Damini Sharma
Computational Content Analysis Final Paper
Link to GitHub Repo: https://github.com/DSharm/american-prison-writing_nlp
June 12, 2020

Racism, Forgiveness, Love, God, Injustice: Content Analysis of American Prison Writing Archive

America has 5 percent of the world's population and 25 percent of the world's incarcerated population. By the latest figures, the US holds approximately 2.2 million people in prisons and jails. This number doesn't capture the full impact of the criminal justice system – another 11 million people are part of the carceral system on probation or parole, not to mention the families and communities that have been ravaged by mass incarceration.[1]

Additionally, massive disparities exist in who is incarcerated. Poverty is a significant predictor of incarceration, as well as an outcome. Mass incarceration also disproportionately affects Black and Hispanic people and those with mental health issues. While Black Americans make up 13 percent of US residents, they make up 40 percent of the incarcerated population.[2]

Despite the astonishing scale of the problem, very little is known or understood about the experience of currently or formerly incarcerated people. There are some explanations for this widespread ignorance – notoriously limited access to carceral facilities, apathy, dehumanization of those labeled criminals, and active censorship by prison and jail authorities. This ignorance leads to two significant consequences: the average person can ignore and dehumanize a prisoner, and also has very limited understanding of the various problems of incarceration - poor living conditions, inhumane treatment, abuse by guards, wrongful convictions, or discriminatory practices.

A few groups have taken the lead on combating this ignorance. One such initiative is the American Prison Writing Archive (APWA) – an open-source database for currently and formerly incarcerated people and prison staff to document their experience. These essays give the authors a voice and give the readers an insight into the lives of people who are easily ignored, marginalized, or forgotten by society.

In this paper I conduct a content analysis of the APWA. Using counting, classification, structured topic modeling, and word embedding techniques, I attempt to do a systematic exploration of this important and unique corpus. None of these techniques can act as substitutes for thoroughly reading and understanding the often-difficult material contained in the essays – doing so would be a disservice to an already-marginalized population. However, content analysis techniques can be used as a useful first step in exploring and generating hypotheses about the essays – what the major themes are, and how those might differ by attributes of the authors such as sex, race, and location.

The paper is structured as follows. Section I describes the corpus and data collection process. Section II provides summary statistics. Section III gives an overview of the methods used. Section IV discusses the findings and interpretations. Section V discusses limitations of this work and concludes.

[1] Sawyer, Wendy, and Peter Wagner
[2] Sawyer, Wendy, and Peter Wagner

Section I: American Prison Writing Archive and Data Collection Process

The APWA is an initiative of Hamilton College, and evolved from a project completed in 2014 titled the *Fourth City: Essays from the Prisons in America*, a collection of non-fiction essays written by currently incarcerated Americans. After the book was published, the submissions continued, and were eventually turned into the current archive containing over 2,300 essays. The essays are solicited through prison-support newsletters, and the collected essays are then scanned and ingested, or transcribed. The group welcomes anyone with first-hand experience of the US carceral system to submit essays – including those who live, work, and volunteer there.[3] The essays also contain self-reported attributes of the authors, including race, sex, sexuality, and state of imprisonment.

For this project, I scraped 2,098 transcribed essays and author attributes. The scraping process was far more involved than I had originally imagined. Most of the essays are handwritten letters or scanned PDF documents, and the APWA has done a fantastic job of transcribing nearly all the essays. The transcriptions and author details, however, only appear after clicking a button on each essay's page, which poses an additional challenge for scraping techniques.

To address this, I use packages in Python such as *selenium* to create a Chrome webdriver that essentially mimick a user – visiting each essay page, clicking on the button to reveal transcription and author attributes, and then scraping the relevant information. Predictably, the scraping process takes hours, and sometimes the lag in opening up a webdriver scrambled author attribute information. As a work-around, I separately iterate through all the author attributes and construct dictionaries of essay titles and author attributes – since this does not involve using a webdriver, this process concludes in minutes. I then merge the essay content and attributes together by essay title. Some essay titles do appear twice – approximately 100 out of 2,300 titles. In all the duplicate cases I checked, the title pointed to the same, unique essay, but was listed under different attributes, usually if someone self-described as having multiple ethnicities. In these cases, I aggregated across the ethnicities to make the attributes data unique on title before merging with the transcriptions.

Finally, a small minority (17) of the essays are written in part or entirely in Spanish. However, most of them have been translated, and there was no way of identifying those that haven't, so they have not been dropped from the sample. The resulting dataset has 2,098 transcribed essays with race, sex, and state information.

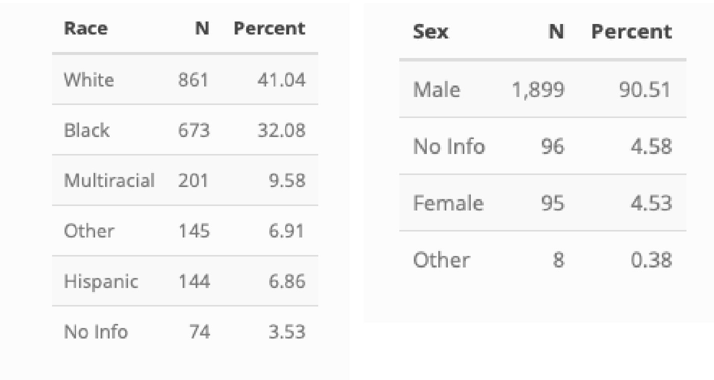
Section II: American Prison Writing Archive and Data Collection Process

Given the relatively small sample size of this corpus, I did not employ any further sampling techniques. However, there is no reason to believe that these essays are necessarily representative of the incarcerated population, since they are openly solicited, and authors self-select into writing them. Looking at some descriptive statistics also helps demonstrate this.

The essays range in page length from 1 to 30 pages, but the majority are between 1 and 10 pages long. The essays are overwhelmingly written by males (90%) and only 4% by women, even though women are a rapidly growing part of the prison population. 41 percent of the essays have White

[3] American Prison Writing Archive

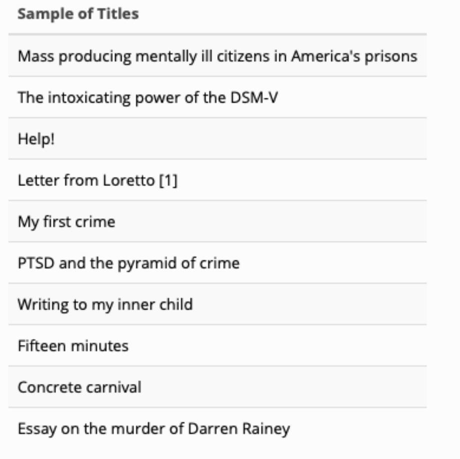
authors, 32 percent have Black authors, and 7 percent have Hispanic authors. The essays cover 47 states, with approximately 44 percent coming from California, Texas, Pennsylvania, New York, and Arizona. Authors can write multiple essays, so the summary statistics should be taken as statistics of the essays themselves, not of the authors overall. Figure 1 summarizes descriptive statistics about the authors.



Race	N	Percent
White	861	41.04
Black	673	32.08
Multiracial	201	9.58
Other	145	6.91
Hispanic	144	6.86
No Info	74	3.53

Sex	N	Percent
Male	1,899	90.51
No Info	96	4.58
Female	95	4.53
Other	8	0.38

Figure 1: Race and sex breakdown of author attributes



Sample of Titles
Mass producing mentally ill citizens in America's prisons
The intoxicating power of the DSM-V
Help!
Letter from Loretto [1]
My first crime
PTSD and the pyramid of crime
Writing to my inner child
Fifteen minutes
Concrete carnival
Essay on the murder of Darren Rainey

Figure 2: Sample of ten essay titles

These statistics highlight that this is not necessarily a representative sample, and the findings should be interpreted as such. While this corpus is unique and important in providing a voice to incarcerated people, the results of this paper should not be generalized to all incarcerated people.

Furthermore, the data on author attributes is very sparse. There are some available demographic categories that I did not collect, such as religion and sexual orientation, because these categories were incredibly messy and thin. For example, the majority of the authors self-reported as "catholic" or "Christian", with the remainder spread across 92 categories. Still, cleaning and using the categories could be an avenue for potential future work.

Figure 2 looks beyond summarizing essay and author demographics to gain some insight about the essays. It displays a random sample of ten essay titles – these are not necessarily representative of the corpus but do show some helpful range in the content of the essays. Some are more personal, others more explicitly discussing or critiquing the criminal justice system.

Section II: Overview of methods

In approaching this corpus, I had two main guiding questions – one was to get an overall understanding of what themes these essays contain, and the second was to see if there were any systematic differences in those themes by author attributes. These differences may or may not reveal differences in how different sexes or racial groups are treated in prison, or even differences in their prison experiences. As Figure 2 above suggests, not all the essays are about actual experiences in the prison – some are general critiques of the system, some are personal narratives of the circumstances that landed one in prison, some are thoughts on current affairs, and so on. These exercises should therefore be seen as an exploration of racial and gender dynamics in this corpus, rather than attempting to make inferences about racial or gender disparities in prison experience generally.

To get a sense of what the corpus contains, I begin by counting the important words in the corpus, visualizing them with a word cloud, and calculating and visualizing differences between the word frequency distributions of the corpora on race, sex, and state dimensions. This is a first pass at seeing if the distribution of words is different across difference demographics. The results from this would not be conclusive for many reasons, including that this method doesn't allow words to have multiple meanings. However, if there was no divergence, there would be reason to think that the gender and race dimensions may not be insightful.

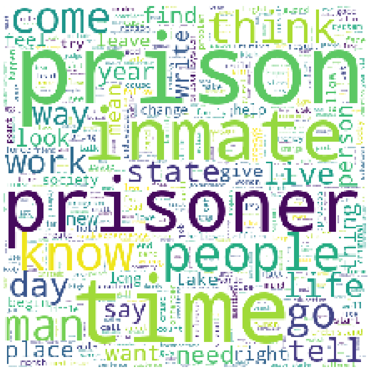
Next, I employ a variety of supervised learning methods as a more precise way of understanding whether there are systematic differences across race, sex, and state dimensions, which could be used to successfully predict the race, sex, or state of a given essay's author.

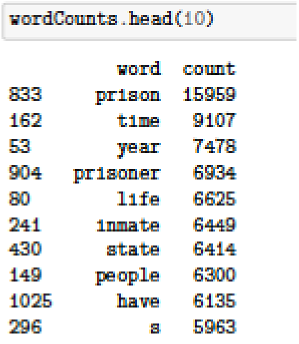
The analyses until now can only shed light on answering the question of whether differences exist in the essays. They fall short, however, in answering the question of what those differences might be. This motivates the use of structured topic modeling and word embedding techniques.

Structured topic modeling builds on traditional topic model methods by accounting for document-level (in this case essay-level) metadata such as the author attributes described above. Word embedding models differ from the above methods–which rely mostly on counting words—by mapping words or phrases to vectors of real numbers. I use these methods to get further insight about the local linguistic context or culture that produced those words.

Section III Findings and Interpretations

Section A: Word Counts and Divergences





```
wordCounts.head(10)

          word  count
833     prison  15959
162       time   9107
53        year   7478
904   prisoner   6934
80        life   6625
241     inmate   6449
430      state   6414
149     people   6300
1025      have   6135
296          s   5963
```

Figure 3: Word Cloud of entire corpus and top ten words, after normalizing

I begin by looking at counts of important words in the corpus. These words are deemed important because they appear after tokenizing and normalization has taken place – that is, after removing stop words such as "the", "of", "a" and so on. Figure 3 shows a Word Cloud of the entire corpus and the ten most common words appearing in the corpus. Figure 3 doesn't offer much insight or surprises but does offer a simple sanity check. Unsurprisingly, "prison," "prisoner," and "inmate" are extremely common in this corpus. Other words that are perhaps more interesting are "life," "time," and "think". A limitation of this method of course is that there is little context on these words. Is "life" a common word because many essays are pondering life, or because they are discussing life sentences?

Next, I look at divergences, or differences between the word frequency distributions of the corpora across race, sex, and state. To do this, I consider three types of divergences: Kullback-Leibler (KL) divergence, Kolmogrov-Smirnov (KS) test, and Wasserstein distance. All three of these methods approach calculating distance differently – KL divergence calculates the difference between two probability distributions, KS divergence calculates the difference between the cumulative distribution function, and Wasserstein distance calculates the minimum possible distance between the words. The results are presented in Figure 4.

Figure 4a: KL Divergence

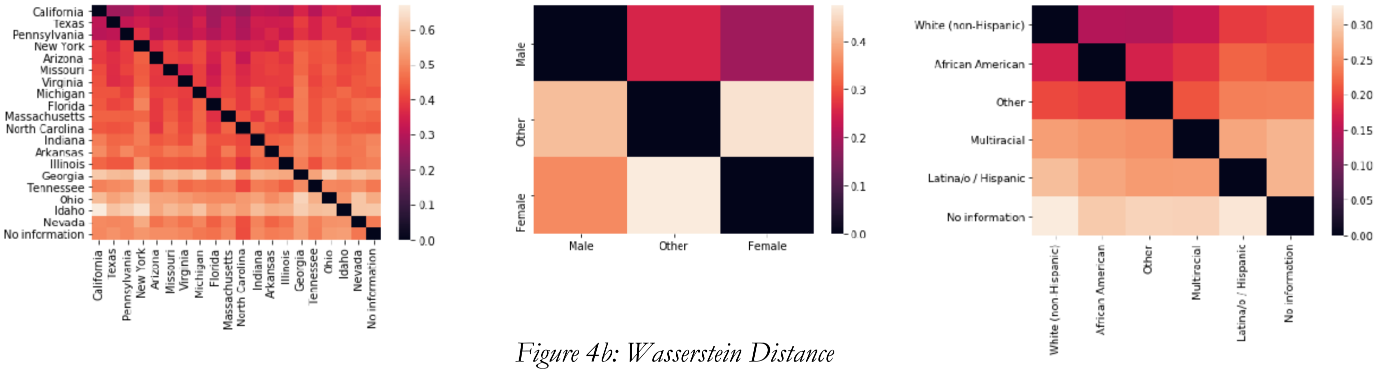


Figure 4b: Wasserstein Distance

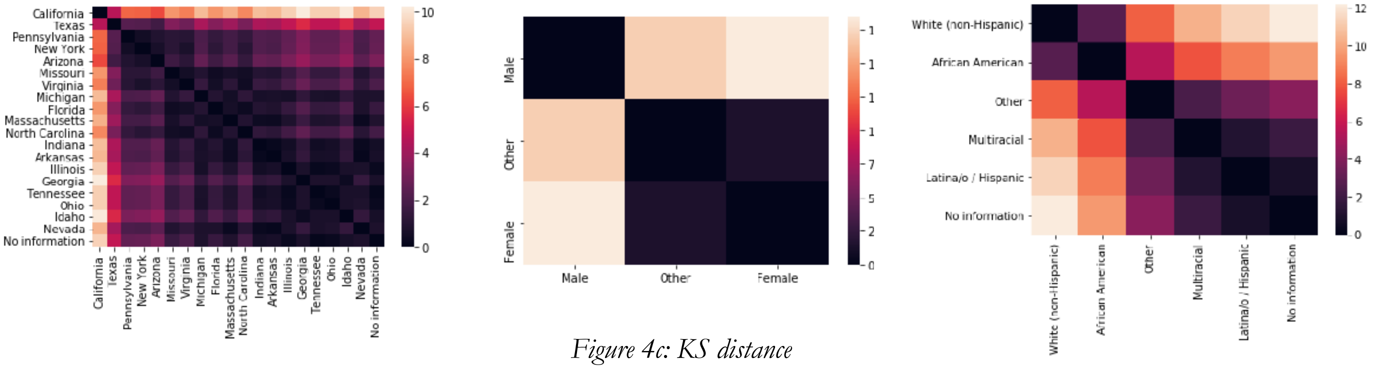
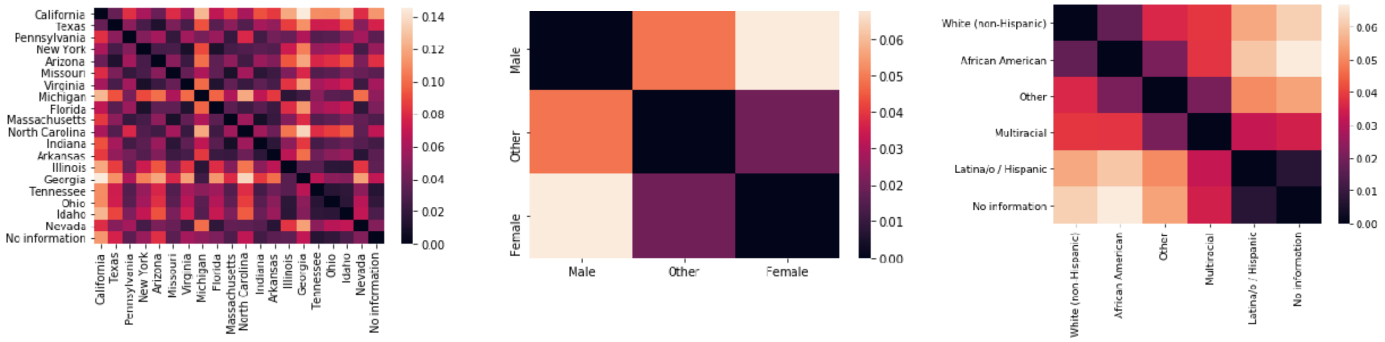


Figure 4c: KS distance



Based on these heat maps, there definitely appears to be divergences among the corpora across race, sex, and state dimensions. However, for KL and Wasserstein divergences, the divergence seems somewhat correlated with number of essays falling in that group. That is, groups with a large number of essays appear least divergent from others, whereas groups with a smaller number of essays appear most divergent. This suggests that the divergences may just be picking up sample size variation – groups with more essays probably contain a diverse set of words which make them more similar to other groups. This is less true for KS divergence, except in the case of race.

This is suggestive preliminary evidence that there are differences among these corpora, but this is just comparing divergences across words. Next, I do a classification exercise to see if I can actually predict with any accuracy and precision which essays come from which groups.

Section B: Classification

To understand whether the corpus shows any meaningful differences by sex, race, or states, I utilize a number of supervised learning methods. The idea is to test whether the corpora has predictive power in classifying the attributes as Black or White, male or female, or from California or not. I turn everything into a binary classification problem, since the sample size of a train/test split becomes increasingly small for other attributes.

In order to perform classification on texts, I first turn the essays into a TF-IDF matrix. TF-IDF (term frequency-inverse document frequency) is a product of two statistics that show the importance of a word with respect to the documents. Term frequency counts how many times a term is used in a document and inverse-document frequency measures how common or how rare this term is across documents.

After creating the TF-IDF matrix, I use logistic regression, naïve Bayes, ensemble methods, support vector machines (SVM) and a simple neural net to classify different author attributes. For each method, I compute the accuracy of the training data, accuracy of the test data, as well as precision, recall and F1 score.

The classification results on sex are not interesting - given the extremely small sample of essays written by female authors, this is unsurprising. The F1 scores for all classification methods range from 0 to 0.08. Results for whether the state of the author was California or not are somewhat clearer, but not very conclusive. The F1 scores of these methods ranged from 0.26 to 0.55.

The classification on race appears to be the most conclusive, when using these methods to predict whether or not the author of the essay is Black. These results are presented in Figure 5, sorted in descending order on F1 Score. Based on the table, the neural net performed the best, with an F1 score of 0.784, precision score of 0.877, and recall of 0.709. Figure 6 shows the confusion matrix based on the best performing method.

Classifier	Training Accuracy	Testing Accuracy	Precision	Recall	F1 Score
Neural Net	1.0000	0.8203	0.8772	0.7092	0.7843
SVM	0.9967	0.8235	0.8991	0.6950	0.7840
Naïve Bayes	0.8518	0.7092	0.6354	0.8652	0.7327
Logistic Regression	0.9723	0.7353	0.9545	0.4468	0.6087
Bagging trees	0.8127	0.6863	0.8169	0.4113	0.5472

Figure 5: Metrics of classification methods predicting whether author of an essay is Black or White

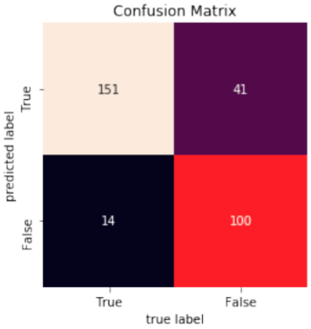


Figure 6: Confusion matrix of neural net predicting whether the author of an essay is Black (True) or White (False)

The neural net did not have the highest testing accuracy, but for the purposes of this exercise I chose a metric that balances both precision and recall, since correctly predicting whether an essay has a Black or a White author are equally important. Additionally, I conducted this analysis primarily to see if *any* classification method would perform well, giving greater confirmation that systematic differences exist in essays written by authors of different races.

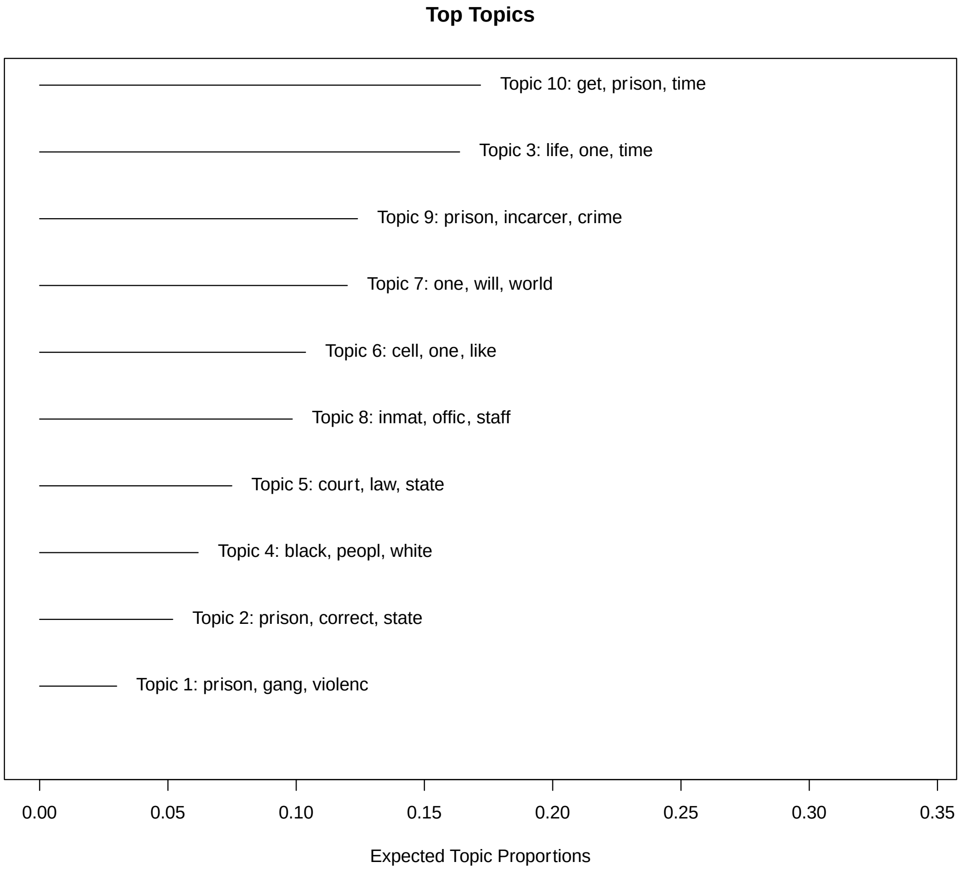
Taking the findings from Sections A and B into account, there is strong evidence that the essays differ by race, some evidence that they differ systematically by state, and much less evidence that they differ by sex. This is likely because of the low variation in the author's sex.

Section C: Structured Topic Modeling

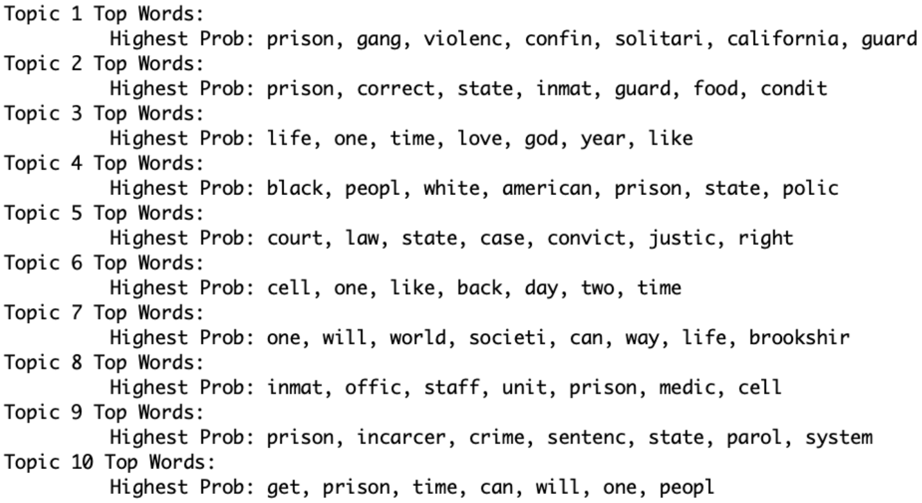
The methods so far only show whether there are differences by author attributes, but do not shed much light on what those differences might be. To uncover those patterns, I employ a variation of topic modeling called Structured Topic Modeling, which allows metadata from documents to be used in constructing topic models.

Topics are defined as a mixture over words, and each word has a probability of it belonging to a given topic. A document in turn is a mixture over topics, and a single document can be composed of multiple topics. To determine topical prevalence and topical content, structured topic modeling brings in the metadata of the document. This allows for a finer tuning of the topic model generation itself and allows analysts to then discover topics and estimate their relationship to the document metadata.

I use the *stm* package[4] in R to perform this structured topic modeling. As a first pass, I chose to model ten topics, and provided race, sex, and state as inputs to the model. The resulting topic models can be seen in Figure 7, showing the ten topics, words most likely to belong to those topics, and the expected proportions of each topic.



[4] Roberts ME, Stewart BM, Tingley D



```
Topic 1 Top Words:
 	 Highest Prob: prison, gang, violenc, confin, solitari, california, guard
Topic 2 Top Words:
 	 Highest Prob: prison, correct, state, inmat, guard, food, condit
Topic 3 Top Words:
 	 Highest Prob: life, one, time, love, god, year, like
Topic 4 Top Words:
 	 Highest Prob: black, peopl, white, american, prison, state, polic
Topic 5 Top Words:
 	 Highest Prob: court, law, state, case, convict, justic, right
Topic 6 Top Words:
 	 Highest Prob: cell, one, like, back, day, two, time
Topic 7 Top Words:
 	 Highest Prob: one, will, world, societi, can, way, life, brookshir
Topic 8 Top Words:
 	 Highest Prob: inmat, offic, staff, unit, prison, medic, cell
Topic 9 Top Words:
 	 Highest Prob: prison, incarcer, crime, sentenc, state, parol, system
Topic 10 Top Words:
 	 Highest Prob: get, prison, time, can, will, one, peopl
```

Figure 7: Results of topic modeling with 10 topics. Top graph shows their expected proportions, and bottom graphic shows the most common words for each topic

Figure 7 shows some pretty interesting results, some that are in line with the results from the counting exercises, and others that are new. For example, it's unsurprising that the most common topic (Topic 10) seems to contain the words "prison" and "time", which were among the top ten most common words. Other topics appear to explore more interesting parts of prison life – for example topic 8 appears related to health, topics 1 and 2 seem to be related to gangs, violence, confinement and guards, and topic 5 seems more related to the criminal legal system. Topics 3 and 5 diverge from the other topics as they seem more related to spirituality and race, respectively.

A concern is determining whether ten is the optimal number of topics for this corpus. This is important because the "quality" of each topic is affected by the number of topics chosen. I looked at some model diagnostics to determine model quality, and these results are presented in Figure 9. The diagnostics considered are held-out likelihood, residuals, semantic coherence, and lower bound, for different values of K (topics). In general, we want to maximize held-out likelihood, minimize residuals, and maximize semantic coherence. Given the results below, increasing the number of topics past ten may have increased the held-out likelihood or lowered the residuals. However, after ten topics the rate of change does appear to change, and semantic coherence declines steeply after 10 topics. For this reason, I continued with the ten topics above.

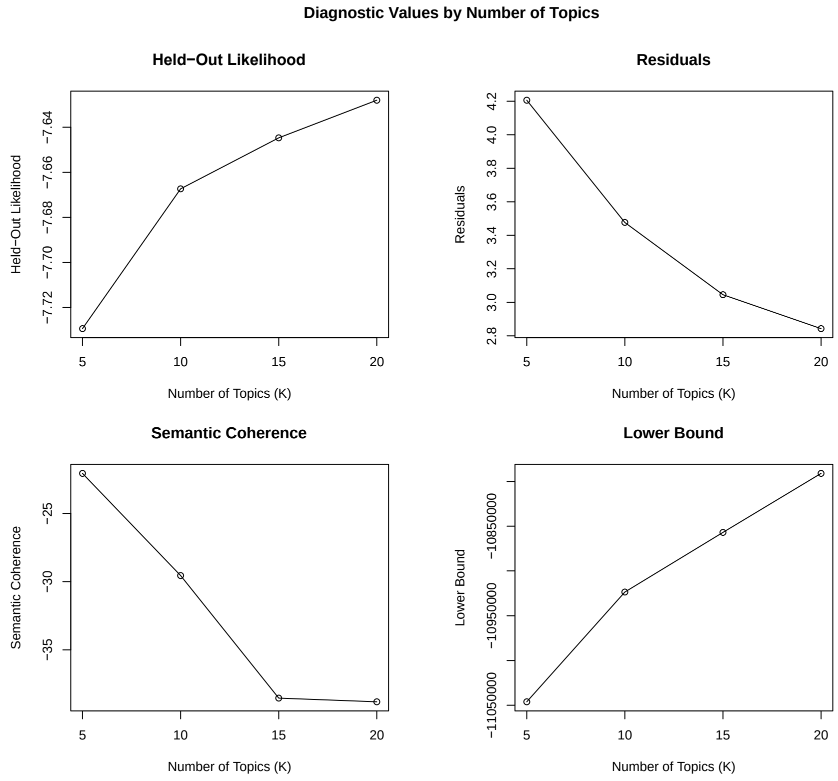


Figure 9: Model diagnostics to find optimal number of topics

Next, I look relationships between the metadata–specifically race and sex of authors—and the topics modeled. I contrast the prevalence of each topic for two groups – Black/White and male/female. The results can be seen in Figures 10a and 10b, where I plot the estimated difference for each topic. I created the topic labels, based on looking at the top words for each topic in Figure 7.

Some results jump out immediately in Figure 10a. Essays written by Black authors are much more likely to be composed of topics 4 (race), 3 (religion/spirituality), and 7 (society). Essays written by White authors are much more likely to be composed of topics 6 (cell), 8 (health), 9 (system), 5 (legal). Other topics don't appear to differ significantly by race. Some of these results certainly make intuitive sense, such as the prevalence of topic 4 (race) in essays written by Black authors, but others are topics I did not necessarily have a prior opinion of, such as topics 3, 5, 6, and 8. These differences likely contribute to the divergences and classifications seen earlier, and work towards answering the question of not only whether there are systematic differences in the essays by race, but what those differences might be.

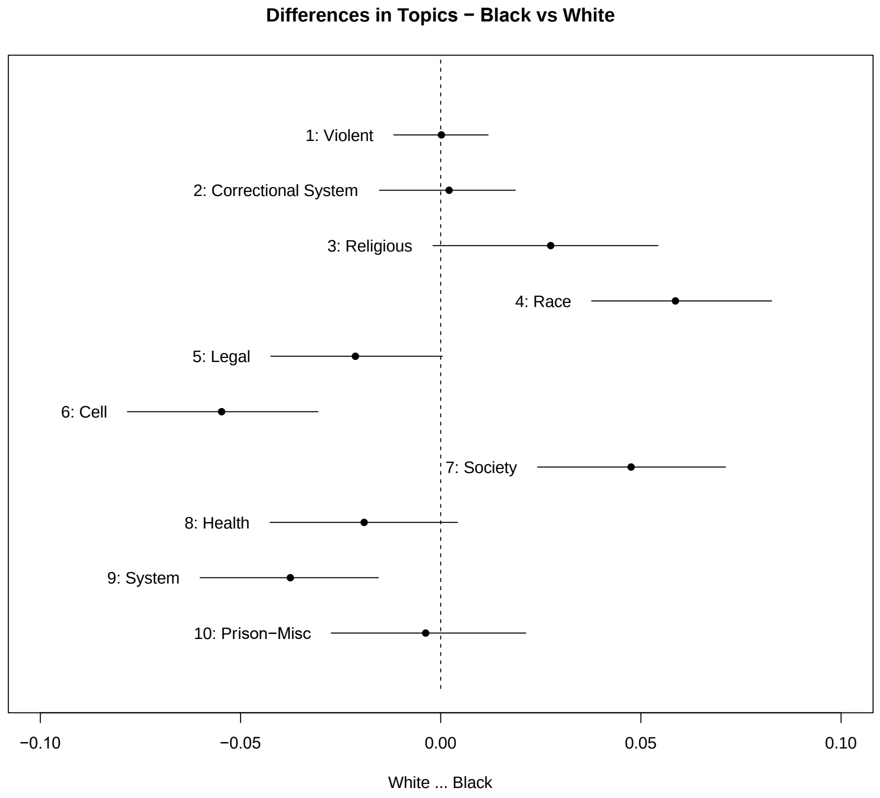


Figure 10a: Topic prevalence contrasted for Black and White authors

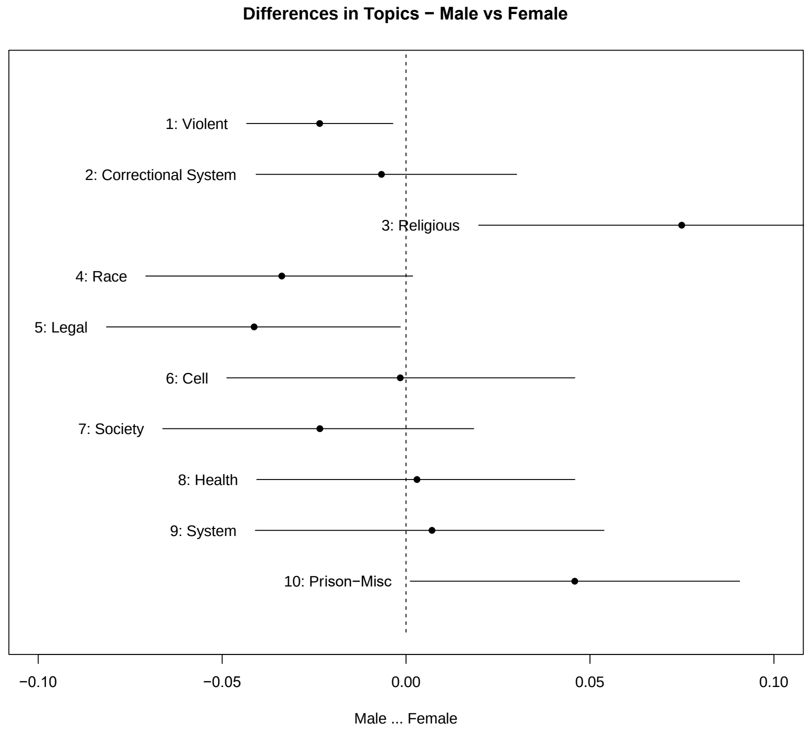


Figure 10b: Topic prevalence contrasted for male and female authors

The results in 10b are not as clear or stark as those in 10a. Even the differences that do appear are not always statistically significantly different, likely because of the small number of essays written by females. Despite this, topics 1 (violence) and 5 (legal) are much more related to male authors, and topic 3 (religion/spirituality) is much likelier to be in essays written by females.

To look beyond top words, I look at quotes from documents that load heavily on the topics with the most divergence by race, topics 3 (religion/spirituality), 4 (race), 7 (society), 6 (cell), and 9 (system). Figure 8 presents these quotes.

"Choosing to love Betrayal and wounds I was the one who betrayed and caused wounds Not an easy thing to live with How do you rebuild recreate begin again once you have caused so much destruction is it even possible to repair broken relationships I say yes God says yes I say yes because God said yes to me I gave my nothingness to God who took it and made something I hurt those closest to me I betrayed them in the worst way possible Yet God has brought us back together My family has shown love to me by extending grace forgiveness and mercy"

- Courtney Kirby-Watts, Texas, Topic 3 (religion/spirituality)

"Six out of eight americans believe that race relations have worsen But we need to look into why has it worsened I believe many Americans are becoming more aware of the conditions in America I think race relations has worsened under Trump because of his normality of white supremacy His lets make America great again slogan to some people was a subliminal message of making America white again"

- Samuel Foster II, Pennsylvania, Topic 4 (race)

"As if I was observing some kind of sickening and sadistic scientific experiment I watched the garbage bag I had wrapped around the toilet slowly rise as the unflushed feces emanated methane gas I was fascinated by the sight of science in action and horrified by the knowledge that soon I would have to remove the plastic covering to take a leak and in so doing get a mighty whiff of my own shit"

- William D. Hastings, Illinois, Topic 6 (cell)

"Lying somewhere at the core of every individual circumstance of life the heart of the lesson may be found Nowhere does this prove more evident than in the dynamic that is contained inside the prison system Within the constant friction that occurs as a result of authoratarian interaction a pattern begins to emerge a truth of human nature This pattern if one conditions himself to recognize it can be utilized in any and all situations as a means to overcome the challenges we face in our environment"

- Todd Alan Leatherland, Texas, Topic 7 (society)

"The yearly operating budget of the Michigan Department of Corrections is over 2 billion dollars consuming a 19 of the State's genera fund This is far more than what Michigan spends on education While no one doubts the necessity of prisons reducing expenditures without sacrificing public safety and security is possible In recent years Michigan's legislators have considered various reforms but have passed few laws to significantly impact the prison system"

- Daniel Pirkel, Michigan, Topic 9 (system)

Figure 8: Quotes from essays that load heavily on different topics

The quotes in Figure 8 show a broad range of topics, emotions, and sentiments that are only glimpsed earlier with the sample titles, word counts, and even topic labels. I present these quotes without presenting the race of the author intentionally, as I do not want to cherry-pick findings to highlight divergence in topic prevalence by race. Instead, these quotes are meant to provide insight into the corpus more generally and help flesh out the topics seen in above figures.

Section D: Word2Vec

Structured topic modeling goes a long way in shedding light on the patterns that emerge from the American Prison Writing Archive. Seeing words grouped together in topics that they are most likely to load on, and by "reading" the documents that load heavily on those topics, provides immense insight into the themes in the corpus, and the differential patterns that may emerge by sex or race of the authors.

Using word embedding models allows digging deeper and learning about the words seen in their local linguistic context. The methods I have employed so far primarily use word counts, including structured topic modeling which is a generative model of word counts. As mentioned above, word embedding models differ from such models by actually mapping words or phrases to vectors of real numbers. Word2vec and Doc2vec are two models that are used to produce word embeddings.

As a first step, I use the *genism* implementation of Word2Vec and provide the model with sentences from the APWA corpus. Unlike the above methods where the models take unordered words, word embedding models take sentences as input, since retaining this structure is important for training the model. Figure 11 shows the result of using principal component analysis to reduce the dimensions of our Word2Vec model and project those onto two dimensions.

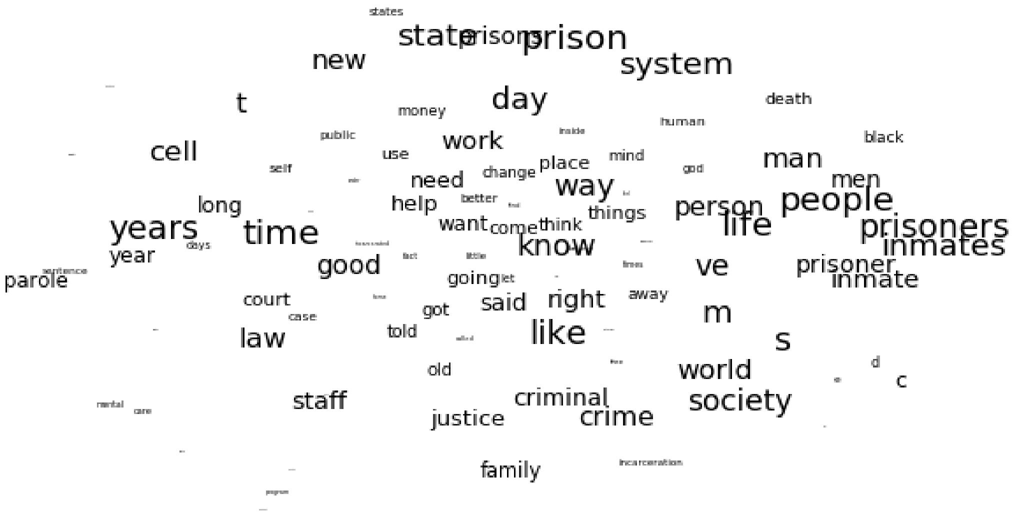


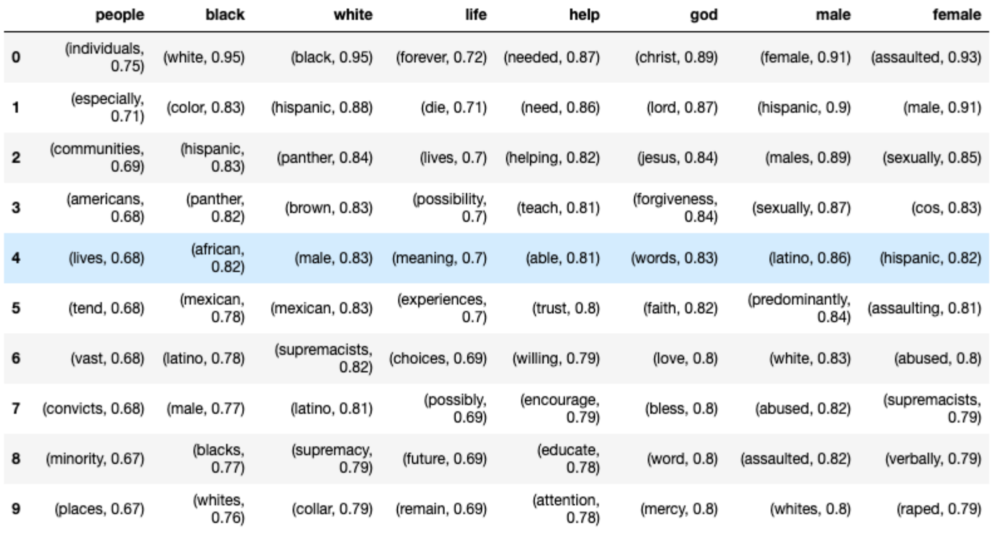
Figure 11: Visualization of Word2Vec model. 100 words, 80 components

Figure 11 provides some useful insight and is also consistent with some of the results seen above. For example, a lot of the same words appear, including in the Word Cloud. However, there is now more context for those words. For example, "prison" and "prisoners" appear in multiple different contexts – sometimes close to "inmates," "people," and "life" and sometimes close to "state," "system," and "money". The former appears more related to the people in prison, while the latter seems to refer to prison as a system, and the prison industrial complex more specifically. This also appears consistent with the topic modeling results, as prison and prisoners are discussed in the context of society, a broader system, and the gory details of daily life in cells.

There's also a set of words that have not come up as often before – "help", "want", "need", "better", and "change". These, combined with the quotes in Figure 8, suggests that a reasonable subset of the essays reflect on the author's current circumstances, express feelings of remorse, or a cry for help.

While useful, Figure 11 is created from a non-deterministic process and does not provide clarity on a lot of words that appear in the corpus which could have various meanings. Luckily, word2vec models allow us to find words that are "most similar" to other words, using cosine similarity.

Figure 12 shows the top ten words most similar to a given set of words. This provides more clarity on the context of some recurring words, like "life". It appears that in this corpus, "life" is more closely related to existential questions, death, and meaning rather than "life sentences".



	people	black	white	life	help	god	male	female
0	(individuals, 0.75)	(white, 0.95)	(black, 0.95)	(forever, 0.72)	(needed, 0.87)	(christ, 0.89)	(female, 0.91)	(assaulted, 0.93)
1	(especially, 0.71)	(color, 0.83)	(hispanic, 0.88)	(die, 0.71)	(need, 0.86)	(lord, 0.87)	(hispanic, 0.9)	(male, 0.91)
2	(communities, 0.69)	(hispanic, 0.83)	(panther, 0.84)	(lives, 0.7)	(helping, 0.82)	(jesus, 0.84)	(males, 0.89)	(sexually, 0.85)
3	(americans, 0.68)	(panther, 0.82)	(brown, 0.83)	(possibility, 0.7)	(teach, 0.81)	(forgiveness, 0.84)	(sexually, 0.87)	(cos, 0.83)
4	(lives, 0.68)	(african, 0.82)	(male, 0.83)	(meaning, 0.7)	(able, 0.81)	(words, 0.83)	(latino, 0.86)	(hispanic, 0.82)
5	(tend, 0.68)	(mexican, 0.78)	(mexican, 0.83)	(experiences, 0.7)	(trust, 0.8)	(faith, 0.82)	(predominantly, 0.84)	(assaulting, 0.81)
6	(vast, 0.68)	(latino, 0.78)	(supremacists, 0.82)	(choices, 0.69)	(willing, 0.79)	(love, 0.8)	(white, 0.83)	(abused, 0.8)
7	(convicts, 0.68)	(male, 0.77)	(latino, 0.81)	(possibly, 0.69)	(encourage, 0.79)	(bless, 0.8)	(abused, 0.82)	(supremacists, 0.79)
8	(minority, 0.67)	(blacks, 0.77)	(supremacy, 0.79)	(future, 0.69)	(educate, 0.78)	(word, 0.8)	(assaulted, 0.82)	(verbally, 0.79)
9	(places, 0.67)	(whites, 0.76)	(collar, 0.79)	(remain, 0.69)	(attention, 0.78)	(mercy, 0.8)	(whites, 0.8)	(raped, 0.79)

Figure 12: Top 10 words most similar to given word

Figure 12 also shows the context of other words – "black" and "white" seem closely related to each other, but other words similar to "black" predominantly feature other race-related words. Other words similar to "white" however range from race-related words ('hispanic', and 'supremacy') to other contexts (e.g. 'white collar'). It is disturbing and yet perhaps unsurprising that most of the words related to "male" and "female" have strong violent connotations – "assaulted", "sexually", "abused", "verbally" and "rape" for female, and similar words (though much less prevalent) for males. Interestingly, "male" seems more related to race-related words, such as "Hispanic", "white", and "latino".

Section E: Doc2Vec

Thus far, the analysis of differences by race and sex has been fairly coarse – divergences, classification, and even structured topic modeling split the authors of these essays into crude buckets, defining them only by their race and sex. In this section, I utilize the powerful doc2vec models to get more nuanced and look at similarities and differences, at the essay level.

I first train a doc2vec model – this differs from the word2vec model by seeing how documents relate to each other within a space, rather than seeing how words relate to each other. I then separate out essays written by Black and White authors, and randomly sample 10 essays from each subset. I plot a heatmap comparing the sample of Black essays with White essays. The heatmap shows the cosine similarity of these documents with each other – a similarity of 1 means the documents are very closely related, 0 means they are orthogonal, and -1 means they are diametrically opposed. The results of this are presented in Figure 13a.

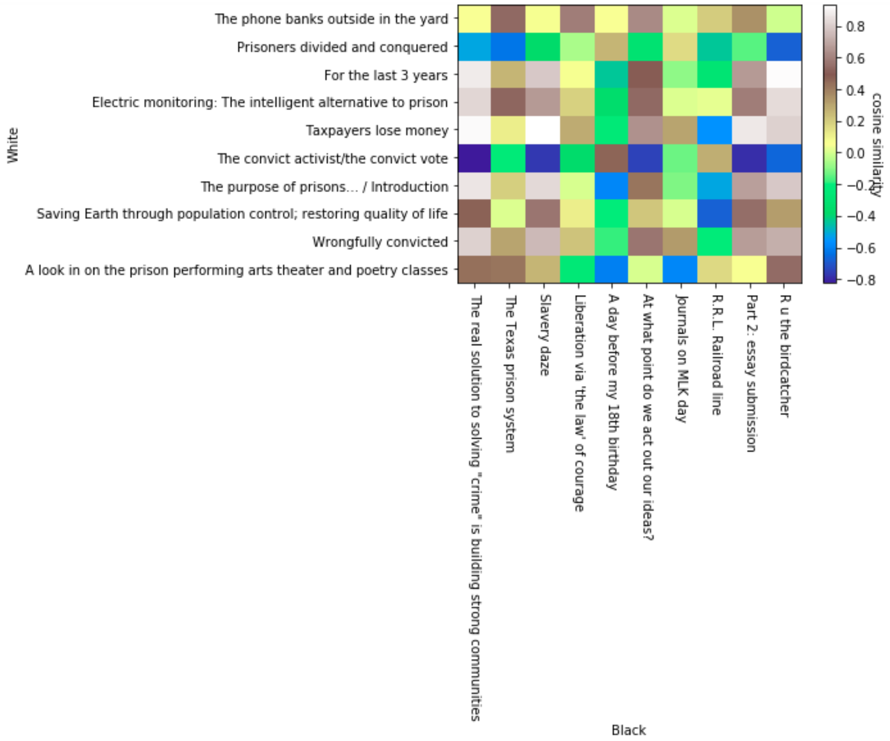
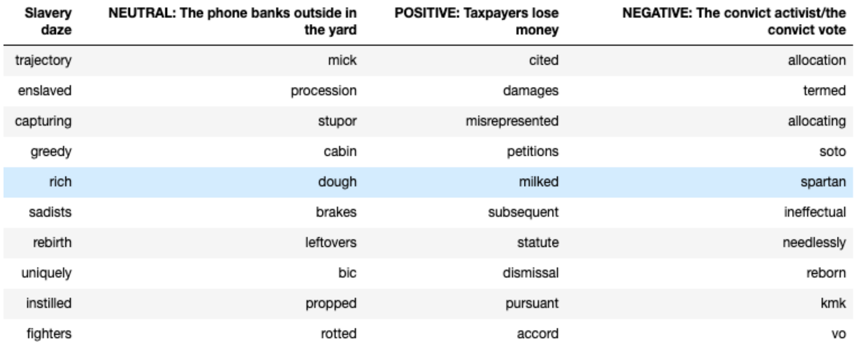


Figure 13a: Document similarity of 10 randomly sampled essays with White and Black authors

It is immediately evident from Figure 13a that there is no clear pattern in the similarity or differences of the essays. Of course, this is a small random sample, and previous analysis has already demonstrated differences across these categories, on average. At the document level, however, there is a great deal of variation in the topics covered, which can often get masked when only considering broad strokes and averages. This doc2vec model can be used to gain further insight on what types of stories are captured by this corpus and generate new hypotheses explaining differences by race at the aggregate level. One theory may be that the essays about prison life tend to be similar regardless of the race/sex of the author, but essays about personal stories or current affairs tend to be more divergent. This is not necessarily the conclusion I arrive at, since I am only exploring document similarity for a small subsample of essays. However, this exercise is illustrative of how these methods can be used to further develop an understanding of the corpus.

To gain insight past the essay titles, I select the essay titled "Slavery daze" from Figure 13a. I chose this essay because it relates very differently to the sample of essay from White authors – it is apparently very similar to "Taxpayers lose money," diametrically opposite from "The convict activist/the convict vote," and orthogonal or unrelated to "The phone banks outside in the yard." To understand why these essays may have been categorized as similar or different to each other, I look at the most similar words associated with each of the documents.

The results are presented in Figure 13b. This figure helps clarify why those essays may have been assigned the cosine similarity values seen in the heat map in Figure 13a. Words in "Slavery daze" seem unrelated to words in "The phone banks outside in the yard", while "greedy", "enslaved", "sadists" do appear similar to "damages", "milked", "misrepresented". They also appear different from more generous-sounding words such as "allocation" and "allocating".



Slavery daze	NEUTRAL: The phone banks outside in the yard	POSITIVE: Taxpayers lose money	NEGATIVE: The convict activist/the convict vote
trajectory	mick	cited	allocation
enslaved	procession	damages	termed
capturing	stupor	misrepresented	allocating
greedy	cabin	petitions	soto
rich	dough	milked	spartan
sadists	brakes	subsequent	ineffectual
rebirth	leftovers	statute	needlessly
uniquely	bic	dismissal	reborn
instilled	propped	pursuant	kmk
fighters	rotted	accord	vo

Figure 13b: Top ten words most similar to a selection of related, unrelated, and opposing documents from Figure 13a

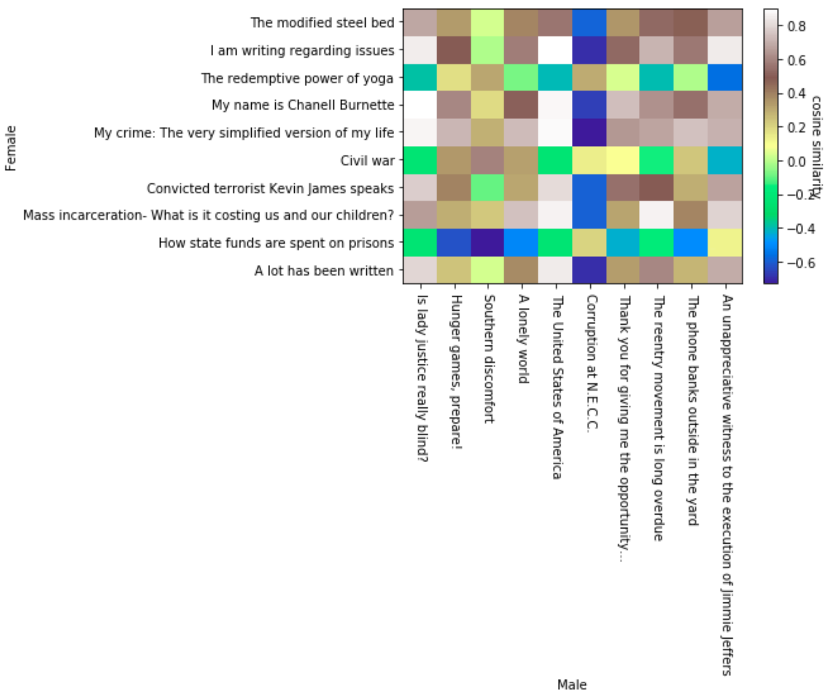


Figure 14: Document similarity of 10 randomly sampled essays with male and female authors

I repeat the exercise above with small sample of essays written by male and female authors. The result is presented in Figure 14. Like Figure 13a, the lack of pattern in document similarity for this small sample is striking.

Section V: Discussion, Limitations, and Future Work

The analysis above reveals many insights about the essays in the American Prison Writing Archive. It shows a very diverse range of themes across the essays – from intensely personal stories to musings about the corrupt prison industrial complex and displays a range of emotions and sentiments including anger, disgust, fear, remorse, sadness, and at times, optimism. Beyond general themes and patterns, there is also evidence that these themes and topics vary systematically by the race and perhaps sex of the author. The classification and structured topic modeling results are the strongest evidence for this variation – there is something about the corpus which allows for high quality predictions and classification on the race of the author. Perhaps the most significant and interesting findings that emerge are the prevalence of different topics by race. Whether it is a function of personal experience, prison experience, or something else, there is a stark difference in the prevalence of race and spiritual topics in essays by Black authors, and the prevalence of more cell and system related topics by White authors.

However, the analysis also reveals that these broad categorizations of Black/White and male/female may be overly simplistic, and the word2vec and doc2vec analyses reveal that such categorizations can cover up incredibly important nuances that make individual essays relate to other essays in ways that cannot be wholly captured by author demographics.

A wide variety of factors influence an individual's experience in prison, including age, medical vulnerability, length of sentence, behavior in prison, type of living environment (e.g. regular population or solitary confinement). It is possible, and likely, that some of these important variables also correlate with race and sex, resulting in the patterns seen above.

Beyond the lack of important covariates, this sample is also not necessarily representative of the incarcerated population. As discussed above, authors self-select into the sample, sometimes repeatedly, and there is little information on the reasons behind this self-selection. Future avenues for this work may therefore involve collecting richer and more granular information about the authors – when this can be done ethically and with the author's consent – to inform clearer conclusions. Another avenue for future work would be replicating this analysis on a similar corpus, and seeing what patterns emerge. The Marshall Project's Life Inside[5] is a series of essays from people who live or work in the criminal justice system. While these essays are not documented with author attributes as systematically as APWA, they do offer insight from a related population.

Given the limitations of this analysis, I do not aim to make strong conclusions about the variation seen across demographic categories. Instead, my goal with this paper was to utilize a range of content analysis techniques – from simple counting to powerful word embeddings – to systematically explore a corpus that reveals unprecedented insight into the lives and experiences of a frequently-forgotten and marginalized population.

[5] The Marshall Project

References

Sawyer, Wendy, and Peter Wagner. "Mass Incarceration: The Whole Pie 2020." Mass Incarceration: The Whole Pie 2020 | Prison Policy Initiative. Accessed June 12, 2020. https://www.prisonpolicy.org/reports/pie2020.html.

Roberts ME, Stewart BM, Tingley D (2019). "stm: An R Package for Structural Topic Models." *Journal of Statistical Software*, **91**(2), 1–40. doi: 10.18637/jss.v091.i02.

American Prison Writing Archive at Hamilton College. Accessed June 12, 2020. https://apw.dhinitiative.org/.

"Life Inside." The Marshall Project. Accessed June 12, 2020. https://www.themarshallproject.org/tag/life-inside.