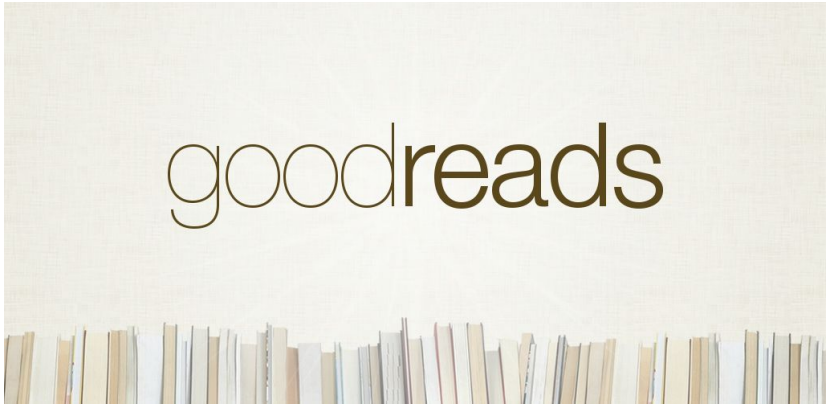
Goodreaders: A Book Recommender System

Catherine Johnston
Yevgeniya “Jonah” Tarasova

Motivation

- **Problem**: The large number of book choices presents a challenge for readers. Reliance on personal recommendations is often insufficient.
- **Solution**: This project developed a systematic book recommendation system.
- **Data Source**: Goodreads data was utilized for system development.

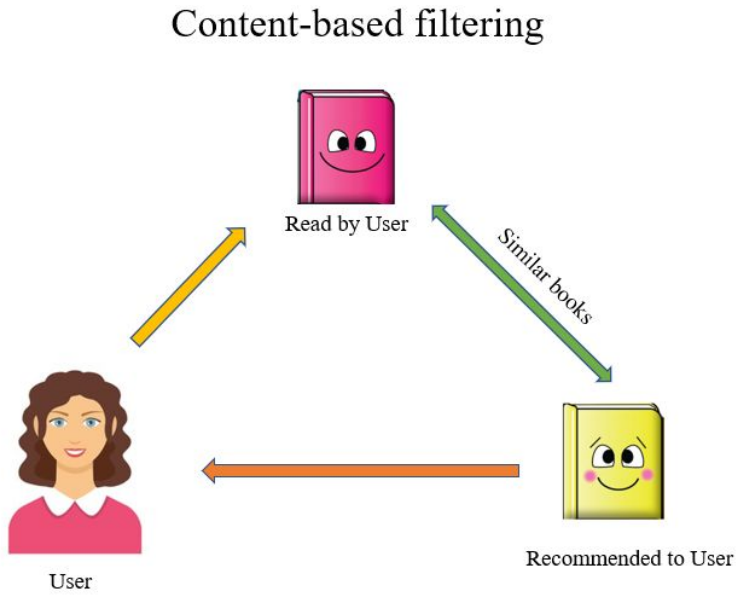


Stakeholders, KPIs, and Data Overview

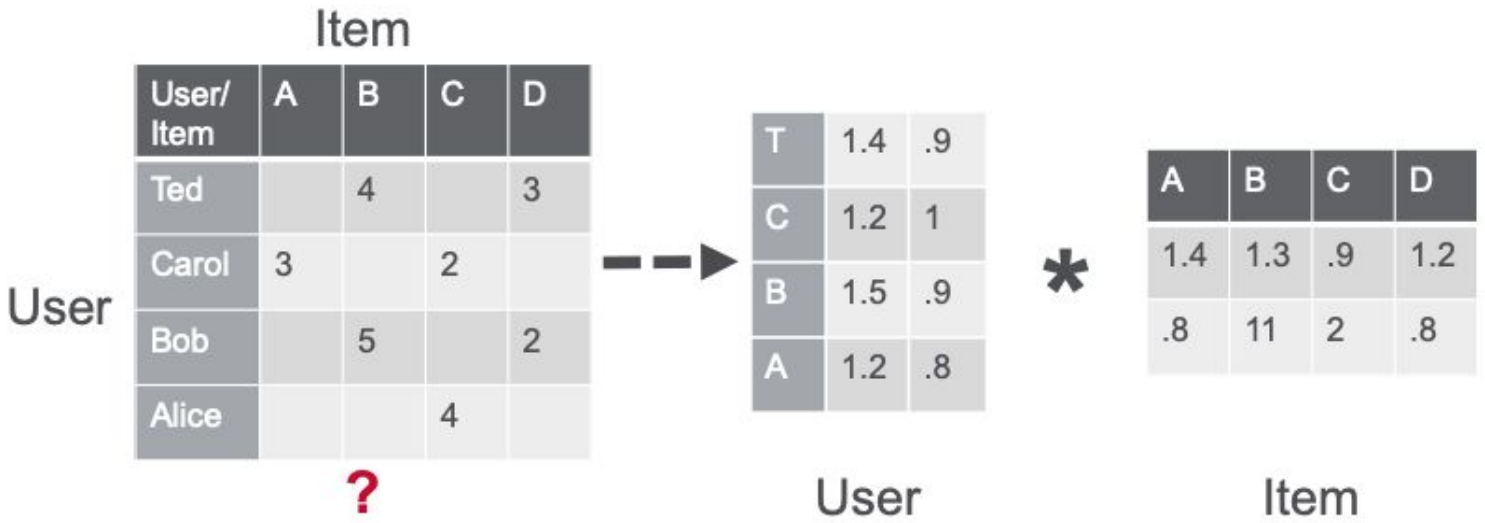
- **Stakeholders:** Readers, Publishers, Booksellers, Libraries, Community Book Providers.
- **Key Performance Indicator (KPI):** "Hits"
  - A "hit" represents a recommended item positively rated by the user, on which the model had not been trained.
- **Data Collection:**
  - Books from Goodreads' yearly Readers' Favorite Books lists (2011-2024).
  - Random sampling of user reviews.
  - Relevant information collected: title, author, description, genres, individual and average ratings, number of ratings.
  - **Dataset Size**: Approximately 71,000 cleaned books.

Modeling Approach (Content-Based Filtering)

- **Content-Based Filtering**:
  - Books are represented as feature vectors (genres, keywords).
  - Similarity is measured using cosine similarity.
  - User preferences are computed via a weighted average of these similarities.
- **Keyword Generation**:
  - Step 1: KeyBERT extracts keywords directly from book descriptions.
  - Step 2: Sentence-BERT embeddings cluster similar keywords semantically.

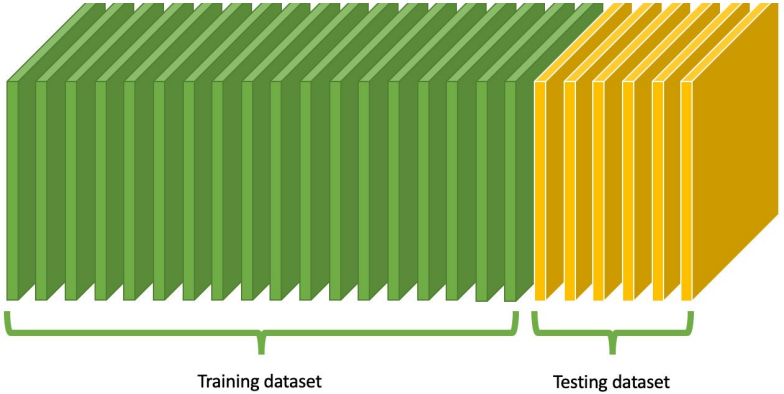


Modeling Approach (Collaborative Filtering)

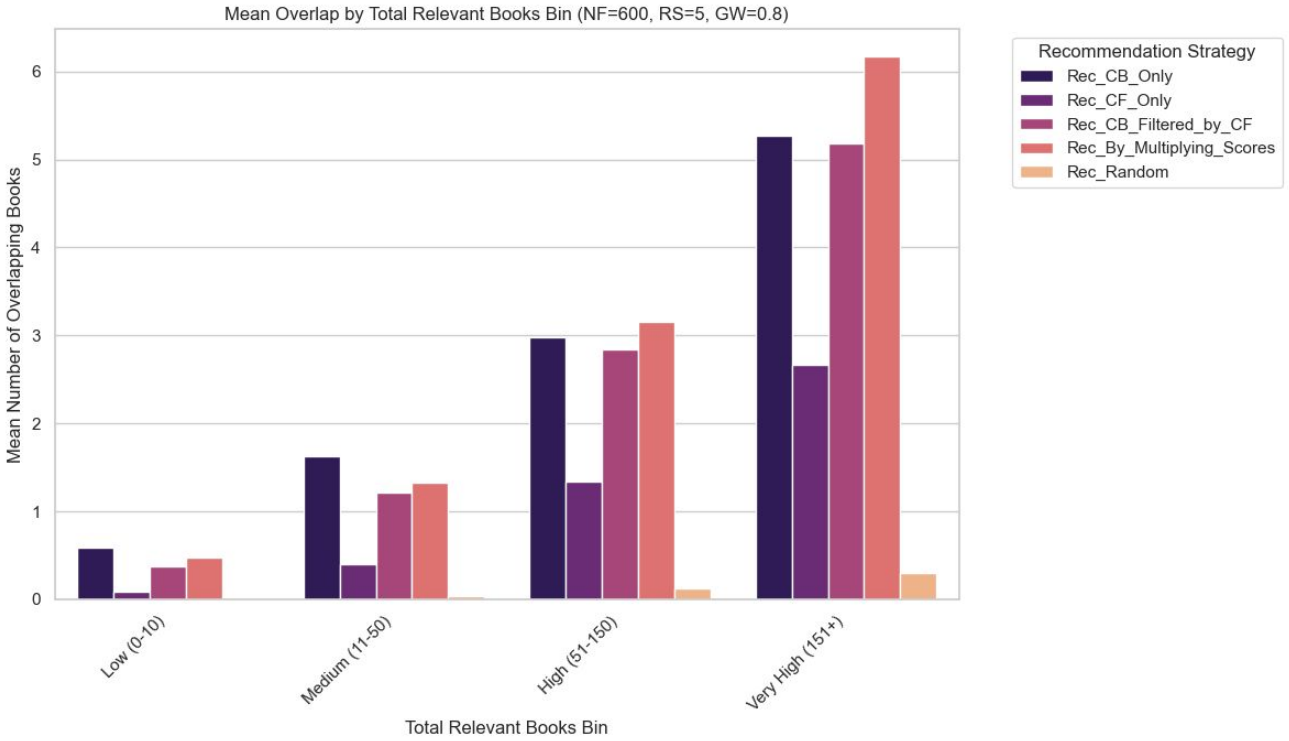


Data Preparation and Evaluation

- Book data was divided into training and testing sets.
- User ratings were split into training, cross-validation, and testing sets.
- For evaluation, half of each user's ratings simulated a 'new user' scenario, with the other half remaining as unseen data.



Results


Mean Overlap by Total Relevant Books Bin (NF=600, RS=5, GW=0.8)
Mean Number of Overlapping Books
0
1
2
3
4
5
6
Low (0-10)
Medium (11-50)
High (51-150)
Very High (151+)
Total Relevant Books Bin
Recommendation Strategy
Rec_CB_Only
Rec_CF_Only
Rec_CB_Filtered_by_CF
Rec_By_Multiplying_Scores
Rec_Random

Streamlit App

Your Books & Ratings

Select a Book You've Read

The Fifth Season by N.K. Jemisin

Your Rating (1-5)

5

1 5

Add Book

Added 'The Fifth Season by N.K. Jemisin' with rating 5!

Books You've Added:

Gideon the Ninth by Tamsyn Muir

Rating

5

1 5

Remove

The Fifth Season by N.K. Jemisin

Rating

5

1 5

Remove

Generate Recommendations

Get Recommendations

✨ Your Personalized Recommendations:

- An Unkindness of Ghosts by Rivers Solomon
- Nona the Ninth by Tamsyn Muir
- The World We Make by N.K. Jemisin
- A Psalm for the Wild-Built by Becky Chambers
- Light Bringer by Pierce Brown

Future Work

- **System Refinement**: Finer parameter tuning and dataset expansion.
- **Addressing Duplicate Books**: Detection of various book versions (e.g., translations) and filtering by user language preference.
- **Optimizing K-modes Initialization**: Explore robust initialization methods (e.g., Huang [1997, 1998], Cao et al. [2009]).
- **Including Additional Features**: Incorporate popularity, engagement metrics (reviews-to-ratings ratio), publication year, page count, or author name.
- **Streamlit Application Finalization**:
    - Models to be trained on full datasets.
    - Application optimization for user-friendliness.
    - Implementation of user accounts for data storage to facilitate further app fine-tuning.


A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining

Zhexue Huang*