Unlocking Your Next Read

Book Recommendation System With Goodreads Reviews

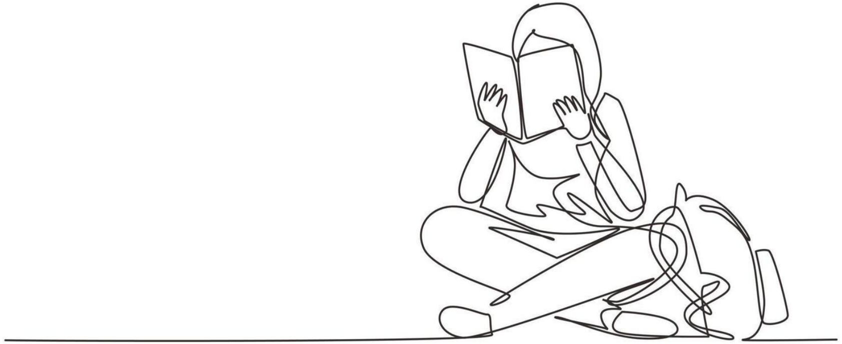
goodreads

A New Reading Journey

Have you ever gotten tired of the generic "top-rated" lists? You may have specific elements you are looking for in a book like "found family tropes" and "dark academia settings"

Our NLP-driven system analyzes readers’ reviews, identifying these nuanced preferences.

Instead of recommending another popular series, it suggests lesser-known titles where reviewers specifically praise similar intricate magic systems or strong "found family" dynamics, leading one to discover a book they truly love. This deepens engagement and fosters a more satisfying reading experience.



Enhancing Book Discovery with NLP

While Goodreads excels in user reviews, its current recommendation system primarily relies on ratings and collaborative filtering. This overlooks the rich textual data within reviews, leading to a limited recommendation experience for users. Our solution leverages Natural Language Processing (NLP) to unlock a deeper level of personalization.

The Core Problem

- Users face choice overload with generic recommendations.
- Valuable textual signals (themes, sentiment) in reviews are underutilized.

Our Goal

- Develop a context-aware system analyzing Goodreads reviews.
- Recommend books based on reader preferences and language nuances.

Expected Impact

- Help readers discover books aligning with specific interests.
- Offer nuanced recommendations for platforms like Goodreads.
- Promote niche books with strong appeal.

Our Data Foundation

Our system is built on a robust dataset of Young Adult reviews from Goodreads, providing a rich source of textual insights for analysis

Data

- **Source:** Publicly available from UCSD website (Goodreads Dataset)
- **Format**: JSON
- **Size**: **~3GB**
- **Description**: Covers **~93K books** and **~2.4MM reviews** (from reviews.json and books.json)

Processing

- **Merging**: Reviews and book metadata merged on book_id
- **Tools**: Python (Pandas, spaCy, Transformers) for efficient processing

Key Metadata (Young Adult)

- review.json
  - 'book_id', 'review_text', 'rating', 'date_added', 'read_at', 'started_at', 'user_id', 'n_votes', 'n_comments'
- book.json
  - 'book_id', 'title', 'title_without_series', 'authors',m'publisher', 'average_rating', 'ratings_count', 'text_reviews_count', 'description', 'popular_shelves', 'similar_books', 'language_code', 'format'

Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys18. [bibtex]
Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL19. [bibtex]

goodreads

Our Design Ecosystem

Data Acquisition & Cleaning

- Data preparation
- Language filtering
- Text preprocessing
- Deduplication

Topic & Aspect Discovery

- Topic modeling with BERTopic
- Identify key aspects for deeper analysis

Analyze & Recommend

- Use ABSA to apply sentiment with aspects
- Extract keywords with TF-IDF
- Recommend books via cosine similarity

Evaluation & Demo

- Evaluate recommendation and accuracy of the model
- Deploy interactive UI in Streamlit
- Plan for next steps

Design Structure

1. Data Processing + Cleaning

- Inspect & combine datasets
- Clean, filter, & normalize text
  - Keep **English reviews** only
  - Convert to lowercase
  - **Tokenization + lemmatization**
  - Remove stopwords & punctuation
- Deduplication (**SimHash + LSH**)
  - Remove near-duplicate reviews

2. Topic Modeling (SBERT embeddings + BERTopic)

- Obtain semantically meaningful sentence embeddings using **SBERT**
- Apply **BERTopic** to cluster reviews into topics
  - Extract topics, keywords, topic probabilities
  - Identify top words per topic

3. Aspect-Based Sentiment Analysis (ABSA)

- Select aspects from BERTopic keywords
- **Extract aspect** mentions in reviews
- Assign sentiment scores (pos/neg/neutral) per aspect with textblob

4. Similarity Search (TF-IDF + Cosine Similarity)

- Extract top keywords per review using **TF-IDF**
- Convert reviews to TF-IDF vectors
- Compute **cosine similarity** between query vector and all review vectors
- Rank & return books with most similar reviews to the user's query

Design Choices & Methodologies Rationale

Method	**Why we chose it**
Preprocessing Tokenization + Lemmatization	Removes morphological noise so TF-IDF & embeddings treat “loved” ≈ “love”, improving both duplicate detection and topic purity
Deduplication SimHash + LSH	Goodreads reviews often get copy-pasted; cutting near-dupes speeds topic modelling and prevents popularity bias in recommendation scores
Embeddings & Topic Modeling (SBERT + BERTopic)	Uncovers nuanced themes and sub-genres within Goodreads reviews that traditional metadata or keyword-only models would miss, enhancing content understanding for personalized recommendations
Aspect-Based Sentiment Analysis (ABSA) using TextBlob scores (-1 to 1)	By aligning aspect sentiment with topics extracted from BERTopic, the model tailors book suggestions to both thematic interests and emotional preferences expressed in reviews
TF-IDF Keyword Extraction	Supplies quick, transparent book descriptors and helps align topics with human-readable tags
Cosine Similarity	Performs content-based filtering using TF-IDF or embeddings to compare book reviews, giving good suggestions even for new or less-popular books without many ratings

Individual System Analysis

Data Quality (SimHash Deduplication)

- Hamming Distance = **10**
- **Similarity Threshold**
  - 1 - 10/64 = **84.4% similarity** indicating strong semantic alignment
  - **Balance:** 10 provides strict deduplication without over-filtering
- **219 reviews (4.38%)** identified as near-duplicates and removed → Reviews with ≥84.4% bit-level similarity are considered duplicates, **effectively catching spam and copy-paste reviews** while preserving legitimate variations

Topic Modeling (BERTopic)

- **98.0%** Coverage Rate **GOOD**
- **3 topics** discovered automatically
- **4,468** reviews successfully clustered
- **2.0%** noise rate

ABSA (Aspect-Based Sentiment Analysis)

- **74.9%** Coverage Rate **GOOD**
- **3,414/4,559** reviews analyzed

Aspect	Coverage	Avg Sentiment	Reviews
characters	32.9%	0.243 (Positive)	1499
story_plot	64.8%	0.264 (Positive)	2956
writing_style	29.7%	0.234 (Positive)	1354
paranormal_romance	13.9%	0.272 (Positive)	635
comparisons	34.6%	0.220 (Positive)	1578
adventure_mythology	5.9%	0.221 (Positive)	269
series_context	61.2%	0.263 (Positive)	2789
pacing_engagement	46.4%	0.254 (Positive)	2116
emotional_themes	50.4%	0.270 (Positive)	2300
nostalgia_connection	6.4%	0.201 (Positive)	293
reading_experience	50.1%	0.250 (Positive)	2283

Cosine Similarity & Evaluation Performance

- **Up to 0.78** score
- **Peak similarity:** 0.475 (Romance: 100% precision)
- **Query-Performance Correlation**
  - **High similarity (0.4+)**
  - **Moderate similarity (0.1-0.3)**
  - **Semantic Discrimination:** >8 relevant books per query

goodreads

Understanding Recommendation Quality

Experimental Design
Training 500 books, 5K reviews → Testing: 100 unseen books, 800 reviews → Sample: 5 queries x 10 recommendation = 50 evaluation points

Precision@k

P@5: 56%
P@10: 56% GOOD

→ Out of 5 books recommended, 3 books are actually relevant to the user's query

→ Shows how accurate our recommendations are; higher precision means users get more relevant suggestions

Recall@k

R@5: 55.8%
R@10: 100% EXCELLENT

→ Our system finds 56-100% of all relevant books that exist for a query

→ Shows how well we discover relevant content; higher recall means we don't miss good recommendations

F1 Score

F1@5: 51.9%
F1@10: 68.9% GOOD

→ Balanced measure combining precision and recall, shows overall recommendation quality

→ Gives a single number to compare different recommendation approaches

Hit Rate@5

Hit Rate@5: 100%
Hit Rate@10: 100% EXCELLENT

→ 5 out of 5 queries get at least one relevant book in the top 5 recommendations

→ Shows user satisfaction - measures if users find anything useful at all

Findings & Next Steps

Evaluation Results

→ **56% precision** on completely unseen books demonstrates strong cross-dataset transfer
→ **100% hit rate** indicates robust query understanding across diverse information needs
→ **Perfect recall@10** suggests comprehensive coverage of relevant items

Query Type Analysis

Romance	Fantasy	Adventure	Quality	Popular
100% P@5	40% P@5	40% P@5	60% P@5	40% P@5

→ **Romance queries** got perfect precision, indicating effective semantic matching

Improvements

- **Enhance Model Accuracy**: Fine-tune similarity models and thresholds.
- **Refine Topics and Layers**: Build richer ground truth and integrate user interaction data for more realistic testing.
- **Improve ABSA Coverage**: Increase coverage beyond current 75% by refining aspect extraction and sentiment handling.

➔ **Next Steps/Scaling**

- **Expand Data Sources**: Incorporate multi-lingual reviews and additional metadata (genres, author networks, etc.) for broader coverage.
- **Implement Cloud Infrastructure**: Deploy scalable cloud-based architecture with GPU acceleration to handle larger datasets and enable faster processing at scale.

➔ **Conclusion**

- Develop a robust/scalable pipeline for data cleaning and semantic similarity search
- Position Goodreads for accuracy optimization/personalization/international market expansion
- Prepare to deploy application framework of delivering updated recommendations