Book Recommendation System:

A Text Analytics Project

By: Manuel Iglesias

Overview

Problem Statement:

- Many book readers struggle to find books similar to their current interests as many current book recommendation systems primarily focus on popularity, publisher deals, or user ratings

Research Goal:

- Develop a recommendation system that utilizes text analysis techniques to identify connections between books to cater to different user preferences

Users:

- ***Readers*** trying to find books that align to their interests and past reading history
- ***Librarians*** who may need to recommend guests books based on individual preferences
- ***Book retailers*** who may want to improve sales using a personalized recommendation system

Benefits:

- Enhanced personalization
- Increased reader satisfaction and reading habits
- Increased support for niche books and authors

Data Description

Data Title: Goodreads' Best Books Ever

Data Source: Kaggle - https://www.kaggle.com/datasets/meetnaren/goodreads-best-books

Dataset Description: 10,000,000 Books scraped from the GoodReads API. Fields include Id, Name, RatingDist1, RatingDist2, RatingDist3, RatingDist4, RatingDist5, pagesNumber, RatingDistTotal, PublishMonth, PublishDay, Publisher, CountsOfReview, PublishYear, Language, Authors, Rating, ISBN, Count of Text Reviews, and Description

Data Size: 2.37 GB

Data Time Period: N/A

Data Types:

Strings - (author, description, edition, format, title, genre(s), image url)

Numeric (pages, rating, rating counts, review count, isbn), etc

Data Size: 57,510 unique values

Data Statistics

Remaining Missing Values: None

Data Shape BEFORE Cleaning: 54,301 titles * 12 columns = 651,612 data points

Data Shape AFTER Cleaning: 22,715 unique titles * 8 columns = 181,720 data points

Data:

Numeric: Pages, Rating, Rating Count, Review Count
Text: Author, Title, Description, Genre(s)

Dropped Columns: ‘image_url', 'book_edition', 'book_format', 'book_isbn’

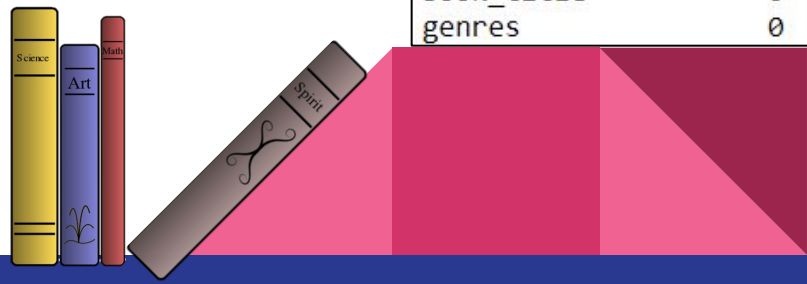
Key Data: ‘Features’ column consists of book descriptions, book authors, & genre

Unclean Data

book_authors	0
book_desc	1331
book_edition	48848
book_format	1656
book_isbn	12866
book_pages	2522
book_rating	0
book_rating_count	0
book_review_count	0
book_title	0
genres	3242
image_url	683

Clean Data

book_authors	0
book_desc	0
book_pages	0
book_rating	0
book_rating_count	0
book_review_count	0
book_title	0
genres	0



Data Engineering

Import:

- Import and display the data
- Clean data, remove non-English text with langdetect, and drop missing values and columns not necessary for analysis
- Join text columns to one “feature” column

Transformation:

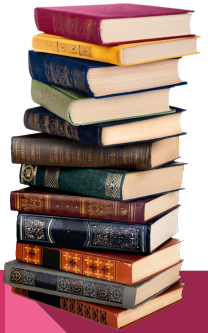
- Standardized text, tokenize, removed stop words, apply stemming and lemmatization
- Normalization of numerical values with sklearn MinMaxScaler

Text Analysis:

- Used TF-IDF to convert text features into vector form
- Apply concatenation to combine TF-IDF and with normalized numeric values
- Apply LSA to capture hidden topics
- Create cosine similarity matrix with sklearn cosine_similarity

Recommendation Feature:

- User to enter title of book to base recommendation off of
- Locate book titles with closest cosine similarities (display in descending order)
- Create a dataframe of top 10 recommended books displaying title, cosine similarity, author, genre, description, pages and rating

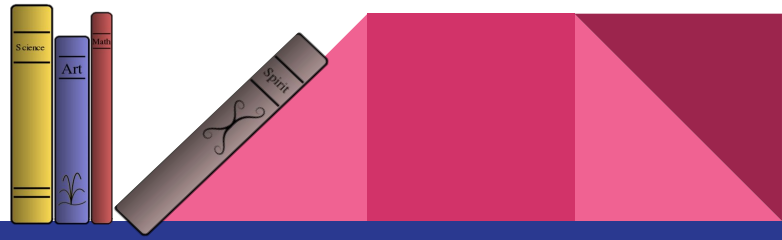


Cosine Similarity Matrix

book_title	The Hunger Games	Harry Potter and the Order of the Phoenix	To Kill a Mockingbird	Twilight	The Book Thief	The Chronicles of Narnia	Gone with the Wind	The Fault in Our Stars	The Hitchhiker's Guide to the Galaxy	The Giving Tree	...
book_title											
The Hunger Games	1.000000	0.802136	0.931796	0.936174	0.845972	0.579381	0.656026	0.955523	0.732517	0.653100	...
Harry Potter and the Order of the Phoenix	0.802136	1.000000	0.893784	0.820182	0.854885	0.888568	0.890608	0.832651	0.955187	0.921542	...
To Kill a Mockingbird	0.931796	0.893784	1.000000	0.928313	0.888616	0.750853	0.838280	0.926610	0.858010	0.814055	...
Twilight	0.936174	0.820182	0.928313	1.000000	0.817228	0.628872	0.690398	0.907677	0.749875	0.692174	...
The Book Thief	0.845972	0.854885	0.888616	0.817228	1.000000	0.764119	0.860092	0.909225	0.826378	0.816582	...
The Chronicles of Narnia	0.579381	0.888568	0.750853	0.628872	0.764119	1.000000	0.889056	0.666447	0.915962	0.926455	...
Gone with the Wind	0.656026	0.890608	0.838280	0.690398	0.860092	0.889056	1.000000	0.725482	0.918032	0.920639	...
The Fault in Our Stars	0.955523	0.832651	0.926610	0.907677	0.909225	0.666447	0.725482	1.000000	0.792194	0.732434	...
The Hitchhiker's Guide to the Galaxy	0.732517	0.955187	0.858010	0.749875	0.826378	0.915962	0.918032	0.792194	1.000000	0.944104	...
The Giving Tree	0.653100	0.921542	0.814055	0.692174	0.816582	0.926455	0.920639	0.732434	0.944104	1.000000	...

Cosine Similarity Matrix:

- In our system cosine similarity was used to quantify book descriptions and group similar books together for more accurate recommendations.
- A score closer to ‘1’ indicates that the books share common themes.
- Due to our large data set it created computational challenges. For usability and lack of processing power, the dataset was shorten.



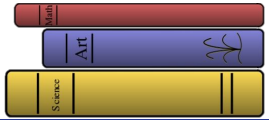
Recommendation 1:

```
recommended_books = recommend_books('The Hunger Games', cosine_similarity_df)
recommended_books
```

	Book Title	Similarity Score	book_authors	genres	book_desc	book_pages	book_rating
0	The Fault in Our Stars	0.955523	John Green	Young Adult, Fiction, Romance, Contemporary	Despite the tumor-shrinking medical miracle th...	313.0	4.24
1	Divergent	0.942101	Veronica Roth	Young Adult, Science Fiction, Dystopia, Fictio...	In Beatrice Prior's dystopian Chicago world, s...	487.0	4.22
2	Harry Potter and the Sorcerer's Stone	0.938962	J.K. Rowling|Mary GrandPré	Fantasy, Young Adult, Fiction	Harry Potter's life is miserable. His parents ...	320.0	4.46
3	Twilight	0.936174	Stephenie Meyer	Young Adult, Fantasy, Romance, Paranormal, Vam...	About three things I was absolutely positive.F...	498.0	3.58
4	Mockingjay	0.935855	Suzanne Collins	Young Adult, Science Fiction, Dystopia, Fictio...	My name is Katniss Everdeen.Why am I not dead?...	392.0	4.03
5	To Kill a Mockingbird	0.931796	Harper Lee	Classics, Fiction, Historical, Historical Fict...	The unforgettable novel of a childhood in a sl...	324.0	4.27
6	Catching Fire	0.926937	Suzanne Collins	Young Adult, Science Fiction, Dystopia, Fictio...	Sparks are igniting.Flames are spreading.And t...	391.0	4.29
7	Gone Girl	0.918955	Gillian Flynn	Fiction, Mystery, Thriller, Mystery, Crime	On a warm summer morning in North Carthage, Mi...	415.0	4.05
8	Nineteen Eighty-Four	0.890731	George Orwell|Thomas Pynchon	Classics, Fiction, Science Fiction, Science Fi...	Alternative cover edition can be found here.NO...	339.0	4.16
9	The Girl with the Dragon Tattoo	0.886953	Stieg Larsson|Reg Keeland	Fiction, Mystery, Thriller, Mystery, Crime	A spellbinding amalgam of murder mystery, fami...	465.0	4.13

1st Book Recommendation Output:

- Our system matches books to the users preference based on textual relevance and the similarity score, ensuring the recommendation is highly tailored to the uders reading history
- This recommendation Includes important information such as the title of the book, author, genre, and a book description.



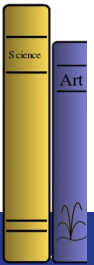
Recommendation 2:

```
recommended_books = recommend_books('The Giver', cosine_similarity_df)
recommended_books
```

	Book Title	Similarity Score	book_authors	genres	book_desc	book_pages	book_rating
0	The Maze Runner	0.954181	James Dashner	Young Adult, Science Fiction, Dystopia, Scienc...	There are alternate cover editions for this AS...	384.0	4.03
1	The Host	0.953100	Stephenie Meyer	Science Fiction, Fiction, Young Adult, Science...	Also see: Alternate Cover Editions for this IS...	620.0	3.84
2	Ready Player One	0.948409	Ernest Cline	Science Fiction, Fiction, Young Adult, Science...	In the year 2045, reality is an ugly place. Th...	374.0	4.29
3	Ready Player One. Movie Tie-In	0.948120	Ernest Cline	Science Fiction, Fiction, Young Adult, Science...	In the year 2044, reality is an ugly place. Th...	608.0	4.29
4	The Handmaid's Tale	0.947340	Margaret Atwood	Fiction, Classics, Science Fiction, Dystopia, ...	Offred is a Handmaid in the Republic of Gilead...	311.0	4.09
5	1984	0.938333	George Orwell|Erich Fromm	Classics, Fiction, Science Fiction, Science Fi...	Among the seminal texts of the 20th century, N...	328.0	4.16
6	Cinder	0.937243	Marissa Meyer	Young Adult, Fantasy, Science Fiction, Science...	Sixteen-year-old Cinder is considered a techno...	390.0	4.15
7	Nineteen Eighty-Four	0.935769	George Orwell|Thomas Pynchon	Classics, Fiction, Science Fiction, Science Fi...	Alternative cover edition can be found here.NO...	339.0	4.16
8	Uljas uusi maailma	0.932761	Aldous Huxley|l. H. Orras	Classics, Fiction, Science Fiction, Science Fi...	Brave New World is a dystopian novel written i...	263.0	3.98
9	The Selection	0.928520	Kiera Cass	Young Adult, Romance, Science Fiction, Dystopi...	For thirty-five girls, the Selection is the ch...	336.0	4.14

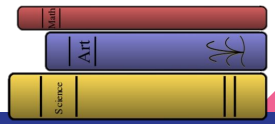
2nd Book Recommendation Output:

- This recommendation is an example of the limitations of our project:
  - Large but limited diversity within the data
    - Leads to lower similarity within recommendations
- Higher accuracy could be achieved either by restricting the dataset or expanding it (ex. Only U.S. created books in English or obtaining a library's giant catalogue)



Insights

- Challenge 1 - Removing non-english texts
  - Used library ‘langdetect’, however some non-english entries remained
    - Affected the quality of our data and the accuracy of our recommendations
- Challenge 2 - Large Data Set
  - Our dataset had 10 million books - presented scalability issues
    - Solution: shortening the dataset to accommodate lack of power processing issues
- ***Conclusion***: System successfully identified books with high textual similarity, demonstrating the effectiveness of our system’s capability to suggest books that closely match users preferences.



Next Steps

- Apply topic modeling
  - Create author profiles where users can see common themes across book descriptions and genre types
    - This could help users find new authors to explore based on common themes
- Perform sentiment analysis on reviews:
  - Scrape “Goodreads” website for book reviews
    - Performing a sentiment analysis on reviews for books and authors in general
      - Amplify our model by incorporating a rating restriction with the help of sentiment analysis


Thank
You!