Predicting Banned Books Using Topic Modelling and Logistic Regression

Project created by Shirley Li on November 14, 2023

Intro

Over July 1, 2021 to June 30, 2022, PEN America recorded **2531 book bans of 1656 unique books** in schools across America. According to PEN America, popular themes for censorship are race, history, and gender and sexuality. Is this in fact the case? Let's build a model that will predict whether a book will be banned, and what themes ("features") are relevant.

Data

- I sourced my banned books from PEN America (spreadsheet here). These book bans occurred from July 1, 2021 - June 30, 2022. I retained information about authors and titles.
- Using BeautifulSoup, I scraped my non-banned books from a private blog, Library of 1000 Books, containing 1000 well-known books. I checked for overlap with the PEN America list and removed duplicates. I retained authors and titles.
- Using the Google Books API, I pulled Book Descriptions for both spreadsheets (banned and well-known books). I was not able to pull descriptions for all books, perhaps due to spelling errors or otherwise.
- After cleaning, I was left with a total of **1477 books**, **938 of which were banned and 539 of which were not banned.**

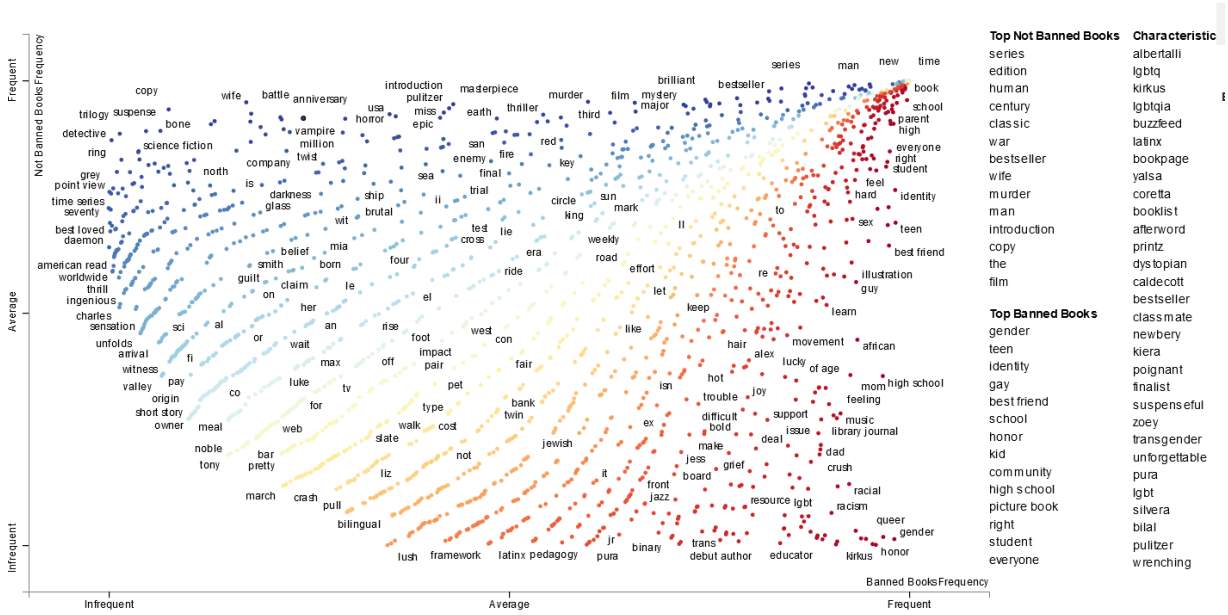
Tools

1. Numpy and pandas for data wrangling
2. requests and BeautifulSoup for web scraping
3. Google Books API for obtaining book descriptions
4. matplotlib, seaborn, and Scattertext for data visualization and analysis
5. NLTK and spacy for NLP analysis (stopwords, tokenization, lemmatization, and more)
6. CountVectorizer and TFIDF for vectorizing words, and TruncatedSVD and NMF for topic modelling
7. sklearn for modelling and evaluation

Data Wrangling and EDA

I cleaned my data ("Book Descriptions") using various NLP techniques:
- removed stopwords
- removed all words except nouns and adjectives
- lemmatized words
- removed descriptions with too few words, truncated descriptions with too many

I used Scattertext to create a visualization of nouns and adjectives appearing in the descriptions of banned and non-banned books. As you can see, words like queer, gender, racial, etc. are more likely to appear in banned books, whereas time series, trilogy, detective, etc. are more likely to appear in non-banned books.



You can see the interactive Scattertext Visualization here. Note that it will take some time to load.

Modelling and Results

I performed topic modelling to create "features" for my book descriptions. In the end vectorizing with TFIDF and singular value decomposition with NMF created the most descriptive columns:

```
nmf_top_fourteen = display_topics(nmf, tfidf.get_feature_names_out(), 10)
nmf_top_fourteen
```

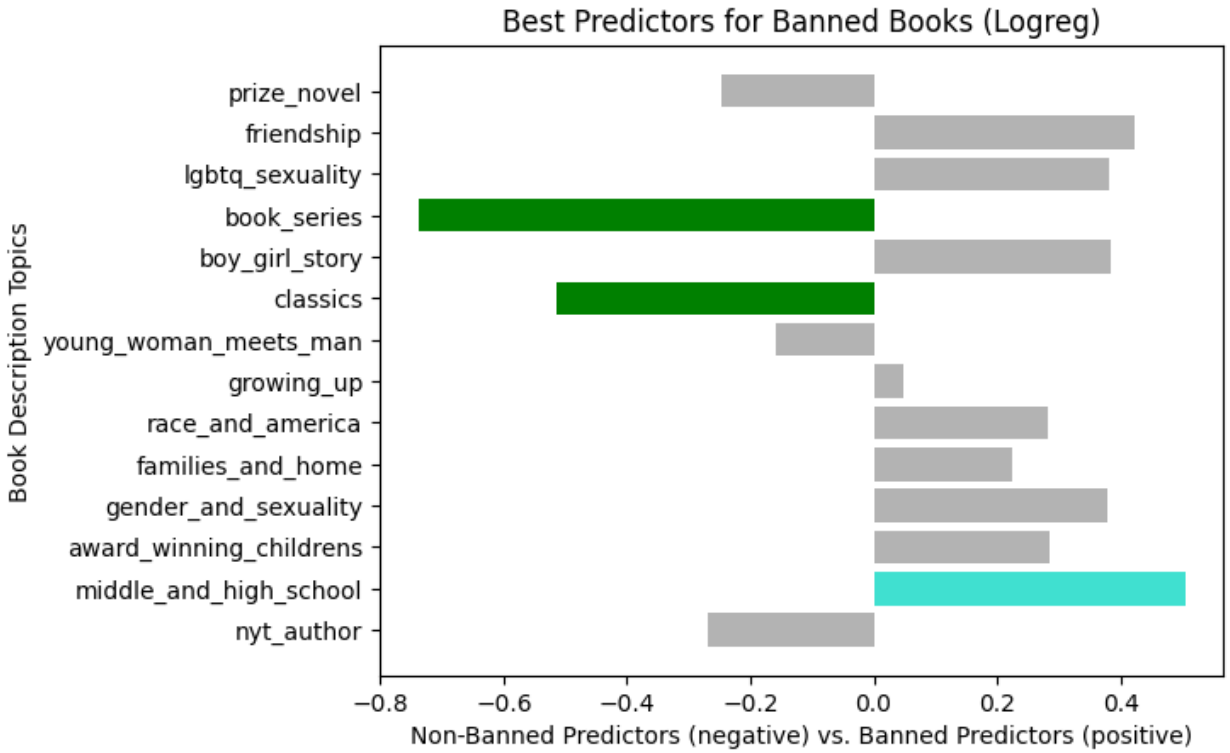
```
Topic  1
time, new, york, review, author, book, the, post, today, day

Topic  2
school, high, student, kid, group, teacher, classroom, senior, middle, new

Topic  3
book, child, picture, award, best, honor, library, illustrator, winner, young

Topic  4
gender, identity, body, binary, sexuality, sex, guide, reader, self, female

Topic  5
family, home, child, parent, mother, house, dad, different, brother, life

Topic  6
american, black, african, white, civil, right, racial, america, history, race

Topic  7
year, old, fifteen, mother, father, seventeen, twelve, life, sixteen, thirteen

Topic  8
woman, life, first, young, men, medicine, career, printing, southern, trace

Topic  9
classic, edition, introduction, penguin, work, reader, literature, note, life, story

Topic  10
boy, girl, story, something, little, young, town, sister, dress, day

Topic  11
world, series, war, human, man, vampire, power, fantasy, death, secret

Topic  12
people, gay, community, lgbt, transgender, teen, lesbian, many, bisexual, issue

Topic  13
friend, best, thing, love, life, everything, secret, summer, friendship, ya

Topic  14
novel, author, award, winning, story, national, love, winner, character, prize
```

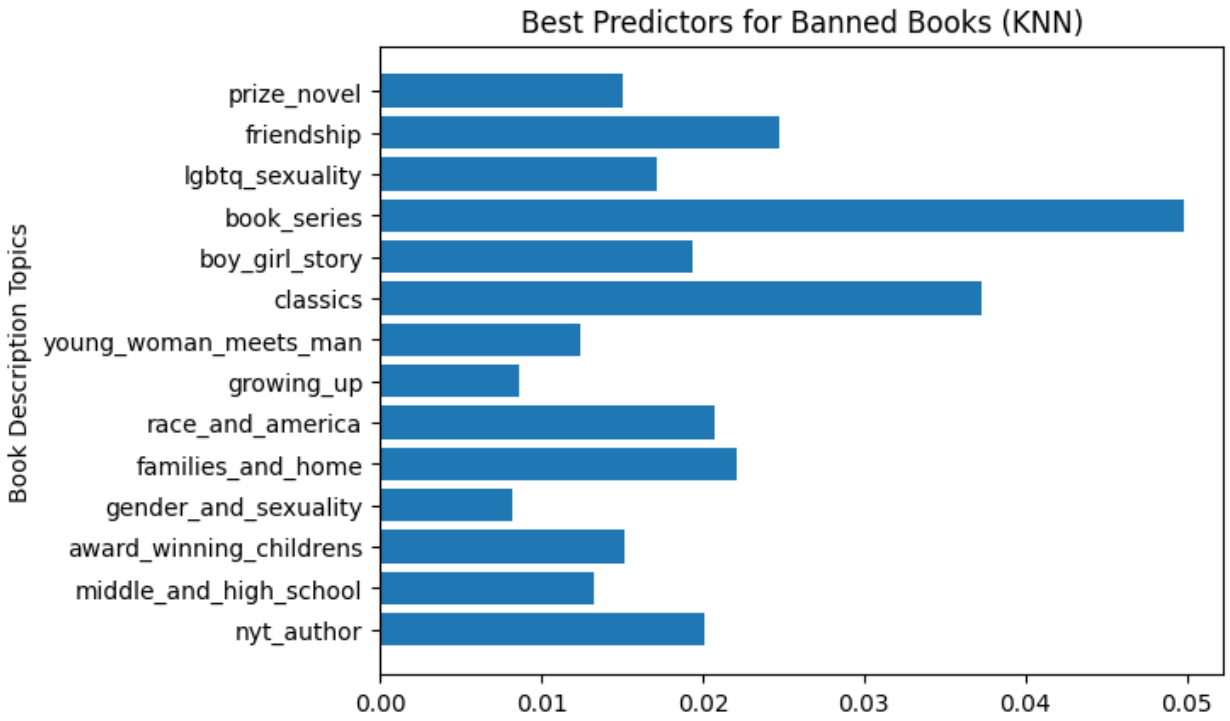
I described these topics with the following feature names:

1. "nyt_author"
2. "middle_and_high_school"
3. "award_winning_childrens"
4. "gender_and_sexuality"
5. "families_and_home"
6. "race_and_america"
7. "growing_up"
8. "young_woman_meets_man"
9. "classics"
10. "boy_girl_story"
11. "book_series"
12. "lgbtq_sexuality"
13. "friendship"
14. "prize_novel"

Using these features, I tried a few models: Logistic Regression, KNN, Decision Tree, and Random Forest. They performed about equally, so for maximum interpretability I stuck with Logistic Regression. It gave me about 78% accuracy on the test set (an improvement from 63% with the dummy classifier). Our parameters were penalty = l2, C = 0.1, solver = lbfgs, class_weights = balanced.

As we can see from both the graphs below, the books that are least likely to be banned seem to be Book Series (fantasy, war, etc.) and Classics. From the Logistic Regression coefficients, the books that are more likely to be banned include books about middle and high school experiences, followed closely by books about gender and sexuality, lgbtq issues, friendship, romance, and children's books.





Future Steps

1. I'd like to scrape more data for non-banned books. Right now I am on the edge of imbalanced data (about 60% banned/40% non-banned). In particular I would like to source more books from elementary, middle, and high school libraries (rather than 1000 popular books in general). Some possible sources include the American Library Association (ALA) and the Association of Library Services for Children's (ALSC) summer reading lists, and the Battle of the Books (national and NC) reading lists. However, scraping these would require quite a bit of data wrangling as they're embedded in multiple webpages and/or pdfs.
2. I performed my topic modelling manually. I wonder: is there a more systematic way to find the best # of topics when working with singular value decomposition for NLP? Or perhaps this is a place for interpretation and judgment.
3. Ideally I would get my model to 80% accuracy. I think that having more data (closer to 1000 banned books/1000 non-banned books) would certainly help with this goal.