Banned Book Recommender System

Team

Prathyusha - prathyusha1231

Nikitha Chandana - NikithaChandana

Prakruti Pruthvi Kumar - ppruthvi22

Introduction

The problem that arises with banning books is restricting information and discouraging freedom of thought. Censors undermine one of the primary functions of education: teaching students how to think for themselves. Such actions, assert free speech proponents, endanger tolerance, free expression, and democracy

Our project proposes to recommend banned books tailored to individual user preferences in regular literature. By providing access to banned books, we empower stakeholders to explore diverse perspectives and exercise their right to free expression, fostering a more informed and democratic society.

As people invested in learning, we believe everyone should have access to a variety of books, even if they tackle tough topics or stir controversy. Offering banned books lets people face challenges, question beliefs, and learn more about the world around them. The stakeholders for our project are students, parents, librarians, and readers. Banning books deprives students of exposure to different viewpoints, limiting their intellectual development and inhibiting their ability to think critically. It undermines parents' authority to guide their children's reading choices, restricting their access to valuable learning experiences and diverse perspectives. The librarian's mission to provide access to information and foster intellectual freedom, impeding their ability to serve the needs of their community gets contradicted. Banning books deprives readers of the opportunity to explore challenging themes, limiting their ability to expand their understanding of the world.

Literature Review

This section will provide additional detail about your problem; feel free to copy some of this from your proposal. I would like you to clearly establish your stakeholder need(s), and then tell me why you chose the methods you chose based on prior work and the nature of your

problem. If no one has worked on your problem (unlikely) you can say that but should then talk about related problems that apply similar methods.

Data and Methods

The Amazon Book Review dataset comprises user ratings for books, including features such as User ID, Book ID, and Rating, sourced from Amazon reviews data posted by UCSD in 2023, with a size exceeding 3 million entries.

The Amazon Book data consists of book information such as Title, Description, Authors, and Categories, derived from Amazon Item Metadata also posted by UCSD in 2023, with a dataset size exceeding 200,000 entries.

The Banned Book data, sourced from PEN America and the Open Library API between July 2021 and December 2023, includes features like Title, Description, Authors, and Categories, with a dataset size exceeding 2,000 entries.

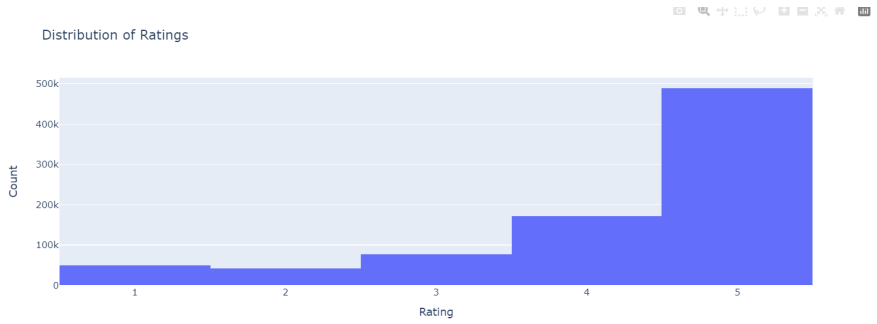
Data

For this project, we utilized three datasets: the Amazon User data, the Amazon Book data, and Banned book data. Given that users typically don't rate or engage with numerous books, our data naturally suffered from sparsity. To address this issue, we opted to merge (based on BookTitle) two of these datasets. The goal was to enhance the quantity of ratings associated with each book.

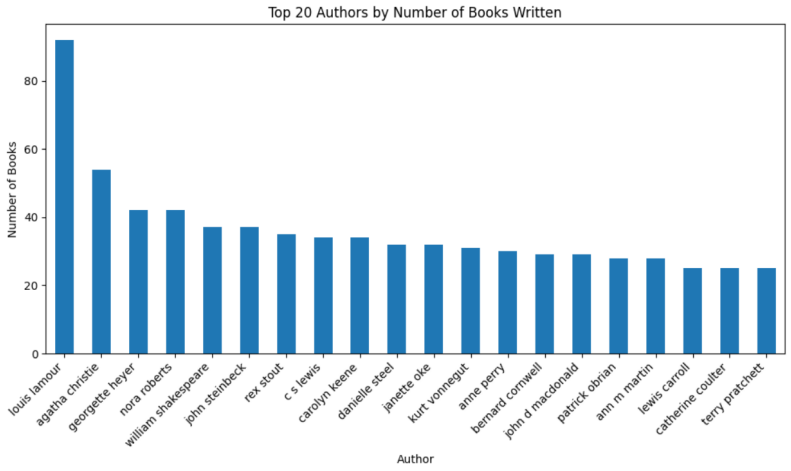
The Banned Book data was extracted using the Open Library API from the website; PEN America. The data extracted is between July 2021 and December 2021, it features the title, Description, Authors, and Categories.

Data Visualization

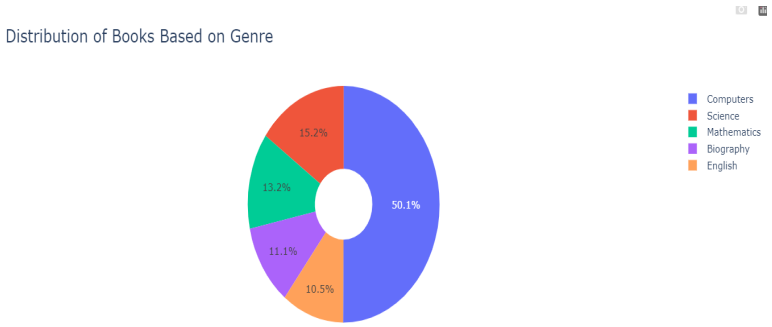
We used library plotly to generate interactive visualizations.



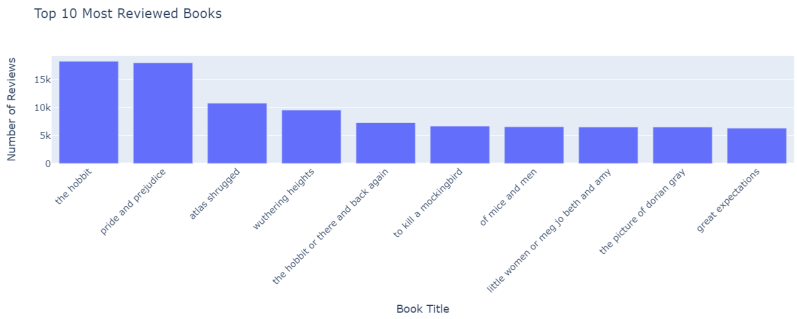
The above chart describes the distribution of the ratings with the count of the review.



The bar chart shows the top 20 authors by the number of books written by them.



The pie chart shows the Distribution of Books based on Genre of the books.



Lastly, the bar chart shows the top 10 most reviewed books.

Methods

Basic Preprocessing of Data:

1. First, we checked for duplicate rows in the datasets and dropped them to ensure data integrity. Then we looked for missing values and dropped rows containing missing values.

2. We dropped unnecessary columns that were not relevant to our analysis and standardized the column titles for consistency.

Text Preprocessing:

1. Lowercasing: All text data (book titles, descriptions, author names) was converted to lowercase to ensure consistency in the text.

2. Removing Inconsistencies: Inconsistencies in the text, such as extra text like "a novel" or "by [author name]", were standardized or removed.

3. Removing Extra White Spaces: Any extra white spaces at the beginning, end, or between words in the text were removed to clean the data further.

After Text Preprocessing:

1. We checked for unique values in each column to ensure data integrity. After performing all the preprocessing, we generated a unique book ID for the Amazon datasets, with IDs starting with "A".

2. To enhance uniqueness and avoid duplicates, we checked for rows with the same title, authors, and description, and we dropped all duplicate entries.

3. We filtered out books based on authors with more than two publications in the dataset to improve model performance and reduce potential biases introduced by over-represented authors.

Banned Books Dataset:

1. We performed similar preprocessing tasks for the banned books dataset, including removing duplicates, handling missing values, text cleaning, and checking for unique values.

2. We generated unique book IDs for the banned books dataset, with IDs starting with "B" to distinguish them from the Amazon dataset.

3. After preprocessing both datasets, we combined them into a single dataset for further analysis.

4. We converted the relevant text columns (titles, descriptions, author names) to numerical embeddings using text embedding techniques (e.g., Word2Vec, BERT, etc.) for use in the subsequent modeling and analysis steps.

Text Embedding Methods:

1. Glove Word Embeddings: We used Glove word embeddings, GloVe embeddings may not capture out-of-vocabulary words or domain-specific terms well, especially if they were not present in the training corpus. Hence, this didn't work efficiently for us.

2. DistilBERT: This pre-trained transformer model was used to convert text data into dense vector representations (embeddings). DistilBERT is a smaller version of BERT, making it faster and more memory-efficient while still retaining much of its performance.

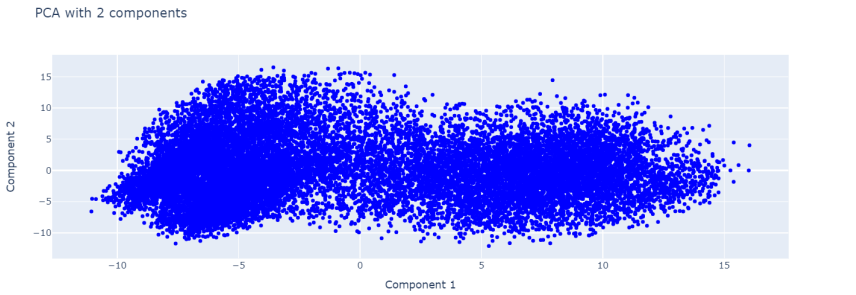
2. T5 Tokenizer: The T5 model and tokenizer were utilized to encode the text into fixed-length vector representations. T5 is a versatile model capable of performing various text-to-text tasks, including text summarization and translation.

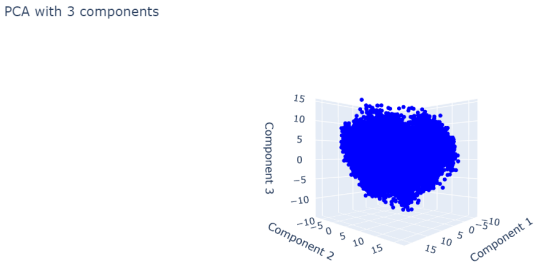
3. Sentence-BERT (SBERT): SBERT was chosen as the preferred method for text embedding due to its ability to preserve semantic relationships between sentences. SBERT fine-tunes the pre-trained BERT model specifically for sentence-level embeddings, making it suitable for tasks requiring semantic similarity.

Dimensionality Reduction Techniques:

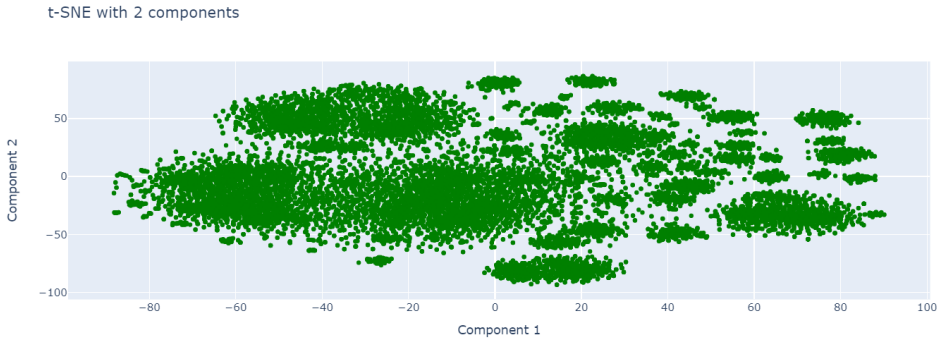
1. Principal Component Analysis (PCA): PCA was applied to reduce the dimensionality of the text embeddings while preserving the variance in the data. PCA transforms the high-

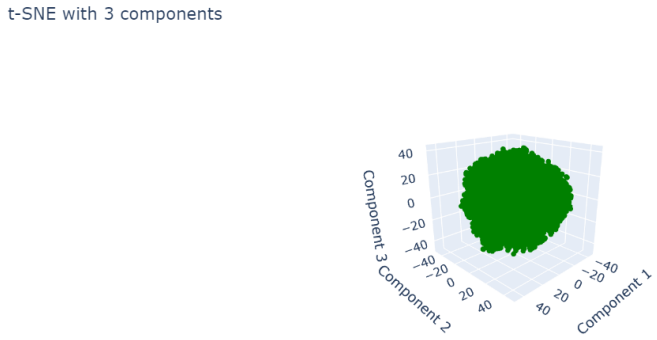
dimensional data into a lower-dimensional space while retaining as much variance as possible.



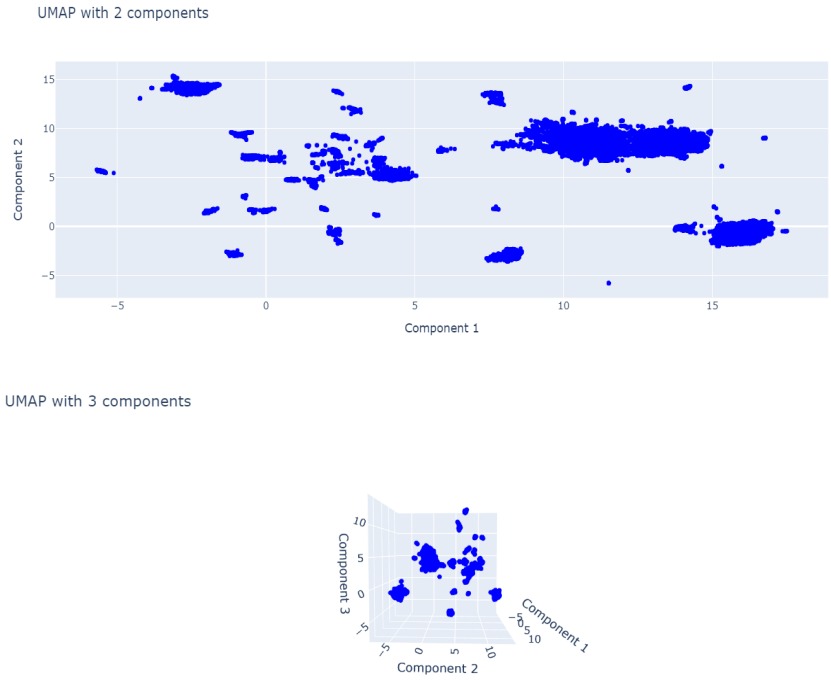


2. t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE is a nonlinear dimensionality reduction technique that is particularly effective for visualizing high-dimensional data in lower-dimensional space. It emphasizes preserving local structures and clustering tendencies in the data.



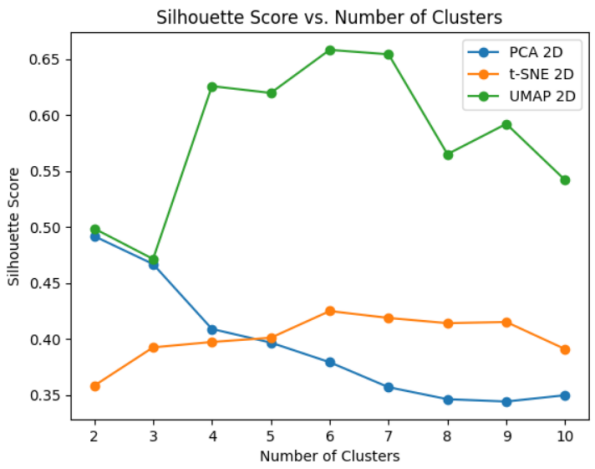


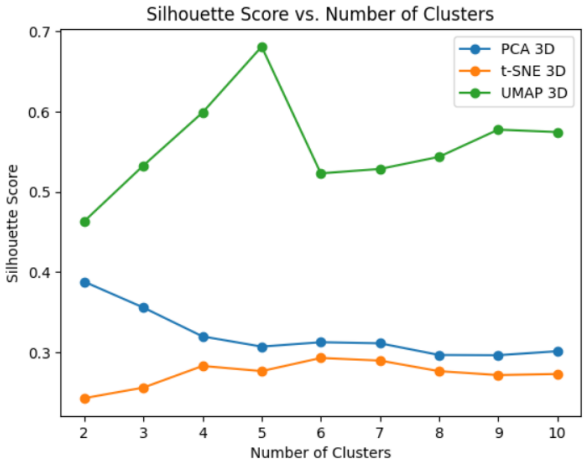
3. Uniform Manifold Approximation and Projection (UMAP): UMAP was chosen as the preferred dimensionality reduction technique for its ability to capture complex global structures in the data while preserving local and global relationships. UMAP provides a more interpretable visualization of high-dimensional data compared to t-SNE.



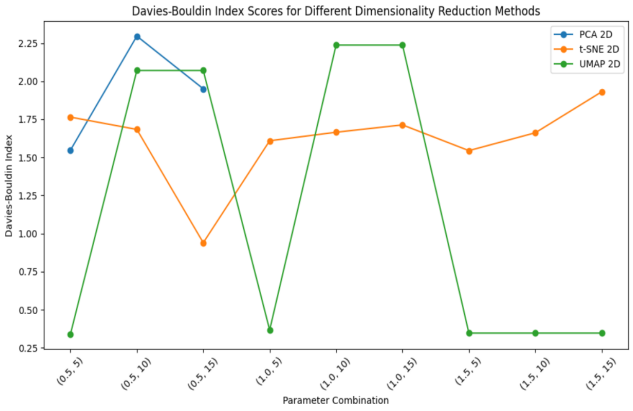
Clustering Algorithms:

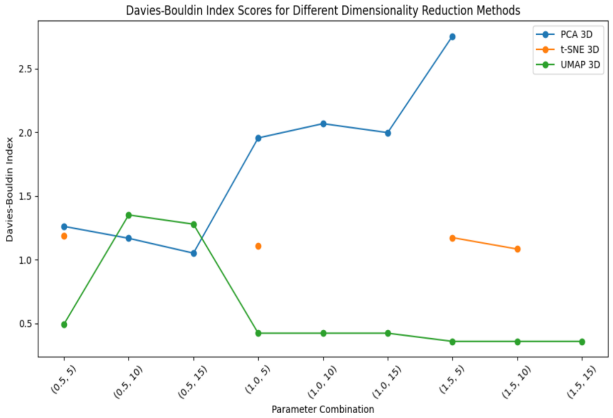
1. K-Means Clustering: K-Means clustering was applied to the reduced-dimensional embeddings to partition the data into clusters based on similarity. K-Means is a popular unsupervised clustering algorithm that assigns each data point to the nearest centroid, iteratively optimizing cluster centroids to minimize the within-cluster variance.



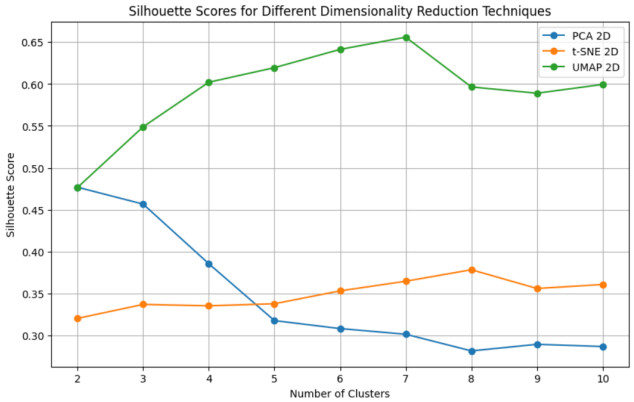


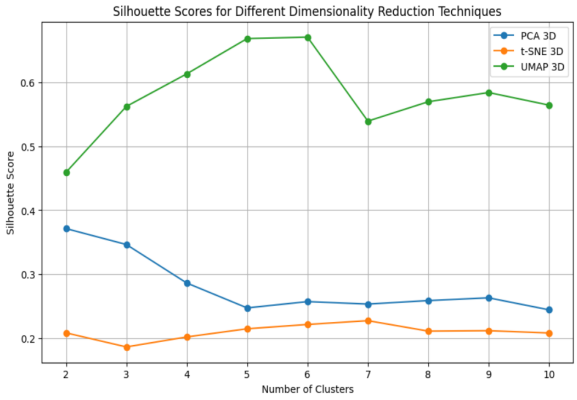
2. DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was utilized to identify clusters of varying shapes and densities in the data. DBSCAN groups together closely packed points as core samples and identifies outliers as noise points.





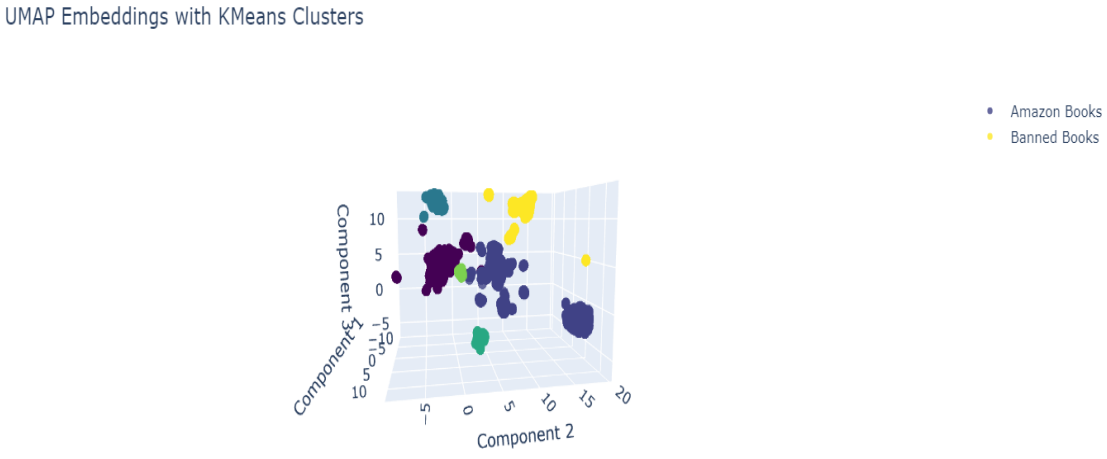
3. Agglomerative Clustering: Agglomerative clustering is a hierarchical clustering method that recursively merges pairs of clusters based on a linkage criterion, such as distance or similarity. It creates a dendrogram to visualize the hierarchical structure of the data.

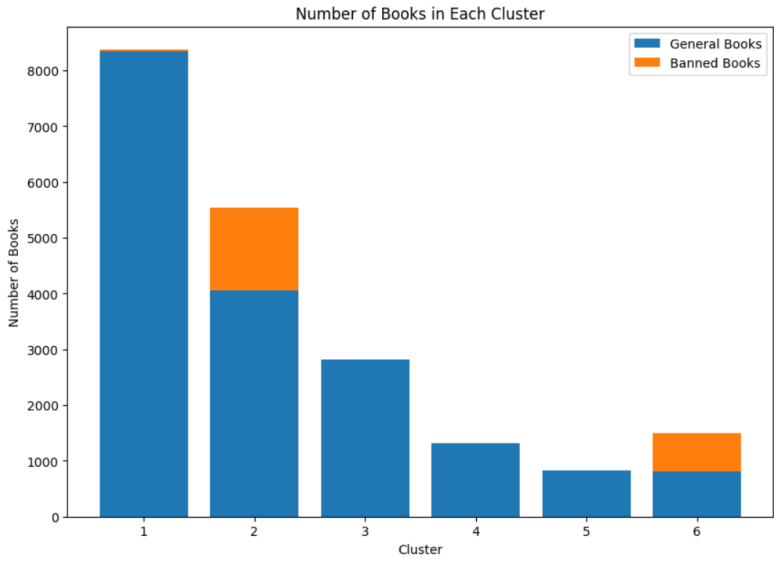




Final Model:

The final model for clustering similar books consisted of applying UMAP dimensionality reduction in 3D followed by K-Means clustering. UMAP effectively captured the complex global structures in the data, while K-Means efficiently partitioned the reduced-dimensional space into clusters based on similarity.





User Similarity Analysis:

To identify similar users interested in the same book clusters, Singular Value Decomposition (SVD) was applied to the combined embeddings of both books and users. SVD reduced the dimensionality of the combined embeddings and revealed latent factors representing user preferences. Similar users interested in the same book clusters were identified based on their latent factors.

Recommendation System:

Based on the book clusters and user preferences inferred from SVD, the system displayed book titles that aligned with a particular user's interests. Users were recommended books from clusters with similar characteristics to those preferred by the user, enhancing personalized recommendations.

Results

Recommended Books: (['A9722', 'A5513', 'A11674'], ['B966', 'B193', 'B347'], 'Success')

- Regular titles: ['how to stop worrying and start living', 'the faeries oracle', 'outwitting writers block and other problems of the pen']
- Banned titles: ['were i not a girl the inspiring and true story of dr james barry', 'girl talk the ultimate body puberty book for girls', 'what to do when im gone a mother's wisdom to her daughter']

The average silhouette score for the Amazon dataset is: 0.69495416

The average silhouette score for the Banned dataset is: 0.7663664

The banned book recommendation system was evaluated using silhouette analysis to measure the quality of the clusters formed. The silhouette score ranges from -1 to 1, where a higher value indicates better-defined clusters.

For the Amazon book dataset, the average silhouette score achieved was 0.69495416. This score suggests the clustering model reasonably separated the Amazon books into distinct groups based on their textual content and metadata features. However, there is still room for improvement in terms of cluster compactness and separation.

On the other hand, the banned book dataset demonstrated better clustering performance with an average silhouette score of 0.7663664. This higher score indicates that the banned books were more effectively grouped into well-defined and separated clusters, reflecting the distinct nature of these often-challenged literary works.

Visual representations of the clustering results are provided above, showcasing the distribution of books within the identified clusters for each dataset.

The SVD algorithm was evaluated for collaborative filtering using 5-fold cross-validation. The best root mean squared error (RMSE) achieved on the test set was 0.8219, with the optimal hyperparameters being 35 epochs, a learning rate of 0.06, and a regularization parameter of 0.01. Across the 5 folds, the mean RMSE was 0.8139, with a standard deviation of 0.0025. The average fit time (training time) across folds was 17.70 seconds, with a standard deviation of 2.70 seconds, while the average test time was 1.38 seconds, with a standard deviation of 0.27 seconds. These results indicate that the SVD algorithm performed reasonably well in predicting user ratings for the collaborative filtering task, with a relatively low RMSE and acceptable computational times. However, there may be room for improvement by tuning the hyperparameters further or exploring alternative algorithms.

```
Best RMSE: 0.8219457576020797
Best parameters: {'n_epochs': 35, 'lr_all': 0.06, 'reg_all': 0.01}
```

```
Evaluating RMSE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.8178  0.8140  0.8143  0.8100  0.8134  0.8139  0.0025
Fit time          15.58   17.63   16.16   16.18   22.93   17.70   2.70
Test time         1.41    1.19    1.21    1.19    1.88    1.38    0.27
Mean RMSE (5-fold CV): 0.8139089477928725
```

Discussion

Our project aimed to develop a recommender system that could suggest banned or challenged books to users based on their preferences for regular literature. By providing access to these often-censored works, we sought to empower stakeholders to explore diverse perspectives and exercise their right to free expression, fostering a more informed and democratic society.

While our recommendation system could generate personalized suggestions, the limited size and scope of our banned book dataset posed challenges in creating larger, more diverse clusters. This restricted the variety of recommendations we could provide, potentially hindering our ability to fully address the stakeholder needs for access to a broader range of banned or challenged literature.

To further improve our model and better meet the stakeholder needs, several enhancements could be explored:

1. Data Expansion: Expanding the banned book dataset by collecting data from additional sources and across a wider range of categories could increase the diversity and representativeness of the banned book clusters. This would enable more comprehensive and varied recommendations.

2. Incorporating Additional Features: Integrating additional features beyond textual data, such as book ratings, popularity metrics, or metadata, could enrich the clustering process and improve the recommendation system's performance.

3. User Interaction and Feedback: Implementing a user interface with feedback mechanisms would allow users to provide input on the recommendations, enabling the system to learn and adapt over time, better aligning with individual preferences.

By addressing these potential improvements, our recommender system could better cater to the diverse needs of stakeholders, providing them with access to a broader range of

banned or challenged literature while encouraging the exploration of diverse perspectives and promoting intellectual freedom.

Limitations

Our banned books dataset was narrowed to very particular categories, which made it difficult to create large clusters that effected in recommending books.

Future work

We look forward to collecting more banned book data, that will help in recommending books to the readers.

Furthermore, we are going to create a user interface for stakeholders; interested in being recommended books according to their liking.