Frequency Analysis and Topic Modeling of Books Banned in 18 US State Prisons

Introduction

Although there is plenty of discussion about and critique of the structural and social place of prisons, there is still much to analyze about the specifics of prison life, especially from a computational standpoint. Prisons control their population through a variety of methods, one of which is limiting the kinds of print materials inmates have access to through book-banning.

For this project, I am studying a dataset of banned books in 18 US states. I analyze the titles, authors, given reason for banning, and state of banning using frequency visualizations and topic modeling. I seek to discover commonalities in the books being banned and to draw conclusions about what this says about how prisons control inmates. Through this I am employing a number of distant reading techniques, particularly Allison's concept of "reductive reading" (2018).

The data, discussed in more detail in the next section, is a CSV from The Marshall Project, which contains information on banned books from 18 US states. As this is a massive amount of data, with hundreds of thousands of items per state, it is not comprehensible without the application of computational methods to sort, clean, and find commonalities.

Research Material and Methodology

As mentioned, the data for this study comes from The Marshall Project. The Marshall Project is a nonprofit prison reform news organization. The banned books data is compiled by Andrew R. Calderón and David Eads (see references). As they discuss on the portal for the dataset, they received information from 24 states but found the data from six of the states too messy to publish without significant cleaning and restructuring. Thus, the database currently only covers 18 - Arizona, California, Connecticut, Florida, Georgia, Iowa, Illinois, Kansas, Michigan, Montana, North Carolina, New Jersey, Oregon, Rhode Island, South Carolina, Texas, Virginia, and Wisconsin. The data is available in per-state CSVs or as a compiled CSV. The latter is what is employed for this project.

Even with the 18 states that presented usable data, the data we're working with is not the cleanest or most consistent. The CSV is divided into the columns "Publication" (the book title), "Author", "Date", "Year", "Month", "Day", "Reason" (the given reason for the ban), and "State". However, not every state reports all of this information. A number of states do not provide author names or reasons for banning. Even when information is provided, it can be vague or incomplete. As we are interested in studying and drawing conclusions from banned books, we need to keep in mind that our data is fundamentally incomplete even with the eighteen states we are working with.

Another major issue with the data is more philosophical: the information The Marshall Project used to form these CSVs was provided by the prisons. As this is a fairly clear conflict of interest - prisons providing information about themselves to an organization whose goal is to criticize them - it's entirely unknown how true-to-life the data is. We can make a fair assumption that the data is for the most part broadly true, but it's impossible to know how accurate it actually is.

With these issues in mind, I proceeded to work with the data. The data was processed and visualized using Python in a Jupyter Notebook. Please read the comments in the Notebook for specifics on the functioning of the code.

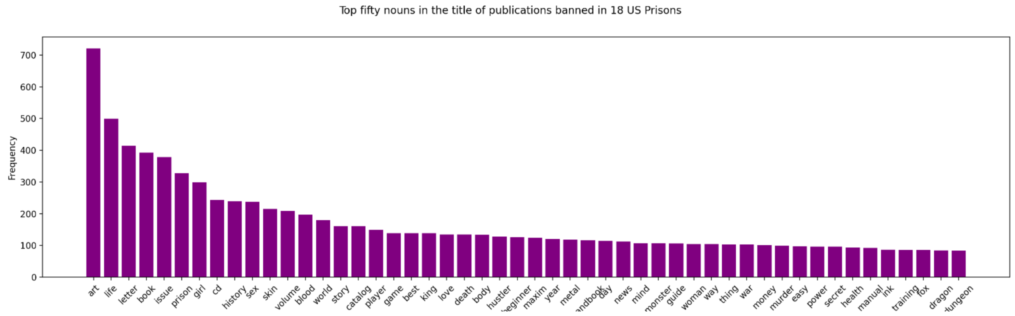
The first decision I made while preprocessing the data was to remove the time-related columns. Although one could perhaps focus on developing a chronological timeline of book banning, I felt this was out of the scope of this project and thus these columns provided no useful information to me.

Next I decided what I would be focusing on for my analysis. I decided to extract the nouns, verbs, and proper nouns from the publication names, the author names, the nouns from the reasons, and the state abbreviation. The rest of the information (stopwords, other parts of speech, non-alphanumeric information) would be left aside, however I did replace all empty cells with the word "None" so I could see how many empty cells there were in the final data.

Finally, I also formed a full list, minus stopwords, of the publication name and reason columns, to be used for topic modeling. Standard stopwords were removed to allow for a focus on "semantically rich" information that would give us more insight about the commonalities amongst the books. This data was output as a series of CSVs as well as visualized in the notebook using the matplotlib and pyLDAvis Python libraries.

Analysis and Visualization

For the purposes of this paper, I will be focusing on a selection of data categories that I felt had the most salient information to a small-scale analysis such as this. These are the publication nouns and publication verbs, the reason nouns, and the states. Author names would require further analysis to pull out usable information, while the publication proper nouns data shows too much noise and would require more manual cleaning and/or stopword modification.


Top fifty nouns in the title of publications banned in 18 US Prisons

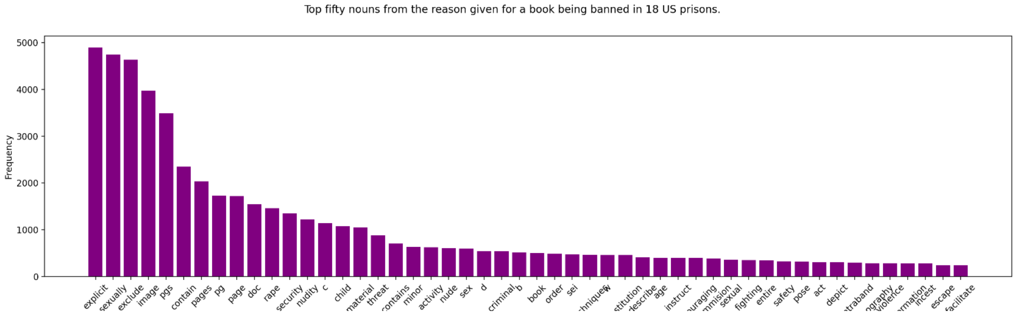
We start with the graph of the fifty most common nouns in the titles of the banned publications. Although we're presented with a good amount of nouns that are general enough we can't say a lot about them out of context (ngram analysis would be a good next step with this data for this reason), we can still begin to see some trends and commonalities among the nature of the banned books.

Perhaps unsurprisingly, there appears to be a strong tendency towards banning books with "explicit" content. Words such as "girl," "sex," "skin," "blood," and so on reflect a banning of material featuring violence and sex. We expect this to be corroborated by the reason data and the topic model.

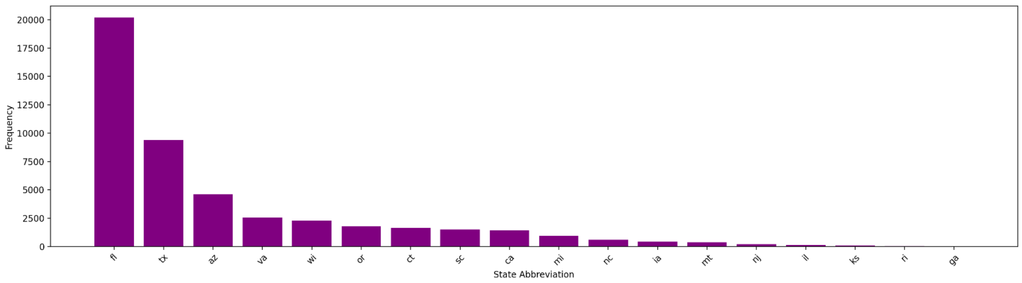
Interestingly, we also see a heavy presence of the words "player" and "game". Consulting the original spreadsheets(important to do here so we don't lose context) shows that some of these books are gambling books. However, we also see by the presence of "dungeon" and "dragon" and "monster" and "manual" together that this ban also includes innocuous manuals for tabletop roleplaying and board games.

Next, we examine the publication verbs. We see a very high preponderance of "illustrate" "color", and "draw" which I found surprising. Returning to the spreadsheets shows that a great number of coloring books are banned, regardless of content.

Elsewhere, the verbs corroborate our conclusions from the nouns. We see a high frequency of words such as "kill" and "die" as well as a high presence of the words "play" and "win".


Top fifty nouns from the reason given for a book being banned in 18 US prisons.

Next we take a look at the reason nouns. As we expect, we see an extremely high preponderance of explicit content, both the word explicit itself and a number of associated words, especially those having to do with sex. We also see the prominence of the word "contraband". This is not a surprise itself, of course, but it is interesting that if we look back at the original data we see that "contraband" is provided without further elaboration when it's listed as the reason. It does not say why something is contraband. Finally we see a few other general terms such as "escape," "instruct," and "safety”.

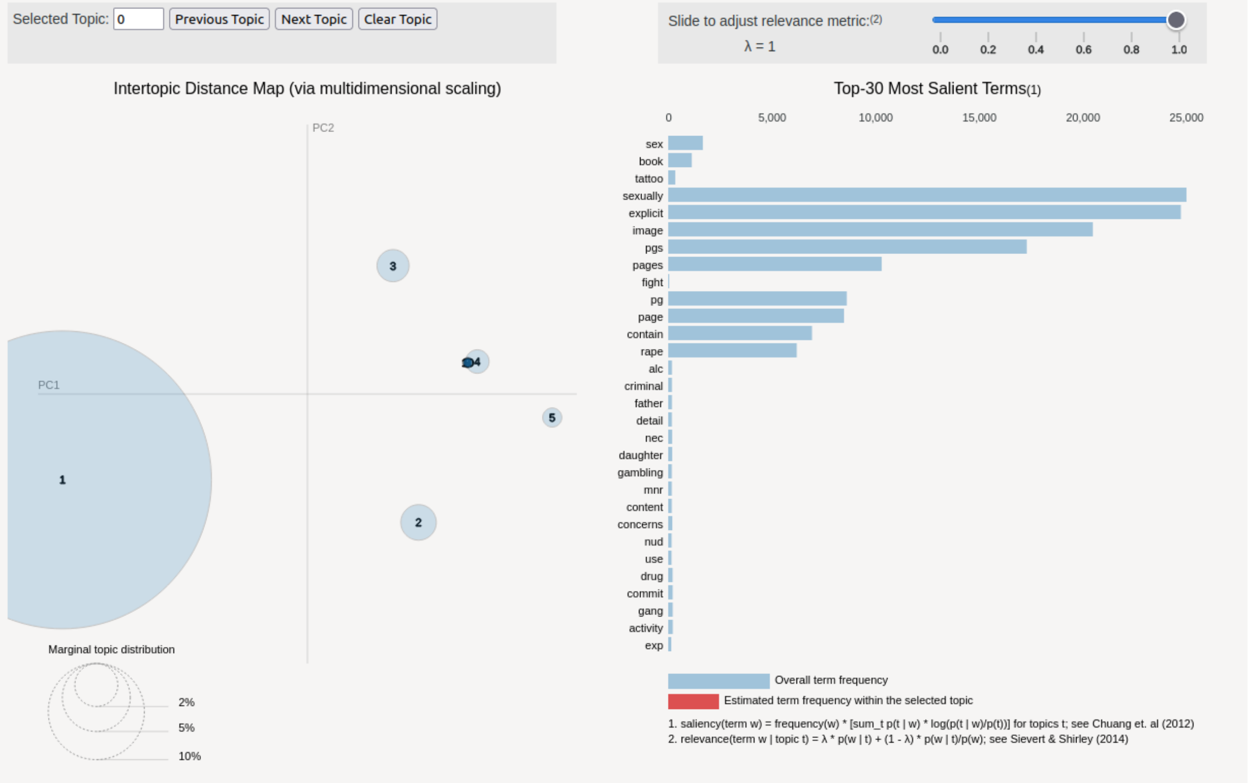

Number of books banned in prison by state according to Marshall Project data.

Lastly, we turn to the graph of states. While the other frequency graphs were indexed to the fifty most frequent items, as there are only eighteen states represented in the data we are able to include them all in the visualization.

In terms of the amount of books banned by each state, Florida takes a substantial lead, with other large amounts shown in Texas and Arizona, after which the graph both steadily decreases and steadily evens out. Georgia is last.

While this information is broadly helpful to get an idea of the dynamics of book banning, it is not highly trustworthy for a deeper analysis. The reason for this is that the statistic we are looking at is the raw numbers of books banned. However, different states have different populations and differently-sized prisons. A further step in visualization to make this information much more useful would be to process it so we are looking at the proportion of books banned to prison population to general population. This would likely say more than the raw numbers would.

The topic modeling, performed on a combined list of nonstop alpha words from the publication and reason columns, corroborates the information we've gathered from the frequency analysis with a larger-scale and more sorted visualization of the data. The largest topic by a huge amount, containing 96% of tokens, has a very high concentration of terms related to sexual explicitness. In topic three, meanwhile, we see a high presence of "book," "war," "role," "play," and "game," suggesting again that books related to gaming are commonly banned. In topic four we see a high concentration of a few terms related to violence.



Selected Topic: 0

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Discussion and Further Analysis

From this analysis, we can make some interpretations. Particularly, what we find backs up much of the current and historical theory about prisons as institutions that restrict humanity (see Foucault, 1977, as well as any number of more recent abolitionist works). They do this in many ways, big and small, and one common way is limiting the sort of materials, information, and narratives prisoners have access to.

Even putting aside illegal or morally questionable material, a vast number of the banned books feature nudity and sexuality or were used in the playing of games or were simply coloring books. Here we see how prisons punish prisoners by restricting their access to human experience. A very classic case of biopolitical oppression (Foucault 1977).

We don't take a data approach to this project with the goal to prove "hard evidence" for previous "soft science" - though data could be useful when pushing for policy change. As Dan (2019) points out, understanding data is itself an interpretive and humanist process, and we need to be careful about treating data as "proof" just because it's data.

As she explains, there is a danger in computational analysis of simply presenting data as the final result of a study without taking the extra step of performing a true humanistic analysis, noting future steps, discussing other literature, and so on. Similarly, Tahmaseb and Hengchen

(2019) urge an approach that fruitfully combines data science and humanities without reducing or prioritizing either. This is why I find it important both to talk about the interpretation of the data itself and to reflect on its relation to other work on the subject and other work about computational humanities in general.

Ultimately, this is simply another approach, one of equal validity to any other (assuming methodological rigorousness). With this approach, we are employing Allison's concept of reductive reading (2018). That is, tactically reducing our text in specific ways that allow us to analyze it and come to new conclusions when relating it back to the original text (Allison, 2018).

I also note this to push back somewhat against Moretti's conception of distant reading as something separate from close reading, where one strategy or the other is employed "because you are convinced that that viewpoint is better." (2000) Rather, I believe these two viewpoints inform each other and co-operate, allowing us to triangulate a richer view of our topic than we would using either by itself. In this, I align with Tahmaseb and Hengchen (2019)'s assertion that "by combining a data-intensive research methodology with traditional and modern humanities, much can be gained."

Also important is the specific nature of what we're analyzing and how. There has been some criticism about computational humanities approaches being a revival of classical structuralism - the idea of trying to force unordered (or unorderable) data into a unified system (see Risam, 2019). However, in the case of a project like this, we are using a systemic analysis approach to analyze an extant system. That is, we are not trying to arrange loosely ordered material into an artificial structure but instead are analyzing a system that already exists in the world - the system of banning books in US prisons.

There are plenty of avenues for further analysis. Simply due to the time constraints of this project I was not able to drill very deeply into the data, and had to stick to pretty broad and general analytics. There are plenty of ways this data could be explored further, with more intricate topic-modeling, named entity recognition, or assembling multiple corpora from different sections of the data. More attention could also be paid to the originating data. It would likely be productive to contact The Marshall Project to discuss the data, any errors or problems they did not state publicly, and so on. It would also be advantageous to spend more time manually going through the data, to familiarize us with any additional ways it needs processing such as more customized stopword filtering.

Through this analysis we have seen how prisons control and punish prisoners through book-banning. Even with relatively simple and broad analysis, we can see the commonalities among banned books and draw some strong conclusions. This is a very generative angle on the discussion around prisons and abolition, allowing us to theorize from multiple directions. However, as discussed above, there are still many ways this analysis could be deepened.

References

Allison, S. (2018), “In Defense of Reading Reductively”, *Reductive Reading: A syntax of Victorian Moralizing*, Baltimore.

Calderón, A. and Eads, D. (2022), *Books Banned In State Prisons.* Available at: https://observablehq.com/@themarshallproject/prison-banned-books.

Da, N. (2019), “The Computational Case Against Computational Literary Studies”, Critical Inquiry, vol. 45:3.

Foucault, M. (1977) *Discipline and Punish: the Birth of the Prison*. 2nd Vintage Books ed. New York: Vintage Books.

Moretti, F. “Conjectures on World Literature,” New Left Review, vol. 40, 2000:1.

Risam, R. (2019). “Introduction,” *New Digital worlds: Postcolonial Digital Humanities in Theory, praxis, and pedagogy*. Evanston, Illinois: Northwestern University Press.

Tahmaseb, N. and Hengchen, S. (2019) “The strengths and pitfalls of large-scale text mining for literary studies”, Samlaren: Journal for Research on Swedish and Other Nordic Literature, vol. 140.