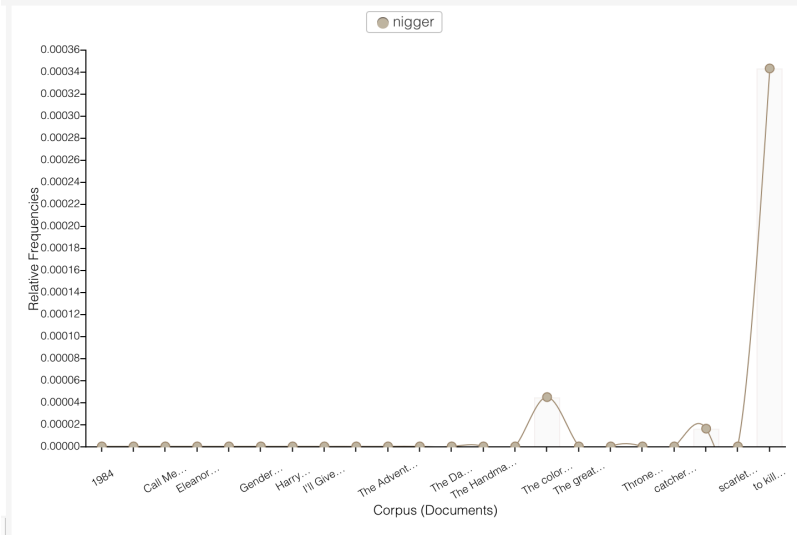
heterosexual characters as well. For instance, portrayed in the trends graph is the word “sexuality”, whose presence is often entailed with questions about it and, therefore, mentions of queerness: it shows predominantly and correctly in *Gender Queer: a Memoir*, and slightly in *The DUFF* and *I’ll give you the sun*, which all explore such themes, but not in other books that famously tackle them, as *Call me by your name*.



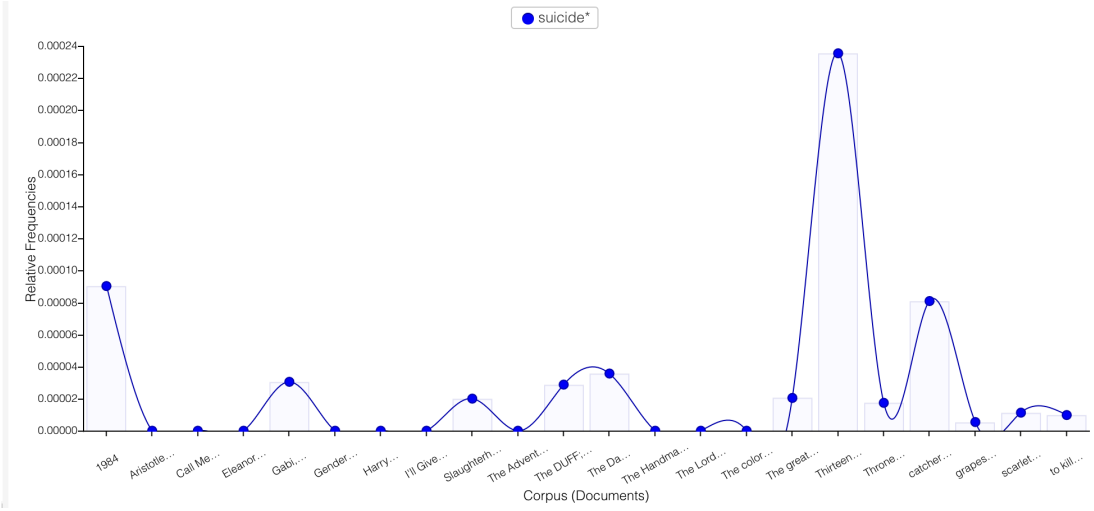
Although not forcefully linked with violence, many books present a high frequency of the words “death” and “blood”, which suggest crude language, another popular topic subject to censorship. The theme persists both in the 1950s to 1990s (*The Catcher in the Rye, The Scarlet Letter, Nineteen Eighty Four, The Grapes of Wrath, Slaughterhouse Five)* and present day (*Eleanor and Park, Throne of Glass, I’ll Give You the Sun, The Handmaid’s Tale)* implying a continuous sensitivity on the matter.

It is challenging to identify words related to the reasons for book bans using Voyant, as many instances require targeted research. For example, the term "nigger" in *To Kill a Mockingbird* has consistently been a focal point of discussion surrounding the book. Its presence can also be noted in *The Color Purple*, banned in the 1980s for

its brutal portrayal of Black men, and in *The Catcher in the Rye,* which depicts the pervasive racism present in society during the time period in which the novel is set.



The same applies to research focusing on the term "suicide," which is undeniably controversial and frequently appears in books that have been banned for addressing mental health issues. The inclusion of this term often reflects the sensitive nature of discussions surrounding mental health, as well as societal attitudes toward such topics. The trend graph shows its peak in *Thirteen Reasons Why, a* popular yet divisive coming-of-age book centred on the topic, often challenged.



For others, it's simpler, because the overall story is based on the very topic that cost it the ban, for instance, "Gender queer: a memoir", whose most common words feature "gender", "pronouns", and "thot", suggesting a focus on questioning one's identity and possible foul language.

Conclusion

The analysis of distinctive words from banned books in the 19th and 20th centuries reveals notable patterns and themes that might have contributed to their censorship.

Based on the provided data:

19th Century Banned Books

The distinctive words in the 19th-century list highlight themes related to **race, religion, morality, and societal norms**:

- **Race and Slavery**: Words like "slave", "slavery", and "nigger" (in *Adventures of Huckleberry Finn* and *An American Slave*) directly address racial inequality and the horrors of slavery, which were highly contentious topics in 19th-century America.
- **Sexuality and Morality**: Words like "louisa", "mrs", and "thighs" (in *Memoirs of Fanny Hill*) suggest the focus on sexuality, which was deemed immoral and indecent for public readership during this time.
- **Religious Critique**: Words such as "chronicles", "jonah", and "isaiah" (*The Age of Reason*) indicate challenges to religious orthodoxy, often leading to bans in a deeply religious society.
- **Philosophical and Political Ideas**: Terms like "modification", "species", and "genus" (in *The Origin of Species*) reflect the challenge posed by Darwin's evolutionary theory to establish religious and cultural beliefs.

These themes illustrate how books were banned for confronting the dominant social hierarchies, questioning religion, and addressing taboo topics like race and sexuality during the 19th century.

20th Century Banned Books

The distinctive words in the 20th-century list suggest modern societal critiques, psychological exploration, and explicit sexual or personal content:

- **Sexual Content and Relationships**: Words like “connie”, “clifford”, and “hilda” (in *Lady Chatterley's Lover*) and “paul” (in *Three Weeks*) reflect explicit exploration of sexuality and relationships, which were deemed inappropriate.
- **Psychological and Existential Themes**: Terms like “stephen”, “mulligan”, “Dedalus,” and “bloom” (in *Ulysses*) point to experimental narratives and existential introspection, often misunderstood or considered subversive.
- **Critiques of Social Structures**: Words such as “gantry”, “mercaptan”, and “ross” (in *Elmer Gantry* and *Oil!*) reflect critiques of capitalism, religion, and societal norms, mirroring broader socio-political tensions.
- **Racial and Class Issues**: Words like “griffs”, “lycurgus”, and “sondra” (in *An American Tragedy*) suggest explorations of class disparities and racial issues, echoing 19th-century concerns but framed in a modern context.

Common Patterns

1. **Challenging Authority**: Both sets of books question authority—be it religious (for example: *The Age of Reason* vs. *Elmer Gantry*), societal, or governmental structures.
2. **Taboo Topics**: Race, class, and sexuality feature prominently in both centuries, showing how addressing these issues often led to censorship.
3. **Moral Panic**: Both eras reflect moral outrage: the 19th century focused on "indecency" tied to race and religion, while the 20th century added concerns about psychological and sexual freedom.
4. **Cultural Evolution**: The shift from critiques of slavery and religion in the 19th century to explorations of psychology and individualism in the 20th century reflects broader societal transformations.

The majority of books subjected to censorship from the second half of the 20th century through the early 21st century tend to focus on the following themes:

Racial issues: the use of racial slurs encountered in some of the novels (*To Kill a Mockingbird, The Color Purple, The Catcher in the Rye)*, mainly banned between the 1960s and 1980s suggests a profound discomfort with the exploration of racism and its implications in society.

Political ideas: as books like *1984* prove, the second half of the 20th century continued to experience censorship based on political themes, influenced by the McCarthy era of the 1950s, especially targeting alleged Communist propaganda.

Identity and Growth: words like “gender”, and “pronouns”, encountered in books of recent publication (*Gender Queer: a Memoir*) suggest a very contemporary focus on finding one’s identity, at the cost of fighting societal norms, leading to bans in conservative States like Texas and Florida.

Sexual contents: as the targeted research for words like "body," "sex," and “sexuality” shows, the latter are some of the most common themes for bans throughout the late 20th century to the present day, with a particular emphasis on more recent books, such as *Throne of Glass, The DUFF: Designated Ugly Fat Friend, Gender Queer: a Memoir, Aristotle and Dante Discover the Secrets of the Universe, Slaughterhouse Five, Gabi, a Girl in Pieces.* As mentioned in the previous point, the conservative nature of the two Southern States taken into consideration quickly resorts to censorship in order to allegedly protect children and young adults.

Violence and foul language, mental health: words such as "blood," "death," and “assassin” (*Throne of Glass, 1984, The Da Vinci Code, Slaughterhouse Five)* evoke a taste for darkness and violence that has consistently faced opposition through bans, from the 1950s to the present day. Foul language, often related to depictions of sexuality or slurs ("thot," *Gabi, a Girl in Pieces),* is also persistently challenged, as well as forms of self-harm, translated with the use of the word “suicide” (*Thirteen Reasons Why, The DUFF: Designated Ugly Fat Friend, The Da Vinci Code, 1984).*

Steps Followed for Zeta Analysis

1. Ensure UTF-8 Encoding:

Before beginning any analysis, it was essential to ensure all text files were properly encoded in UTF-8 format. This ensured that no encoding issues would interfere with the processing of text, especially in texts containing special characters or symbols from non-English languages.

2. Data Organization:

To facilitate a comprehensive analysis, the text data was organized into three main folders based on two key criteria: the century the books were banned and the thematic content of the texts.

3. Analysis by Century of Banning:

For the first analysis:

1. Primary Set: This folder contained 19th-century banned books, organized for initial analysis as a primary set.
2. Secondary Set: This folder contained early 20th-century banned books until 1949.

For the second analysis:

1. Primary Set: This folder included banned books from 1900-1949, allowing for comparative analysis with the 19th-century corpus.
2. Secondary Set: The secondary analytical segment included books banned between 1950 and present day for a progressive investigation along the decades.

4. Analysis by Century of Banning and Theme:
    - For each of the six themes identified within the corpus, we further divided the texts into primary and secondary sets based on the century of censorship:

For the first analysis:

1. Primary Set: For each theme, this folder contained 19th-century banned books within that theme.

2. Secondary set: For each theme, the folder featured early 20th-century banned books.

For the second analysis:

1. Primary Set: For the second part of the analysis, the folder featured early 20th-century banned books.
2. Secondary Set: This folder included 20th-century and beyond banned books for the same theme, facilitating comparison across time periods.

5. Including Data in Zeta Analysis:

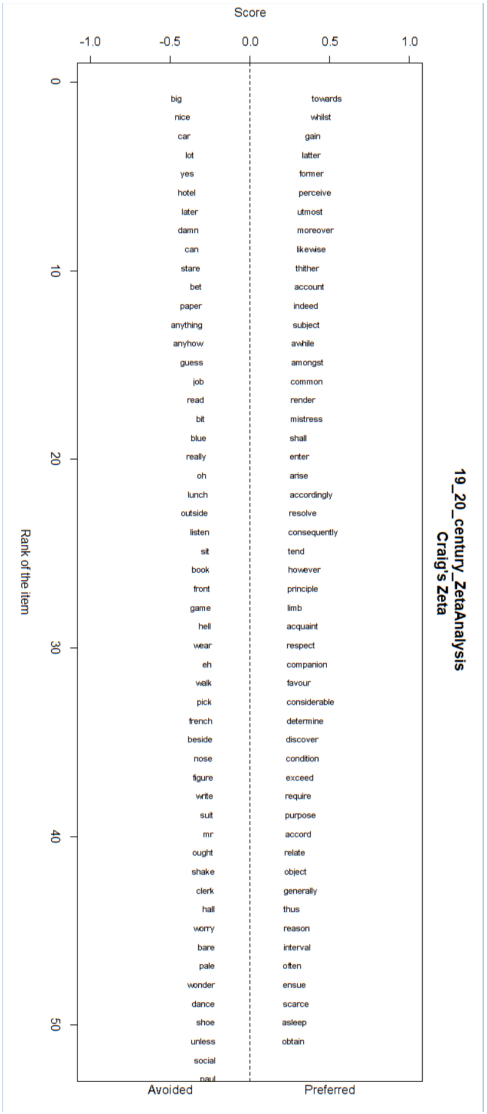
After organizing the texts by time period and theme, the oppose function in R project was used for more in-depth stylistic and thematic clustering. This analysis allowed us to uncover dominant topics and trends that linked specific time periods with particular themes of banning.

6. Data Analysis:

Using Zeta Analysis, dominant topics were extracted for each cluster, enabling to observe which themes were most frequently censored and how they varied between time periods. This step helped in drawing connections between stylistic features, thematic content, and their connection to censorship.

Division Based on the Century the Books Were Banned: 19th vs. 20th Century

To begin the analysis, the dataset was divided based on the century in which the books were banned: the 19th and the 20th centuries. This division allows for a comparative exploration of the themes, patterns, and societal concerns that led to censorship during these distinct periods. By analyzing the distinctive words associated with each century's banned books, we can uncover shifts in cultural priorities, moral standards, and the types of ideas deemed controversial or threatening across time.

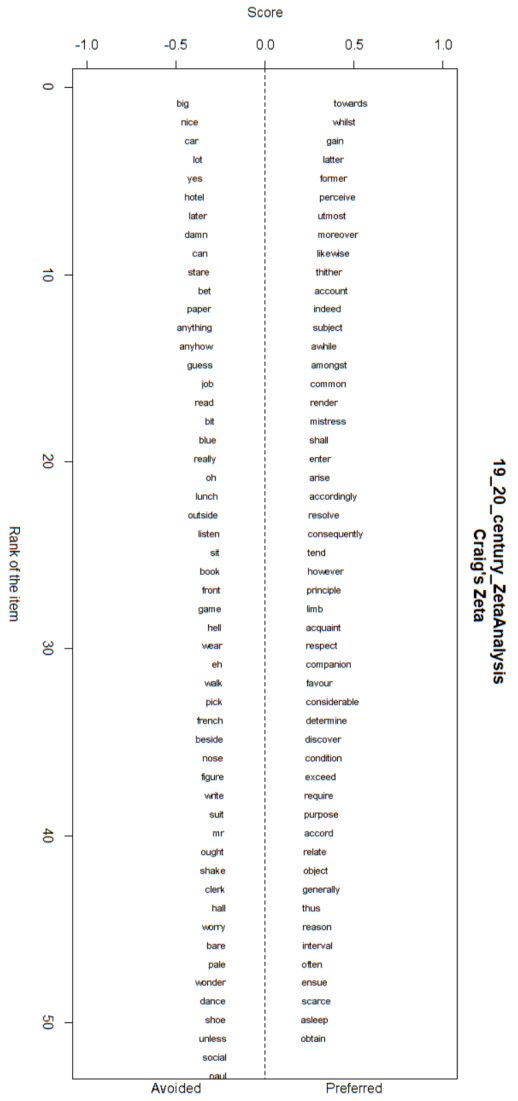


For this analysis, the only adjustment made was setting the occurrence threshold at 2. This means that terms or words appearing at least twice in the corpus were considered for the analysis. As the occurrence threshold increases, the focus narrows to more commonly appearing terms, reducing noise from rare or unique words but potentially excluding nuanced or less frequent associations.

What are Correlation Thresholds?

Correlation thresholds in Zeta Analysis determine what qualifies as a “strong” association between terms based on their co-occurrence. By adjusting this threshold, you can control how often two terms must appear together to be considered related. A higher threshold highlights stronger and more consistent relationships, emphasizing

terms that co-occur frequently across the dataset, but it may also exclude subtler or less common connections. Conversely, a lower threshold captures more varied associations, though it risks including weaker or coincidental relationships. These adjustments allow the analysis to balance between identifying broad, dominant patterns and uncovering finer, less frequent connections.



Setting the correlation threshold to 10 highlights only the strongest and most frequent associations between terms. While this reduces noise and emphasizes dominant patterns in the dataset, it may also overlook less frequent but potentially meaningful connections.

General Structure of Zeta analysis

1. y-axis (Rank of the item): Items (likely words or phrases) are ranked by their distinctiveness in either the 19th or 20th-century banned books.
2. x-axis (Score): Indicates the degree to which an item is "preferred" (positive values) or "avoided" (negative values) in the primary set (19th-century banned books) compared to the secondary set (20th-century banned books).

Analysis

Both the Zeta Analysis diagrams, one with an occurrence threshold of 2 and the other with 10, display identical words and scores, indicating no observable differences between them. This suggests that adjusting the occurrence threshold in this case did not impact the results or reveal additional insights. Consequently, it is not possible to derive any meaningful distinctions or variations between the two analyses, which may indicate that the dataset's structure or term frequencies remain consistent across these thresholds.

Preferred words (preferred from 19th-century Banned Books/ avoided by 20th-century ones)

The words on the right, with positive scores, are ranked as more characteristic of 19th-century texts. These include terms like “towards”, “whilst”, “moreover”, “consequently”, “principle”, “obtain”, and “respect”. They suggest a more formal, older, or literary style typical of 19th-century language.

Avoided words (preferred in 20th-century Banned Books/ avoided in 19th-century ones)

Words on the left, with negative scores, are characteristic of 20th-century texts. These include terms like “big”, “nice”, “car”, “hotel”, “later”, “lunch”, “game”, “damn”, “beside”, and “shoe”. These words reflect more casual, conversational language (“big”, “damn”) and focus on modern settings and objects (“car”, “shoe”). They suggest a shift from the formal tone of 19th-century texts to modern, more casual or colloquial language, with a focus on personal or immediate concerns.

Themes Across Time: Concluding Observations on 19th and 20th-Century Texts

Formal and Abstract Language in the 19th Century: Words like “peculiar” and “consequently” reflect the formal, structured style typical of older literature. Banned books from the 19th century likely contained more philosophical, reflective, or academic content, as seen in words such as “accordingly”, “considerable”, “principle”, and “consequently”.

Modern Language in the 20th Century: Words like “car”, “city”, and “damn” suggest a casual, everyday style focused on realism. Banned books from the 20th century appear to address more relatable or contemporary life scenarios, with words like “car”, “hotel”, “game”, and “big” pointing to a more modern, accessible language.

Findings

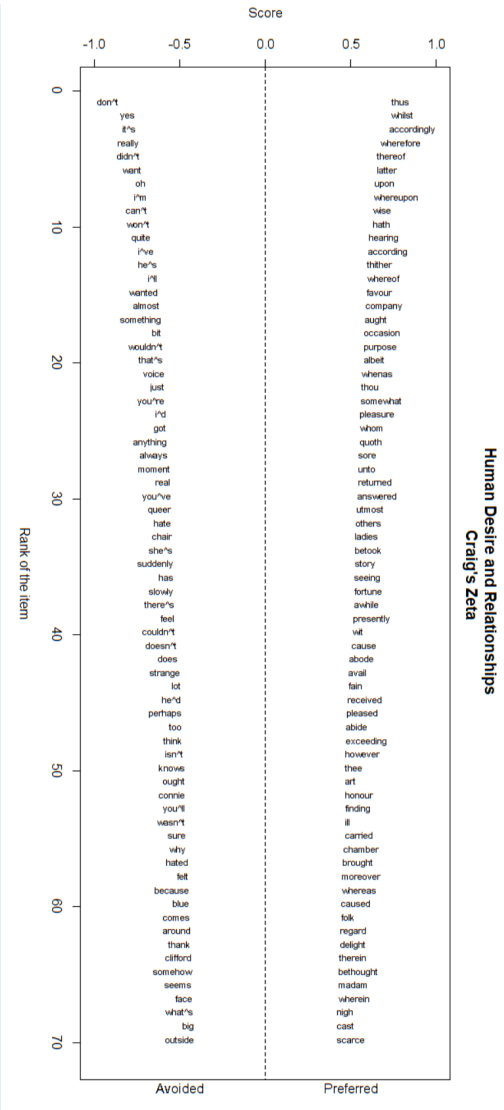
Cultural evolution is evident in banned books, which reflect societal shifts from abstract, philosophical discussions in the 19th century to more personal and materialistic themes in the 20th century. Censorship motivations also evolved: in the 19th century, it targeted books addressing social critique, political ideology, or intellectual challenges, reflected in the formal language used. By the 20th century, censorship focused on books that addressed modern life, controversial topics like race, gender, and sexuality, or cultural taboos. Stylistically, 19th-century texts were formal and historical, while 20th-century works adopted a conversational style with modern settings. This shift reflects a societal change, with 19th-century books aiming to prevent intellectual dissent, and 20th-century books challenging societal norms and taboos. The criteria for banning books also evolved, moving from moral or ideological concerns to challenges against contemporary societal norms.

Division of Banned Books by Century and Theme: 19th vs. 20th Century

For the purpose of topic modeling and identifying common patterns, the corpus was divided into the following thematic categories: Human Desire and Relationships, Social Critique and Class Dynamics, Religion and Morality, Slavery, Racism, and Power, Identity and Growth, Mental Health and Existential Struggles, and Nature, Science, and Philosophy. The comparison of these themes will be made across the two centuries, with the 19th century as the primary focus and the 20th century as the secondary. This approach allows for an analysis of how these themes evolved between the two centuries.

Analysis

THEME: **HUMAN DESIRE AND RELATIONSHIPS**



Preferred words

Words with positive scores appeared more frequently in 19th-century banned books than in 20th-century ones. Examples include “thus”, “whilst”, “accordingly”, “therefore”, and “whereof”. These words suggest a more formal, archaic tone typical of 19th-century literature. The themes in 19th-century books often focused on moral teachings, religious references, and structured societal norms, which could lead to their banning when these themes were seen as outdated or overly didactic.

Avoided Words

Words with negative scores were more common in 20th-century banned books and less frequent in 19th-century ones. Examples include “don’t”, “yes”, “really”, “didn’t”, and “wanted”. These words reflect a shift toward more conversational, modern, and direct language in 20th-century literature. The focus often moved to realism, personal expression, and controversial social topics like race, sexuality, and war, which may have led to their banning for challenging societal norms.

Distinct Patterns in Common Themes

19th-Century Themes: The themes in 19th-century books often focused on morality, religion, and propriety. The use of formal, structured language reflected the literary style of the time. Books were sometimes banned for challenging religious dogma or promoting ideas that went against conservative social norms.

20th-Century Themes: In the 20th century, themes explored individuality, rebellion, and taboo topics. The language became more conversational and modern, reflecting a focus on relatable, contemporary issues. Books were frequently banned for addressing controversial topics such as sexuality, race relations, and political ideologies.

Shared Patterns Leading to Bans

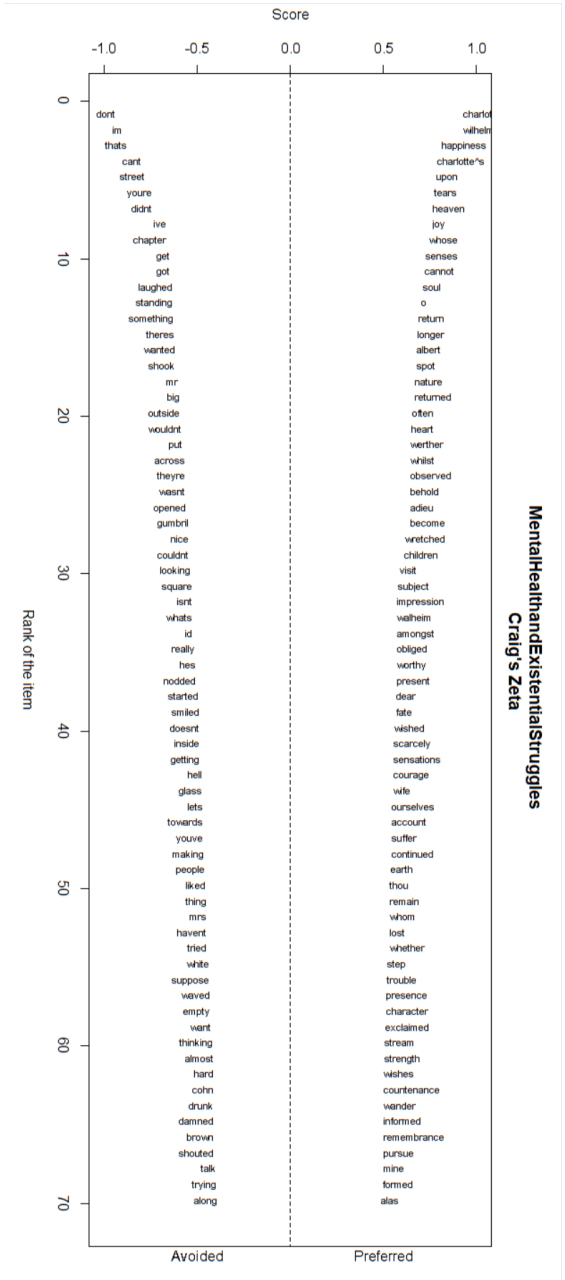
Despite the stylistic differences, both centuries saw bans for challenging established norms. Books that questioned authority, religion, or societal expectations were controversial in both periods. The Zeta analysis shows that while the language of banned books evolved from formal to conversational, the reasons for banning remained the same: questioning authority and societal norms, challenging views on

relationships and human desires, and addressing sensitive or taboo topics that society wasn't ready to confront.

THEME: **IDENTITY AND GROWTH**

The theme of "Identity and Growth" was exclusively represented in 20th-century banned books, with no such books being banned in the 19th century. This absence of books on this particular theme in the earlier period suggests that issues related to personal identity and individual growth were not viewed as significant concerns by those responsible for censorship at the time. As a result, the absence of this theme in 19th-century banned literature indicates that societal preoccupations with these topics had not yet emerged or gained prominence. Consequently, a Zeta analysis cannot be applied to compare this theme across both centuries, as it appears that the focus on identity and personal development became a critical issue for censorship only in the 20th century. This shift reflects a broader societal change in the concerns and values of the time, as issues of personal identity and individual growth began to be more widely debated and scrutinized.

THEME: **MENTAL HEALTH AND EXISTENTIAL STRUGGLES**



Preferred Words

Words like “happiness”, “heaven”, “joy”, “soul”, “upon”, “return”, “fate” and “thy” are more common in 19th-century banned books. These words reflect a focus on philosophical, spiritual, and existential themes, often related to morality, religion, and human purpose. The formal and poetic language, such as “thou”, “shall” and “observed” is typical of the 19th-century literary style.

Avoided Words

Words like “don't”, "I'm”, “that's”, “you're”, “didn't”, “get” “laughed” and “shook” are more common in 20th-century banned books. These words reflect a conversational and casual tone, indicating more personal and relatable stories. The use of direct speech and informal phrases suggests a shift toward intimate storytelling, addressing individual struggles in a more immediate and less abstract way.

Distinct Patterns in Common Themes

The 19th-century banned books lean towards formal, reflective, and poetic expressions of mental health and existential struggles.

The 20th-century banned books emphasize realism, contemporary life, and raw depictions of personal struggles, using modern and accessible language.

Shared Patterns Leading to Bans

19th-Century Patterns: Books from the 19th century often focused on religious and moral concerns, with words like “soul”, “heaven”, “fate”, and “happiness” reflecting a deep engagement with religious and philosophical themes. These works might have been banned for challenging dominant religious ideologies or questioning the nature of morality and existence. Additionally, the formal, abstract language of the time indicates that these books could provoke intellectual debate or challenge social norms, making them targets for censorship.

20th-Century Patterns: In the 20th century, there was a shift toward realism and directness, with the use of modern, colloquial language such as “don't”, “I'm”, and “you're” reflecting a focus on mental health and existential issues in a more personal, grounded manner. These books often tackled social taboos, addressing personal identity, emotional struggles, and societal critique, which could lead to censorship for confronting topics like gender roles, mental illness stigma, and existential crises. The conversational style also made these books more accessible to younger audiences, raising concerns about their potential influence on youth.

Both sets of books appear to have been banned for tackling topics that challenge societal norms, whether through abstract philosophical concepts in the 19th century or direct portrayals of mental health and existential struggles in the 20th century.

Findings

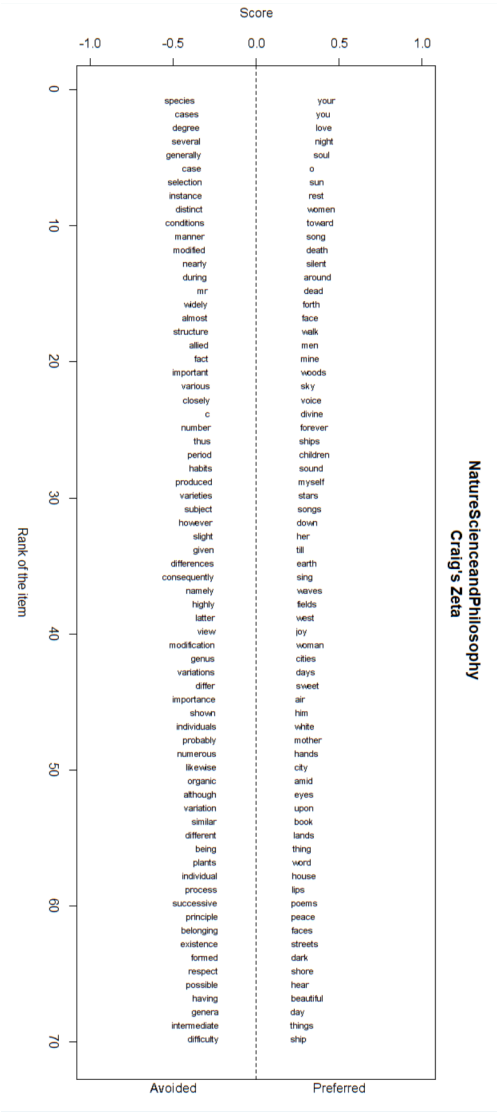
Censorship as a Reflection of Society: The language and themes of banned books reveal a consistent societal discomfort with existential questions and mental health issues, spanning across centuries. However, the reasons for censorship and the methods of addressing these themes have evolved over time.

Shift in Censorship Targets: While 19th-century bans were largely driven by ideological and moral opposition, 20th-century bans appear to focus more on the perceived social and cultural impact of these works.

Historical and Cultural Contexts: These patterns underscore the shifting tension between literature that challenges societal norms and the authorities or groups that seek to suppress such challenges, reflecting the changing cultural and historical contexts in which these books were banned.

THEME: **NATURE, SCIENCE AND PHILOSOPHY**

Since *The Origin of Species* is present in both the 19th and 20th centuries, many words will be similar, especially when discussing themes related to nature, science, and philosophy.



Preferred Words

Words like “love”, “soul”, “rest”, “toward”, “woods” and “song” are more characteristic of 19th-century banned books. These words reflect a romanticized and philosophical approach to nature and science, often expressed in poetic or literary language. The themes in these books emphasize individual contemplation, spirituality, and an emotional connection to the natural world, which are key elements of Romanticism and transcendentalist influences of the time.

Avoided Words

Words like “species”, “selection”, “variation”, “individual”, “process” and “modification” are common in 20th-century banned books. These terms highlight a scientific and analytical perspective, focusing on biology, evolution, and empirical

systems of thought. The language reflects the influence of Darwinian theory and the broader scientific advances of the late 19th and 20th centuries.

Distinct Patterns in Common Themes

19th Century: Dominated by poetic and spiritual language with philosophical underpinnings. The preferred words reflect nature as a medium for emotional and existential exploration.

20th Century: Focuses on technical, scientific, and material descriptions of nature and its processes. Words like “genus”, “modification”, and “individuals” highlight a more objective, research-driven approach.

Shared Patterns Leading to Bans

19th-century banned books often reflected religious and philosophical conflicts, with language suggesting works that merged philosophy and nature in ways that may have challenged traditional religious views, such as questioning divine creation or promoting pantheistic ideas. Additionally, the romanticized and transcendentalist themes in these books likely clashed with the growing industrialization and rigid moral codes of the time, creating tensions between idealized visions of nature and the more orthodox societal norms.

20th-century banned books often reflected scientific materialism, with words aligning closely with scientific ideas about evolution, genetics, and natural history, which were likely controversial to religious or conservative groups. Terms like “species”, “selection” and “process” directly reference Darwinian concepts from *On the Origin of Species*, which faced opposition in both centuries for challenging creationist doctrines.

***On the Origin of Species* as a Bridge**

This book's presence in both sets reflects its consistent challenge to prevailing religious and societal norms across both centuries. In the 19th century, Darwin’s work was banned for its initial shock to creationist beliefs and its philosophical implications. In the 20th century, it became a focal point in cultural wars over science education and the broader acceptance of evolutionary biology.

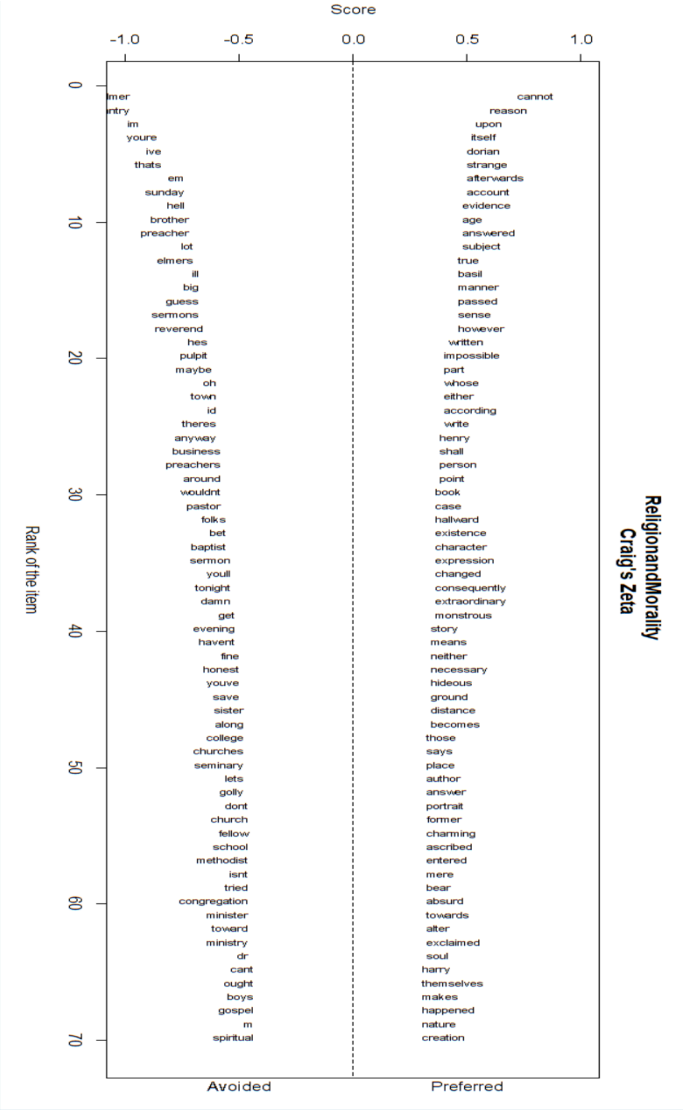
Common Themes Across Centuries

Banned books in both centuries often explored the conflict between science and religion, with 19th-century works framing this tension philosophically, while 20th-century books focused more on practical and educational concerns. These books also challenged authority, questioning dominant narratives about humanity's place in the world and threatening established ideologies. Additionally, there was a shift in language, from poetic and spiritual expressions in the 19th century to more scientific and technical language in the 20th, reflecting changing approaches to nature, science, and philosophy.

Findings

The banning of books focused on nature, science, and philosophy shows society's discomfort with ideas that challenge established norms, whether through poetic exploration in the 19th century or empirical science in the 20th century. Works like *On the Origin of Species* continue to provoke debate, showing how certain ideas can spark controversy across different times and cultures.

THEME: **RELIGION AND MORALITY**



Preferred Words

Words like “reason”, “evidence”, “character”, “necessary” and “existence” are associated with logical thinking, critical examination, and abstract concepts. These words appear more frequently in 19th-century banned books, reflecting themes of intellectual inquiry, challenges to religious beliefs, and progressive moral ideas that conflicted with the dominant societal values of the time.

Avoided Words

Words like “pastor”, “sermons”, “church”, “baptist” and “seminary” are associated with organized religion and traditional moral authority. These terms appear less frequently in 19th-century banned books but are more common in 20th-century

banned books. This shift may indicate a growing focus on critiquing or challenging religious institutions in the later period.

Shared Patterns Leading to Bans

19th Century: The preference for intellectual and philosophical terms in banned books from this era suggests that these works dealt with ideas that contradicted dominant religious beliefs, such as the conflict between reason and faith. Themes related to free thought, rationality, and questioning authority—whether religious or political—were likely targeted for censorship.

20th Century: In the 20th century, the shift toward words tied to institutional religion implies that banned books may have directly critiqued religious practices or figures. Terms like “seminary”, “pastor”, and “sermon” suggest works that challenged or ridiculed the personal and public roles of clergy and the institution of the church.

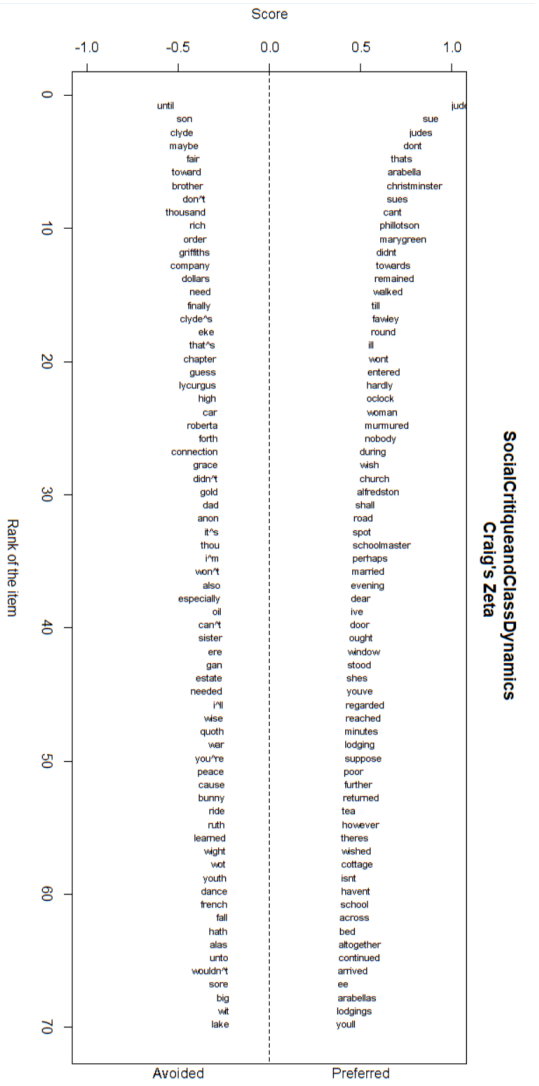
Findings

The Zeta Analysis highlights how the themes of banned books shifted from challenging abstract religious and moral ideas in the 19th century to critiquing specific religious institutions and practices in the 20th century. This progression mirrors broader societal trends: from philosophical debates on morality to addressing social power structures tied to religion. Both sets of bans reflect cultural anxieties over maintaining control of moral and ideological narratives during their respective periods.

THEME: **SLAVERY, RACISM AND POWER**

The theme of slavery, racism, and power is primarily present in 19th-century banned books, with little to no mention in the 20th century. This absence in the 20th century can be interpreted in two ways: either that slavery was no longer a legal institution by that time, or that its impact had been minimized or was no longer a central issue in the literary works that faced censorship. This shift suggests that by the 20th century, the focus had moved away from the direct discussions of slavery, possibly due to the abolition of slavery and changing societal attitudes towards race and power dynamics.

THEME: **SOCIAL CRITIQUE AND DYNAMICS**



Preferred Words

Words like “Jude”, “Sue”, “Arabella”, “church”, “woman” and references to specific characters like “Fawley” suggest a focus on individual struggles against societal norms and religious constraints, as well as challenges related to class and gender. 19th-century banned books often criticized rigid social structures, questioned traditional gender roles, and portrayed characters navigating oppressive systems. These works highlighted personal conflicts with the social expectations of the time, shedding light on issues such as class inequality and the limitations imposed by gender and religion.

Avoided Words

Words like “company”, “dollars”, “firm”, “order” and “car” are associated with economic structures, corporate influence, and modernization. This suggests that 20th-century banned books shifted focus to critiquing capitalism, industrialization, and class inequality, highlighting the systemic and institutionalized nature of class dynamics.

Shared Patterns Leading to Bans

19th Century (Preferred Themes): Terms like “Jude”, “Arabella”, and “woman” suggest that 19th-century banned books often focused on personal struggles with societal expectations, particularly around religion, marriage, and morality. Words like “church” indicate that these books were banned for critiquing the role of religion in perpetuating class and gender oppression. Additionally, many of these books addressed sensitive issues like premarital relationships, class mobility, and challenges to traditional gender roles, which were highly controversial at the time.

20th Century (Avoided Themes): In contrast, 20th-century banned books shifted focus toward critiquing class systems and capitalism, with terms like “dollars”, “firm”, and “company” highlighting the issues of economic inequalities, corporate greed, and capitalist structures. Words like “car” and “order” reflect the rise of industrialization and its impact on societal dynamics, often portraying the alienation of individuals in a capitalist world. These books may have also extended their criticism to larger power systems, including corporations, governments, and social elites.

Shifts in Social Critique

19th Century: The focus of 19th-century banned books is primarily on personal and moral dilemmas, often tied to class and gender roles. These books reflect concerns about characters who challenged religious authority or societal norms, suggesting an anxiety over maintaining traditional hierarchies. Themes like premarital relationships, class mobility, and the questioning of gender expectations highlight the tension between individual desires and the established societal order.

20th Century: In contrast, the critique in 20th-century banned books evolves into a more systemic perspective, focusing on economic inequality and issues like corporate exploitation. These works reflect growing concerns about class struggles, labor rights, and the inequities within industrial society. The shift indicates a move from personal moral dilemmas to a broader critique of societal structures and the power dynamics that perpetuate inequality.

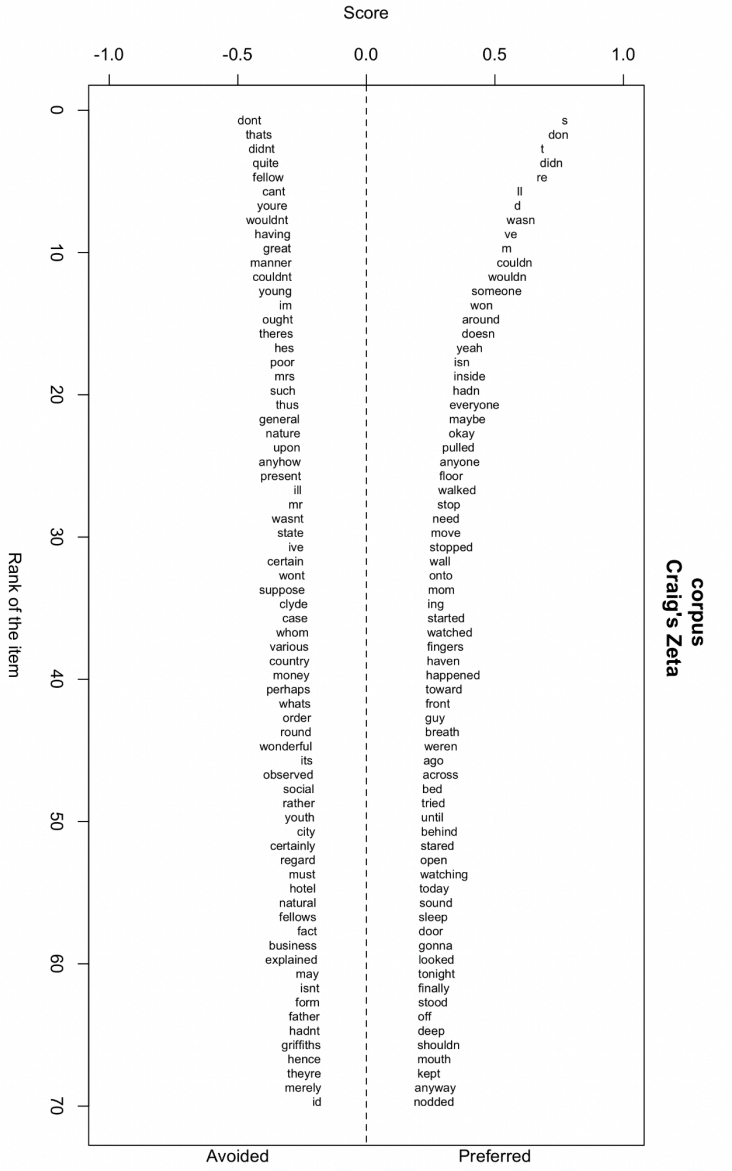
Findings

The Zeta Analysis demonstrates a clear shift in themes from the individual moral struggles of the 19th century to the systemic critiques of economic and class structures in the 20th century. Both sets of banned books reflect societal discomfort with challenges to dominant power structures, whether religious, moral, or economic. This progression mirrors broader societal changes, including the rise of industrial capitalism and the increasing visibility of class-based inequalities.

Division Based on the Century the Books Were Banned: Early 19th century vs. 1950 - present day

The following analysis aims to investigate themes, patterns, and societal issues that contributed to book censorship between 1900 and present day. The dataset is divided into a primary set, comprising books banned between 1950 and present day, and a secondary set including books banned in the first half of the 20th century,. The analysis is intended in continuity with the previous Craig's Zeta test, building up on the results and expanding them to have a comprehensive overview of the history of book bans until present day. As the secondary set comprises an almost 75 year period, which bridges between two millennia and has undergone significant developments, a second analysis will follow, targeting books banned in the second half of the 20th century versus the 21st century, to gain specific contemporary insights.

The analysis that yielded the following results was performed setting the occurrence threshold at 2.



The graph displays, on the x-axis, words preferred or avoided by the primary set (1950s to present day) compared to the secondary set (early 20th century).
A preliminary examination offers interesting insights.

Both lists present negative forms and contractions (“don't," "didn't," "wasn’t,", “couldn’t”, “that’s”, “you’re”), suggesting, at first glance, a more informal language. However, the primary set features colloquial forms such as “yeah” or “gonna”, absent in the secondary set, which in turn sets a more formal tone with terms like “ought”, “upon”, “thus”, “hence”. Moreover, the early 20th century set includes words like

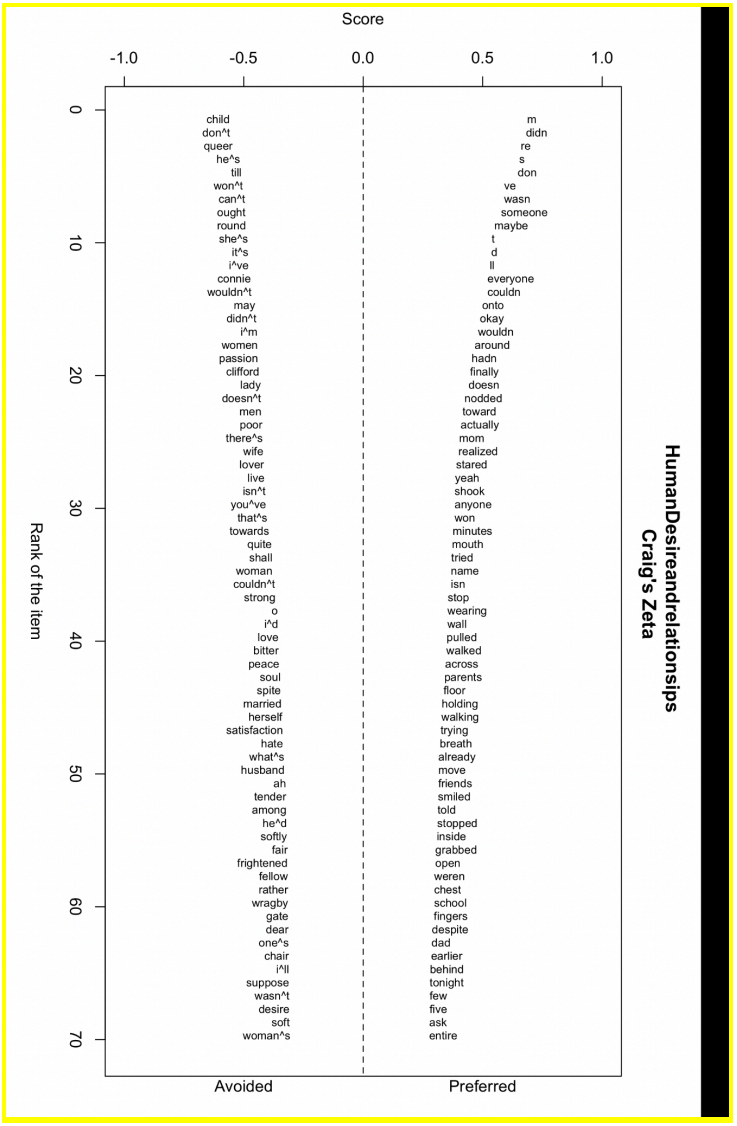
"hotel," "city,", “business”, and “money”, suggesting a focus on broader societal and economic themes. Additionally, the primary set shows a strong preference for first-person and immediate narrative styles. Words like "pulled," "walked," "watched," "stared," and "nodded" indicate a focus on direct action and personal experience, which is still present in the secondary set through verbal forms such as “I’m” or “I’’d” . However, the latter presents more detached, observational language ("observed," "explained," "present").

Findings

Although early 20th century books already employed an informal style and modern settings, departing from the formality of the 19th century, the lists suggest a shift towards even more conversational, accessible writing styles in banned books over time. Personal stories and experiences are recounted in books of both time periods, with early 20th century’s books challenging social norms for the first time. Nevertheless, banned books from the 1950s to present day reflect the aforementioned direct action in the narratives, including personal and intimate stories that aim at further subverting taboos.

Here follows a more in depth analysis of the corpus by themes.

THEME: HUMAN DESIRE AND RELATIONSHIPS



Preferred words: the list, featuring terms appearing more frequently in 1950s to present day banned books, does not seem to include terms overtly connected to the domain of passion and love. Oppositely, it features familiar names, such as “mom”, “parents”, “dad”, suggesting the importance of family dynamics in the topics treated in the books. Nevertheless, body parts are mentioned (“chest”, “fingers”, “mouth”), as well as action verbs (“stopped”, “grabbed”, “stared”), which could imply the presence of directly described physical scenes.

The tone is informal, including terms such as “yeah”, “okay”, and contractions.

Avoided words: the avoided list includes many traditional terms related to relationships and human desire, including “wife”, “husband”, “love”, “passion”, and “desire”, as well as ones expressing emotions, such as “tender” or “satisfaction”. The tone is still quite formal, compared to the preferred list, with terms such as “ought”, “rather”, “shall”, yet contractions are used, suggesting a shift from the pompous formality of the previous century.

Distinct patterns

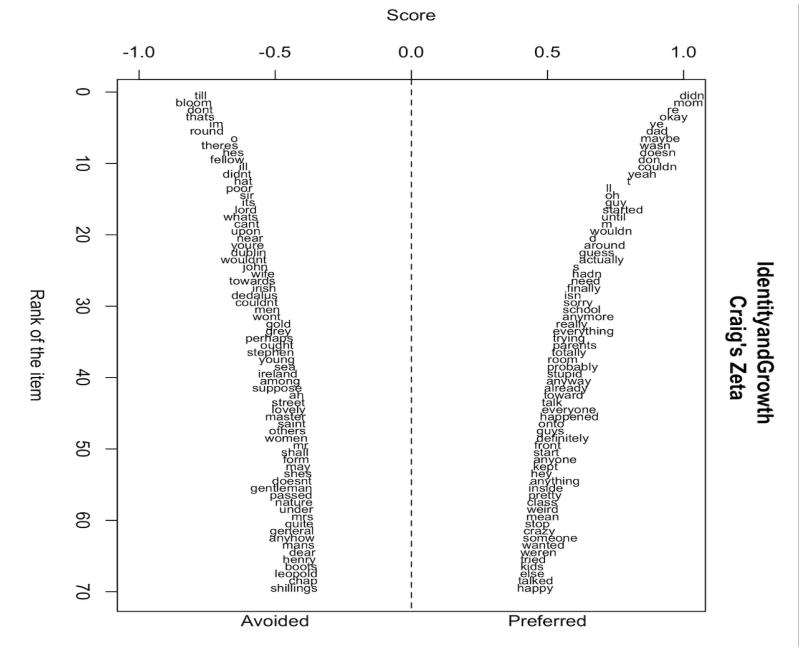
While both lists present familiar bonds of some sort, early 20th century books focus on traditional marriage roles, adding insights into the passionate side of love, which was definitely a break from a more puritan tradition. Late 20th century to present day books include vertical relationships, such as those between parents and children: this emphasis might derive from the high percentage of banned “young adult, coming-of-age” books, which treat “risky” topics along with the struggles of teenage years.

Shared patterns leading to bans

The books banned in both considered periods of time appear to progressively shift to more and more informal tones, possibly to appeal and resonate to broader audiences. This hypothesis is corroborated by the apparent evolution from tackling love relationships with words such as “passion” and “desire”, to directly depicting physical scenes. The earlier period (1900-1950) was characterized by rigid societal norms regarding love and relationships, yet authors often found ways to challenge these norms through their narratives.

Both clusters of books, therefore, seem to challenge contemporary taboos, explaining the reason for bans.

THEME: IDENTITY AND GROWTH



Preferred words: the preferred list still features informal and casual language like "mom," "dad", "guys", and "okay", indicating a trend toward a more relatable narrative style that resonates with modern readers. The presence of words like "myself", "inside", and "trying" suggests that contemporary literature emphasizes personal experiences and self-discovery. The list also includes terms like "kids", "school" and a rose of adjectives evoking different emotions: "happy", "stupid", "crazy", "weird", "sorry". These characteristics highlight themes related to youth and personal development, indicating that more recently banned books often explore formative experiences that shape identity. Finally, the preferred words suggest a focus on interpersonal relationships ("friends," "talking," "everyone") as integral to identity formation. This reflects a shift toward understanding identity as relational rather than solely individualistic.

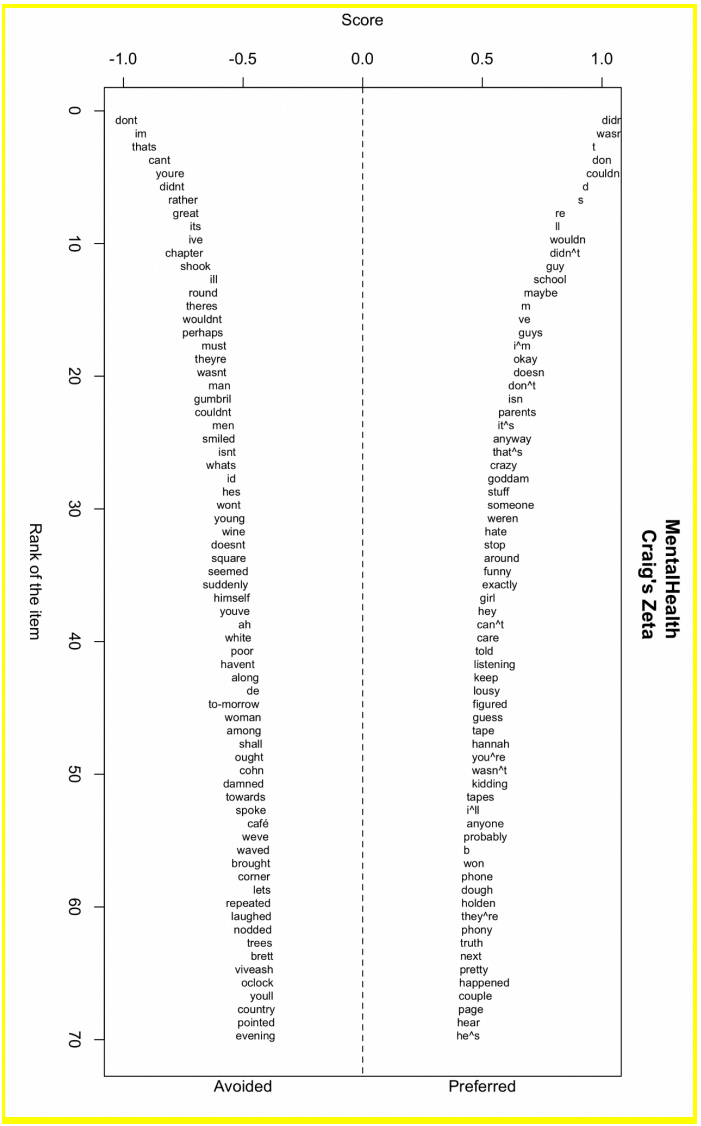
Avoided words: The avoided list includes formal terms such as "lord," "gentleman," and "sir," which reflect a more traditional view of identity and societal roles. The first person "I'm", "I'll", nevertheless, hints at an intimate storytelling, addressing

individual struggles, which, along with informal language and contractions, is likely supposed to appear as relatable to readers.

Findings

In the early 20th century, the mere fact of including the notion of identity and shifting the perspective onto individuality was enough to be considered challenging a social norm, resulting in immediate bans. Earlier works often adhered to traditional notions of identity tied to societal roles. In the following decades, with the concept of identity becoming more and more broad, and bringing perspectives of young people for other young people into the stories, the boundaries of social norms continued to be pushed, following bans especially in more conservative States. In fact, many of the texts featuring vocabulary addressed to young people, are part of the ensemble of books of very recent publication banned from school curricula.

MENTAL HEALTH AND EXISTENTIAL STRUGGLES



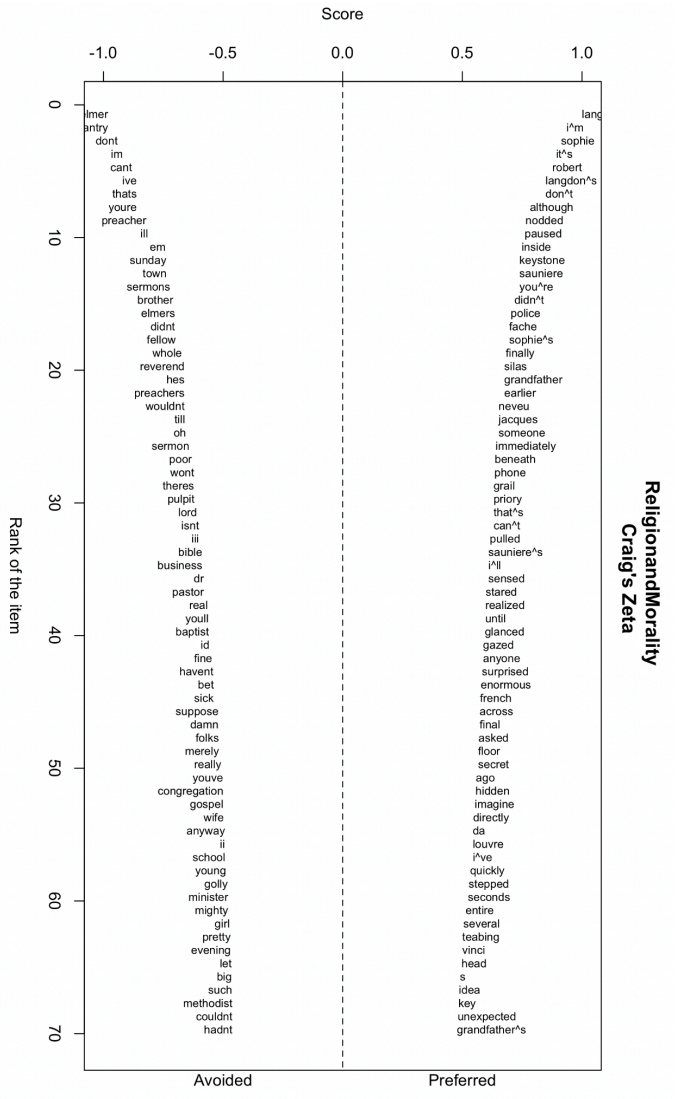
Preferred words: along with the usual casual language, the preferred list includes negatively charged words such as “hate”, “stop”, “goddam”, “lousy”, “crazy”, “phony”, suggesting narratives that openly account for negative experiences and existential struggles in their plots, pursuing a quest for authenticity. “Phone”, “tapes” refer to a modern era of technology that can be a double ended sword, damaging individuals, especially young people, when used to cause harm.

Avoided words: the typical contracted forms are present, as well as first person forms “I’m”, “I’d”, indicating an informal style that favors an intimate storytelling, shifting to addressing individual struggles in a more immediate and less abstract way.

Findings

Both clusters of books seem to have faced bans for addressing topics that confront societal norms, especially by directly exposing mental health issues and existential struggles. In fact, the perpetual bans on books addressing these topics demonstrate how society persistently reveals a deep discomfort with them. This is especially true when the texts delve into negative emotions and are potentially directed at young, impressionable readers due to their accessible linguistic style.

RELIGION AND MORALITY



Preferred words: texts banned from 1950s onwards do not seem to tackle religious or moral themes greatly, nevertheless, the word “grail” can be observed in the list. In fact, while more recent books rarely appear to challenge religion per se, are more inclined to use religious terms indiscriminately, to add a layer or mystery to the plot.

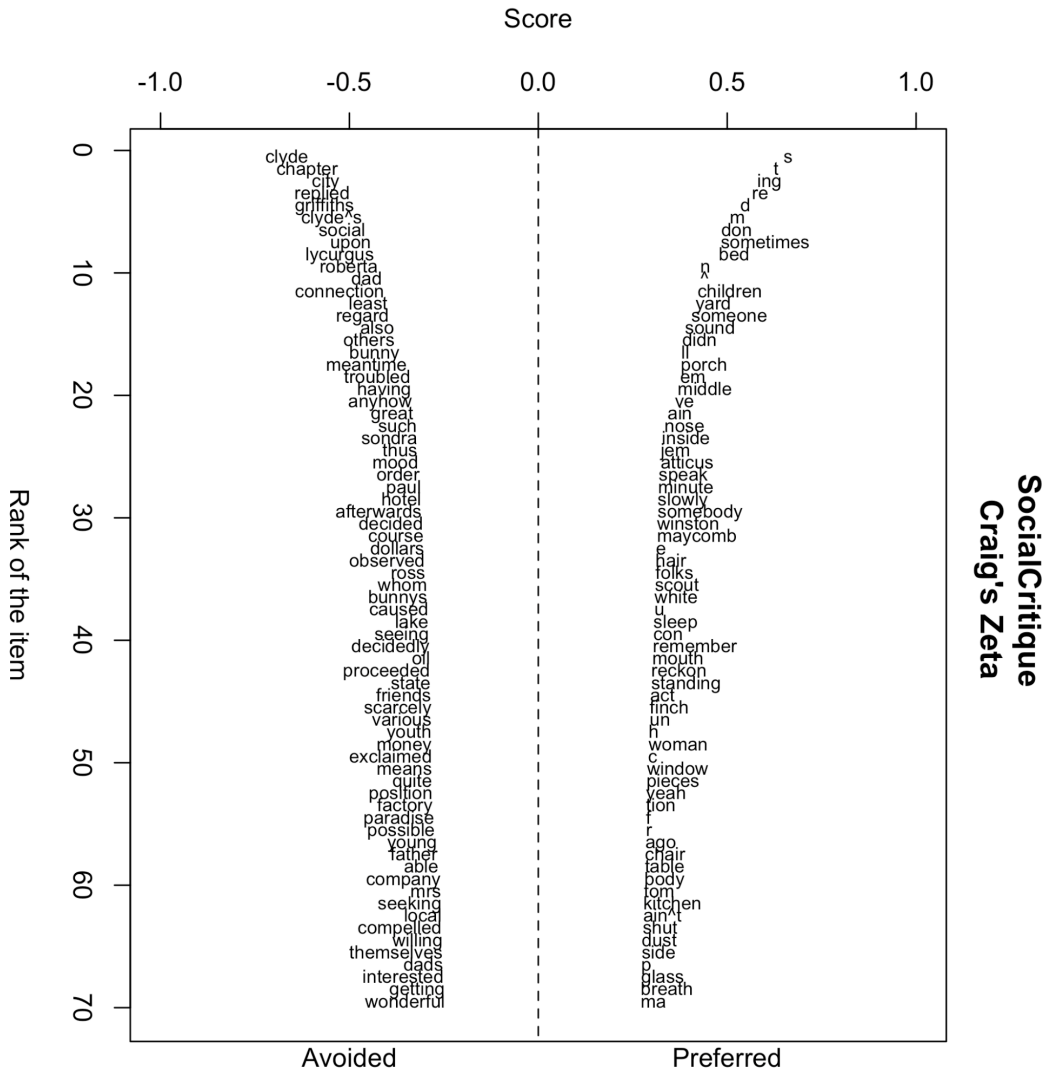
Avoided words: The avoided list includes words directly associated with religious practices (“sunday”, “gospel”, “sermon”) and figures (“pastor”, “reverend”, “congregation”, “minister”), as well as “bible”. The connections are very strong, suggesting novels challenging the institutions directly.

Findings

Early 20th century books faced bans for openly addressing and criticizing religious institutions, challenging the social norms that deemed them untouchable.
More recent books apparently underwent censorship for incorporating religious terms into profane plots, disregarding the sacrality and disrespecting beliefs, being labeled as anti-Christian. The progression between the two periods highlights a more and more challenged vision of religion, which has shifted from an entity to criticize to a powerless institution, not to be overtly feared. The continuous bans, on the contrary, demonstrate its maintained importance and authority as an establishment.

THEMES: SOCIAL CRITIQUE AND CLASS DYNAMICS
SLAVERY, RACISM AND POWER

The topics of slavery and racism are not widely touched upon within books banned in the early 20th century, as mentioned in the previous analysis. Nevertheless, some instances of racism, or violent depictions of black people, reappear in more recently banned books, but within stories aimed at challenging societal norms and exposing uncomfortable truths about systemic injustices.



Preferred words: no specific terms suggest a direct correlation with racism, except for the presence of body parts (“nose”, “hair”), and the term “white”, which could allude to differentiation between races, possibly in derogatory terms. The words “woman” and “kitchen”, if related, could reinforce gender norms and power dynamics.

Avoided words: The presence of words like "money," "dollars," and "factory" in the avoided list suggests that banned early 20th century novels were more likely to directly address issues such as economic disparity and class struggles, with an overt critique of social hierarchies and capitalism.

Findings

In the early 20th century, banned books prominently critiqued class systems and capitalism, highlighting economic inequalities and the alienation stemming from industrialization. This emphasis on class and corporate power dynamics indicated a growing awareness of systemic issues beyond race.

From the 1950s to the present, while explicit discussions of racism have diminished, terms related to body parts and gender roles suggest a continued critique of societal norms and power structures.

These topics are often banned due to concerns that they may promote uncomfortable discussions about race, power, and social norms, which some individuals or groups find threatening or inappropriate.

The themes and issues discussed in the banned books vary from early to late 20th century through to present times, reflecting changes in concerns and taboos. Early 20th-century bans focused on works challenging traditional religious institutions, critiquing economic disparities, and introducing more open discussions of love and passion. Conversely, contemporary bans focus on LGBTQ+ themes, gender identity issues, racial justice topics, explicit depictions of sexuality, and mental health struggles, especially in young adult literature. Despite such divergence, both periods similarly show a trend toward informal writing styles and a challenge to current social norms, with an increased emphasis on personal experiences and individual struggles. The shift in banned literature is indicative of evolving societal anxieties and the tension between established norms and emerging representations of diverse experiences.

Geospatial Visualization with ArcGIS: Mapping the Geographic Distribution of Banned Books

Objective

Geospatial visualization using ArcGIS enables the exploration of regional patterns and historical shifts in censorship practices, highlighting how banned books vary by state and theme over time.

Working with ArcGIS: Data Collection and Map Creation

The following steps outline the process of creating a geospatial visualization of banned books using ArcGIS Pro.

Step 1: Organize the Data

Before creating the map, the data were structured into a CSV file to include essential details about each banned book. The dataset consists of the following columns:

Column Name	Description
State	The state where the book was banned
Title	Title of the banned book
Author	Author's name
Year_Banned	The year(s) the book was banned
Theme	The thematic category of the book
Reason	The reason for banning
Latitude	Latitude coordinate of the state
Longitude	Longitude coordinates of the state

Data Sources: Metadata was gathered from the Jupyter Notebook and supplemented with spatial data (latitude and longitude) sourced from https://www.latlong.net/.

Step 2: Aggregate Data for Visualization

Since multiple books may be banned in a single state, the structure of data was made sure to be effective for visualization in ArcGIS Pro:

- Pop-Ups for Detailed Information: Each state's banned books list will be available as an interactive pop-up, providing detailed book titles, authors, years of banning for each book, theme of each book, and reasons for bans.

Step 3: Importing and Preparing Data in ArcGIS

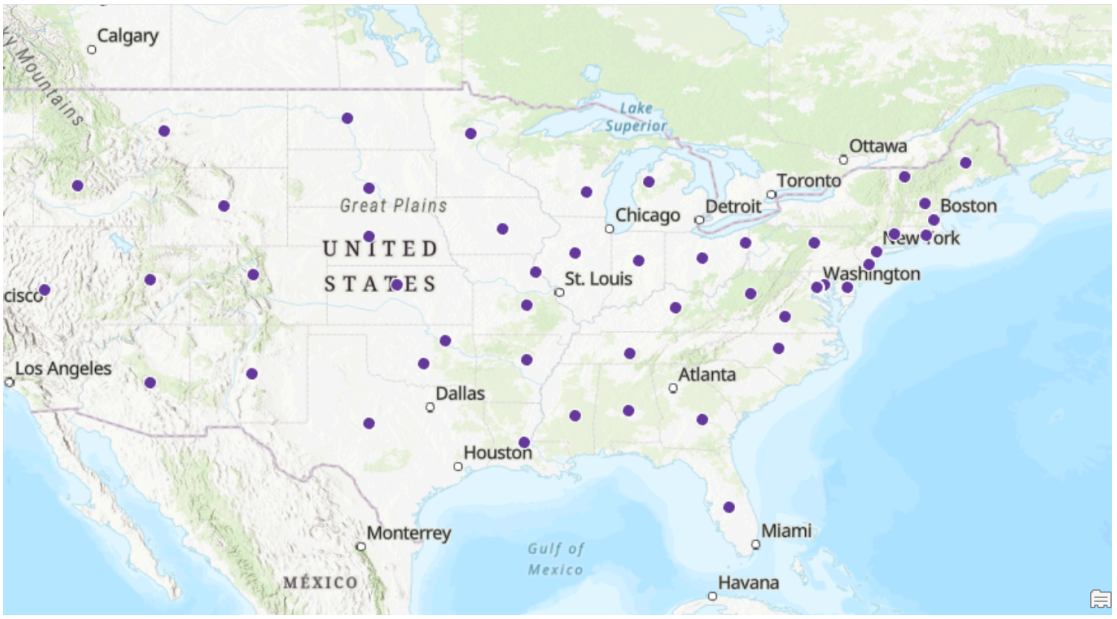
Once the data was structured, these steps were followed in ArcGIS Pro to create the map:

- Import CSV Data: Use the Add Data tool (XY table to point) to load the CSV file containing banned book details.
- Define Spatial Representation: Using the XY Table To Point tool, we mapped the geographic coordinates (latitude and longitude) to generate a point layer for each book's location.
- Apply Coordinate System: We applied the WGS 1984 coordinate system to ensure that the data aligns correctly with the other spatial layers.

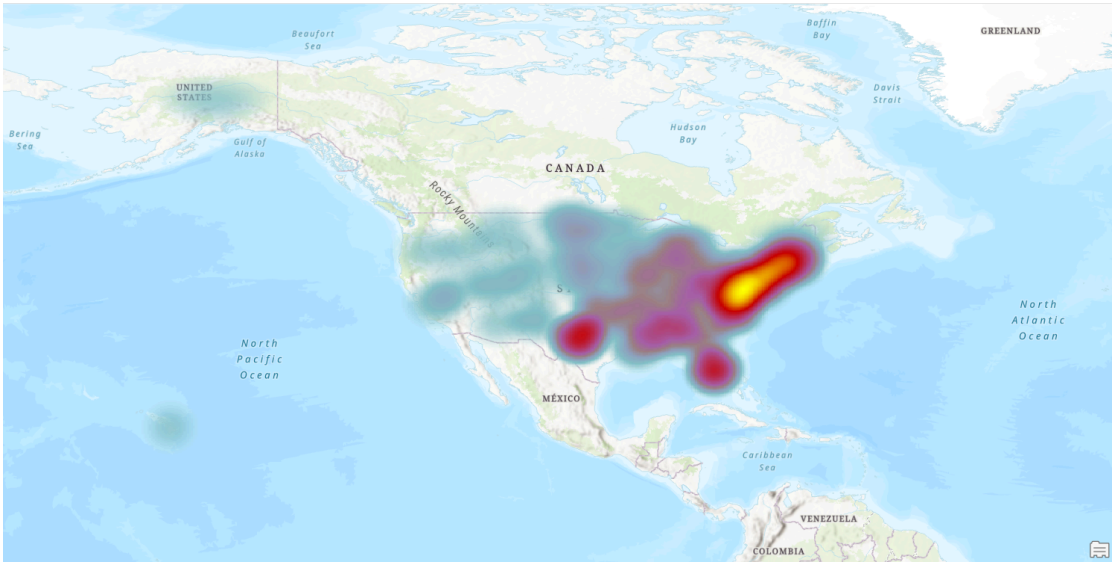
Step 4: Customize the Map

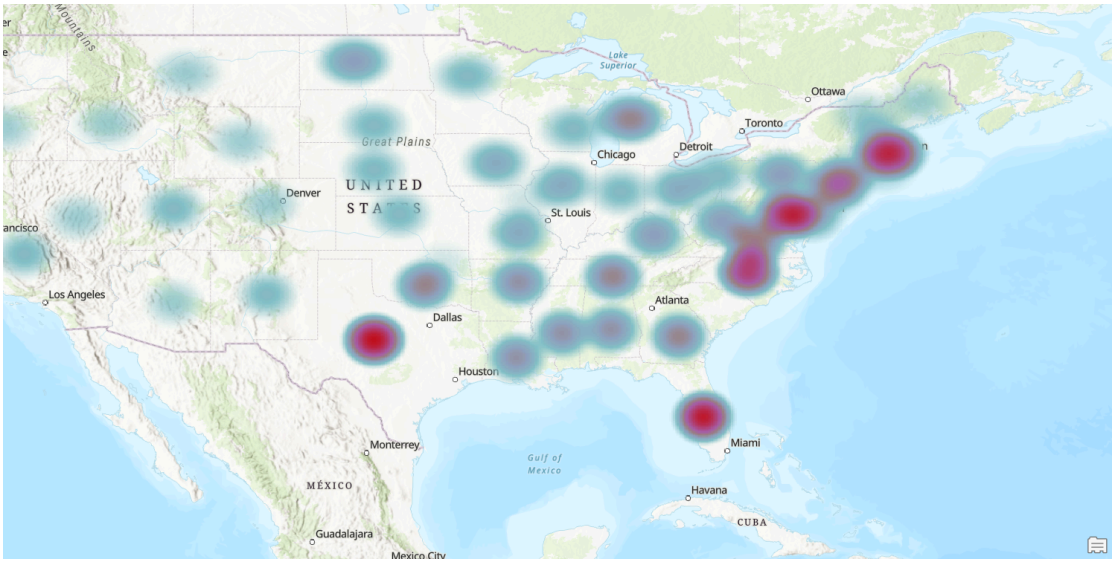
Once the data was prepared, we customized the map to visualize the distribution of banned books. Key elements of the map customization included:

- Single Symbol Visualization: To create a uniform representation of features, we applied a single symbol style to the layer. This approach assigned the same visual appearance to all features, using consistent color, size, and shape. It provided a clean and simple way to display the data without emphasizing any specific attribute, making it ideal for highlighting the overall distribution of features in the study area.



- Heat Map Visualization: To enhance the geographic patterning, we applied a heat map layer. The heat map visually represented areas with a higher concentration of banned books, using color gradients to show intensity. This enabled a clearer understanding of the geographic regions where censorship was most prevalent and provided a quick visual representation of “hot spots”.

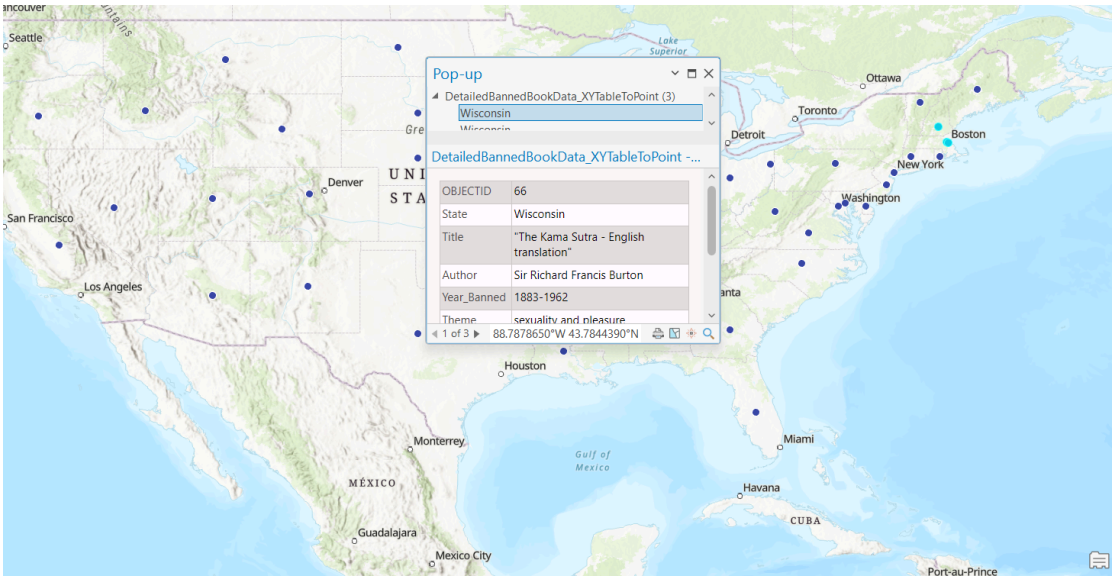




By incorporating both single symbol and a heat map, we were able to highlight both individual book details and regional censorship trends more effectively, enhancing the map's interpretability and impact.

- Pop-Up Configuration: For each state, we configured interactive pop-ups to display detailed information about the banned books within that region. The pop-ups included:
    - Title
    - Author
    - Year Banned
    - Theme

- Reason



Step 5: Sharing or Exporting the Map

Once the map was created, we had several options for sharing or exporting:

We chose the Static Map: For publication, the map was exported as an ArcGIS project file and PDF.

By using ArcGIS, we were able to create a detailed, interactive map of banned books in the U.S. that visualizes both the geographic distribution and thematic patterns of banning.