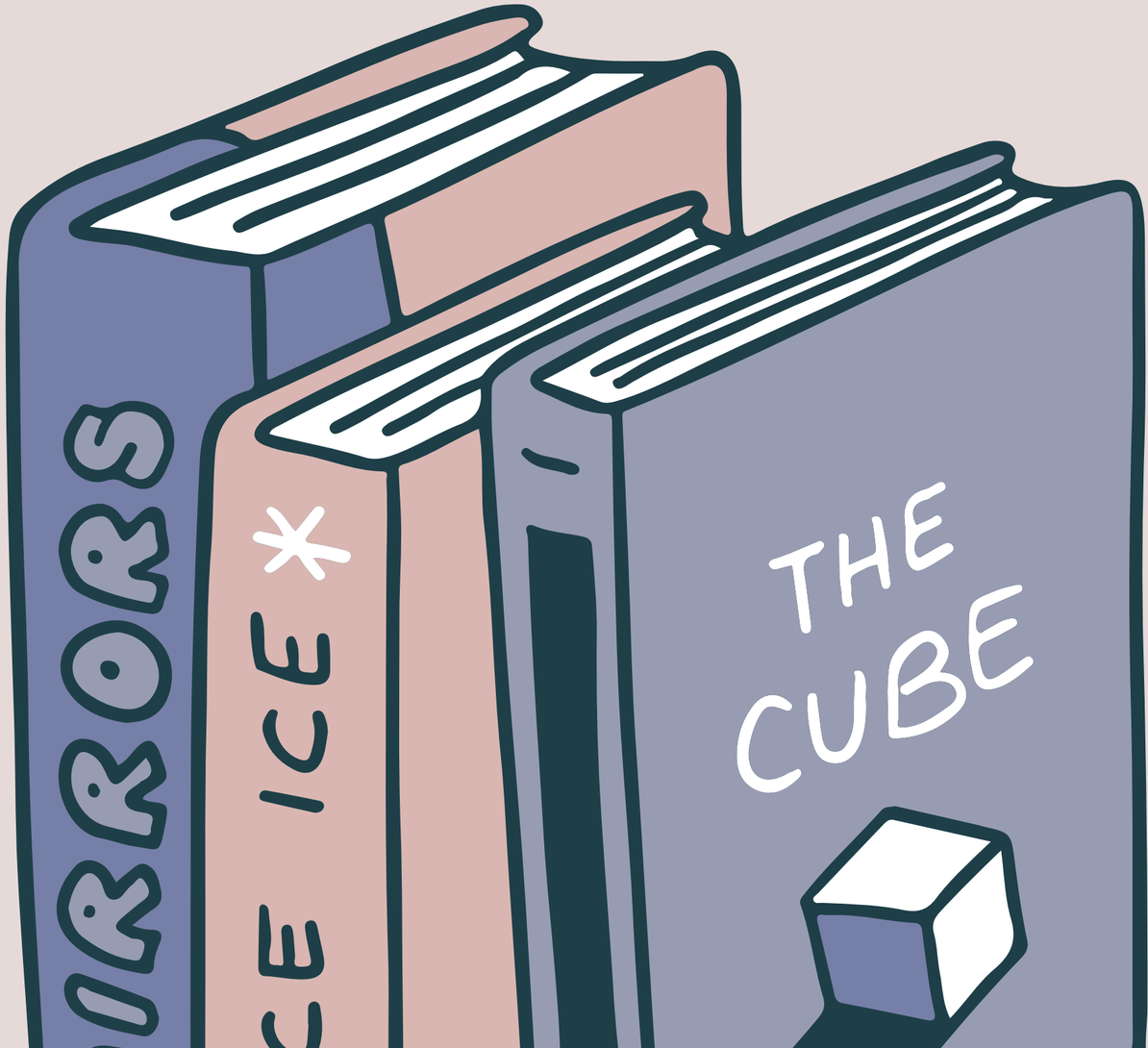
Machine Learning Project

Book Recommendation System



Beatriz Veiga

Project Overview:
Objectives and Potential Impact

Help users discover books related to History and Biography.

Suggest books that are thematically or contextually similar to the description provided by the user.

This system relies more on the user's interests by capturing deeper connections.

Data Selection

History and Biography

Books

ISBN, Format, Average Rating Publication Year, Book ID Rating Count, Title, Description, Authors, etc.

+

Authors

Average Rating, Author's ID, Author's Name, etc.

=

Final DF

- ISBN
- Book ID
- Title
- Description
- Average Rating
- Author's Name
- Average of Ratings per Author
- Publication Year
- Format

Data Cleaning & Preprocessing

- Removed missing/irrelevant data
- Merged relevant data
- Generated sentence embeddings
- Applied NLP techniques

NLP Techniques

Tokenization

+

Stop word Removal

+

Stemming

Lemmatization

Feature Engineering

- Transformed book descriptions using SBERT embeddings with a pretrained model - all-mpnet-base-v2 (best performance although a bit slow)

- Used BERTopic to cluster books into meaningful topics using UMAP for dimensionality and HDBSCAN for clustering

Model Building and Evaluation

Model Building

- NLP
- Bert
- Cosine Similarity

- NLP
- SBert
- BERTopic
- Cosine Similarity

Model Evaluation

- Similarity-Based Evaluation

Model Comparison

NLP + SBERT + BERTopic + Cosine Similarity

Title	Author	Similarity	Topic
Liberty's Daughters: The Revolutionary Experience of American Women, 1750-1800	Mary Beth Norton	0.65	21
Lives of Extraordinary Women: Rulers, Rebels	Kathleen Krull	0.64	21
The Most Powerful Women in the Middle Ages: Queens, Saints, and Viking Slayers, From Empress Theodora to Elizabeth of Tudor	Michael Rank	0.62	21
Clarina Nichols: Frontier Crusader for Women's Rights	Diane Eickhoff	0.61	21
Woman in the Nineteenth Century	Margaret Fuller	0.6	21

SBERT + BERTopic + Cosine Similarity

Title	Author	Similarity	Topic
The Reading List: Literature, Love and Back Again	Leslie Shimotakahara	0.38	49
James Tiptree, Jr.: The Double Life of Alice B. Sheldon	Julie Phillips	0.37	49
Classics: Why and How We Can Encourage Children to Read Them	Fiza Pathan	0.36	49
What There Is to Say We Have Said: The Correspondence of Eudora Welty and William Maxwell	Suzanne Marrs	0.35	49
The Color Purple	Alice Walker	0.34	49

Hyperparameter Tuning and Model Optimization

UMAP (Uniform Manifold Approximation and Projection)

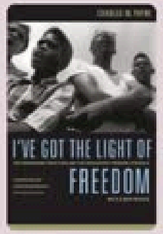
Reduced embedding complexity while preserving meaningful relationships

HDBSCAN (Hierarchical Density-Based Clustering)

Grouped books into topics based on description embeddings, allowing for better-structured recommendations

📖 Recommended Books:

I've Got the Light of Freedom: The Organizing Tradition and the Mississippi Freedom Struggle



I've Got the Light of Freedom: The Organizing Tradition and the Mississippi Freedom Struggle

🌟 4.26

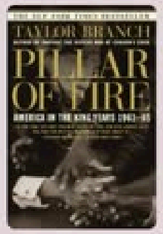
🖋️ Charles M. Payne

📖 Paperback

📊 Similarity: 0.62

This momentous work offers a groundbreaking history of the early civil rights movement in the South with new material that situates the book in the context of subsequent movement literature.

Pillar of Fire: America in the King Years 1963-65



Pillar of Fire: America in the King Years 1963-65

Real-World Application and Impact

APPLICATION:

- Can be extended to other domains, such as academic paper recommendations or movie suggestions
- Helps users find books based on content rather than just popularity, expanding their reading choices

ETHICAL CONSIDERATIONS & LIMITATIONS:

- The system does not incorporate user-specific preferences beyond book content
- New books with limited descriptions may not be well-represented in recommendations

Challenges and Learnings

Challenges Faced:

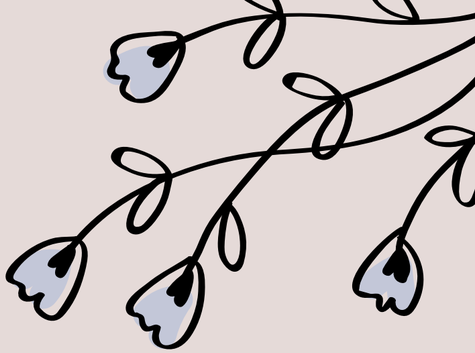
- Choosing the right model
- Combining NLP techniques and models
- Time management and process integration

Key Learnings:

- Combining different models improves recommendation quality
- Model selection is context-dependent
- Balancing exploration and focus

Future Work and Improvements

- Incorporate user ratings and feedback to improve personalization
- Introduce multi-modal recommendations by integrating metadata and user behavior
- Generate book recommendations beyond the constraints of a predefined dataframe



Thank you!