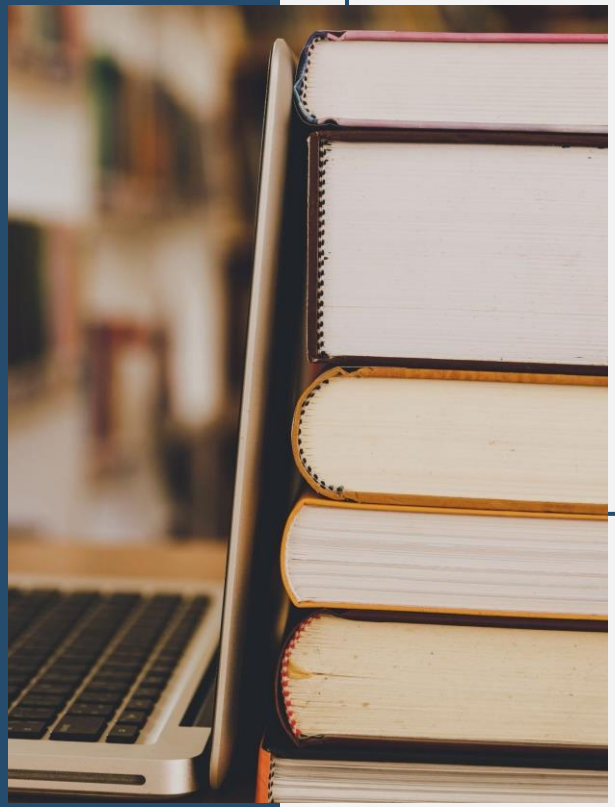


AN EXPLORATION OF ALGORITHMS AND PERSONALIZATION TECHNIQUES FOR EFFECTIVE BOOK RECOMMENDATIONS

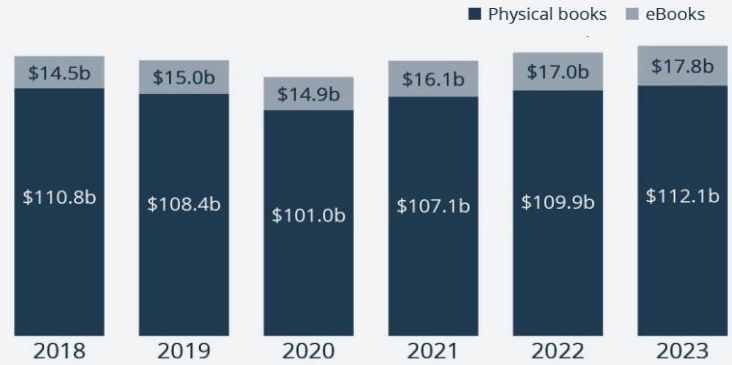
Joffin C Clement

CONTENTS

1. INTRODUCTION
2. OBJECTIVE
3. DATA UNDERSTANDING
4. DATA PREPROCESSING
5. EXPLORATORY DATA ANALYSIS
6. RECOMMENDER SYSTEMS
7. CONCLUSION

INTRODUCTION

Worldwide estimated revenue with eBooks and physical books



Source: Statista Advertising & Media Market Outlook

According to a survey conducted by Pew Research in 2019, the popularity of digital books is anticipated to rise over the coming years. The sales of e-book readers is increasing along with the popularity of e-books.

Need For Building An Online Book Recommender System

amazonkindle

goodreads

Enhancing conversion rate for e-book sellers

Readers can find books that suit their taste

OBJECTIVE

GOAL

Design book recommendation system by several methods and explore the strengths and weaknesses of each method

RESEARCH QUESTION

1. How can we recommend book for a new user?
2. How can we predict the rating that a user will give to books that they haven't read? How can we apply it to a recommendation system?
3. How can we calculate the similarity between books based on their content? How can we apply it to a recommendation system?

DATA UNDERSTANDING

We are using Book-Crossing dataset to train and test our recommendation system. Book-Crossings is a book ratings dataset compiled by Cai-Nicolas Ziegler, which is available in Kaggle.

Datasets:

- Users
- Book
- Ratings

2,78,858 users

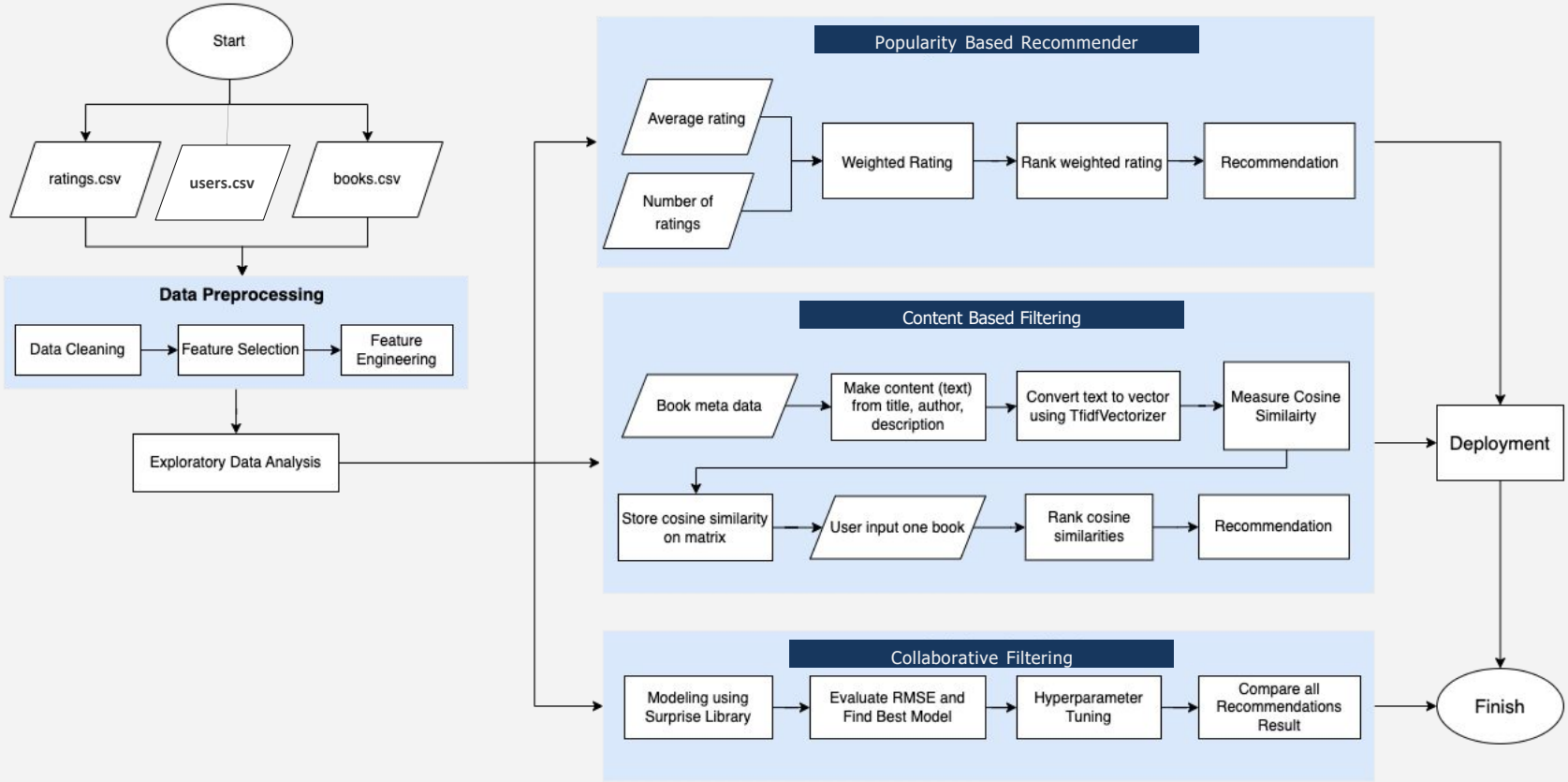
1.1M+ ratings

2,71,360 books

Metadata

1. Book
   - ISBN
   - Book Title
   - Book Author
   - Publication year
   - Publisher
   - Image url (3 cols)
2. Users
   - User Id
   - Location
   - Age
3. Ratings
   - User Id
   - ISBN
   - Book Rating

SYSTEM FLOWCHART



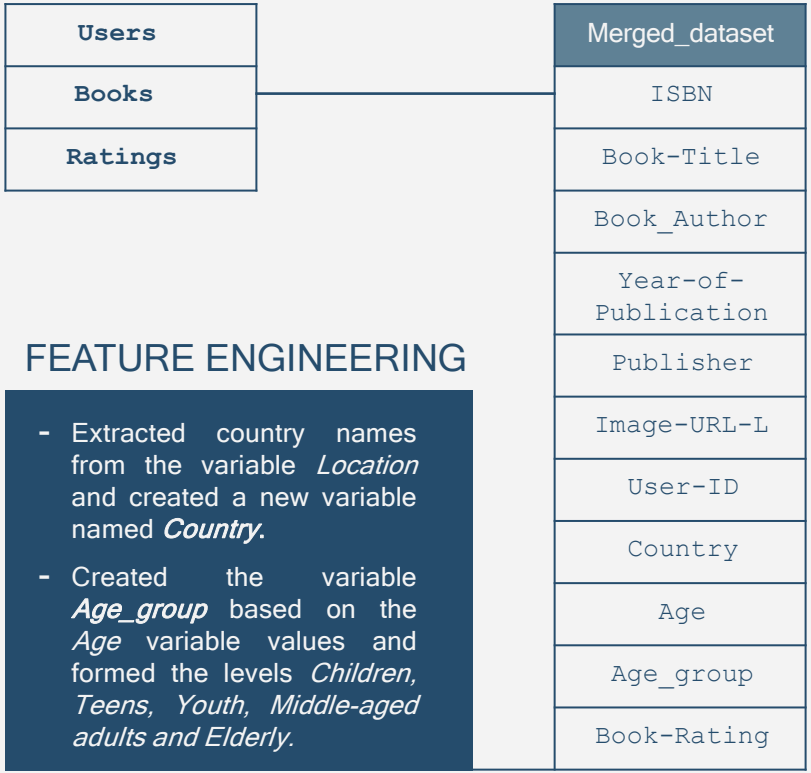
DATA PREPROCESSING

MISSING VALUE HANDLING

Column Name	percent_missing
Book-Author	0.000369
Publisher	0.000737
Image-URL-L	0.001106
Age	39.719857

- Irrelevant informations in the variables *Age* and *Year-of -Publication* considered as missing.
- Imputed *Age* variable by creating a normal distribution based on the mean age value and assigning randomly across the null points.
- Imputed *Book-Author, Publisher,* and *Year of Publication* by searching relevant information through internet.
- *Imputed Book-Rating* for each book using mode and dropped rest of missing observations.

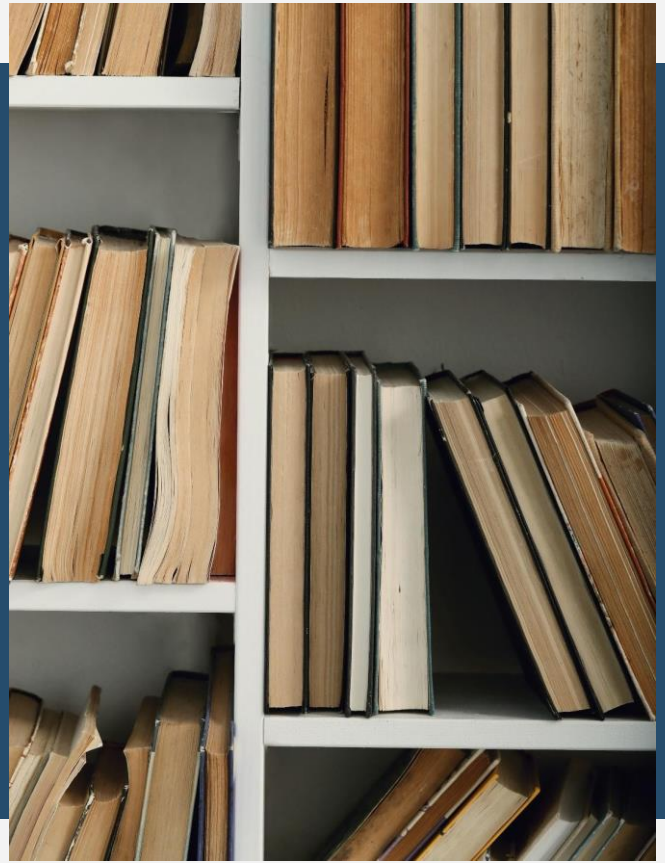
MERGING DATASETS



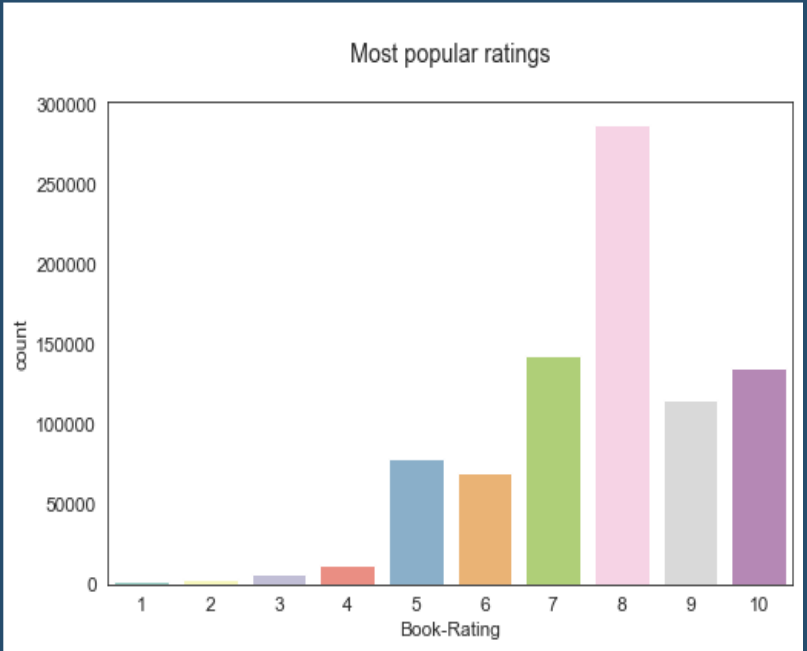
FEATURE ENGINEERING

- Extracted country names from the variable *Location* and created a new variable named ***Country***.
- Created the variable ***Age_group*** based on the *Age* variable values and formed the levels *Children, Teens, Youth, Middle-aged adults and Elderly.*

Exploratory Data Analysis

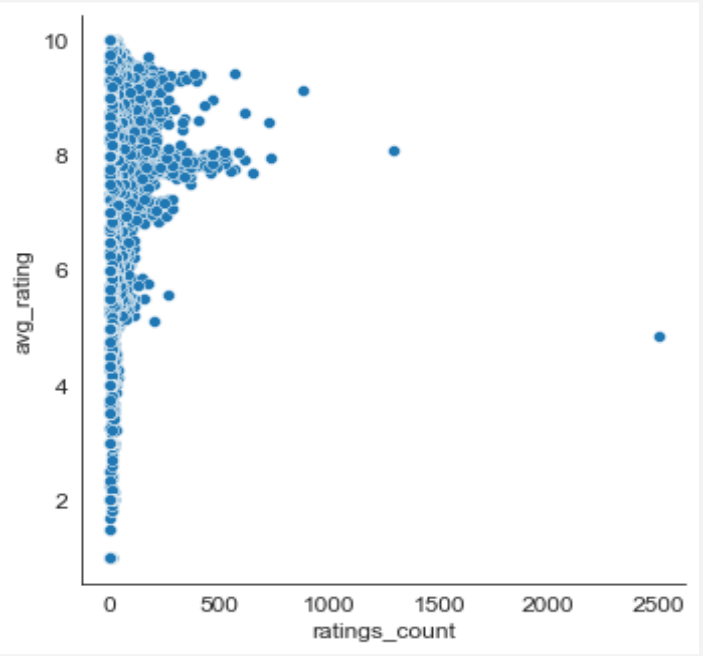


How is the distribution of ratings?



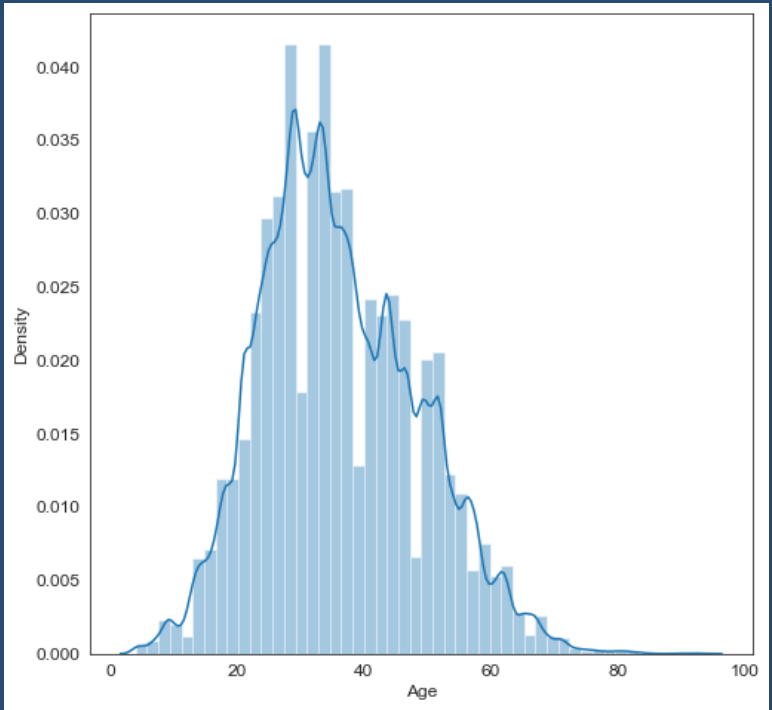
We observe that 8 is the most common rating given by users.

Does the number of reviews affect ratings?



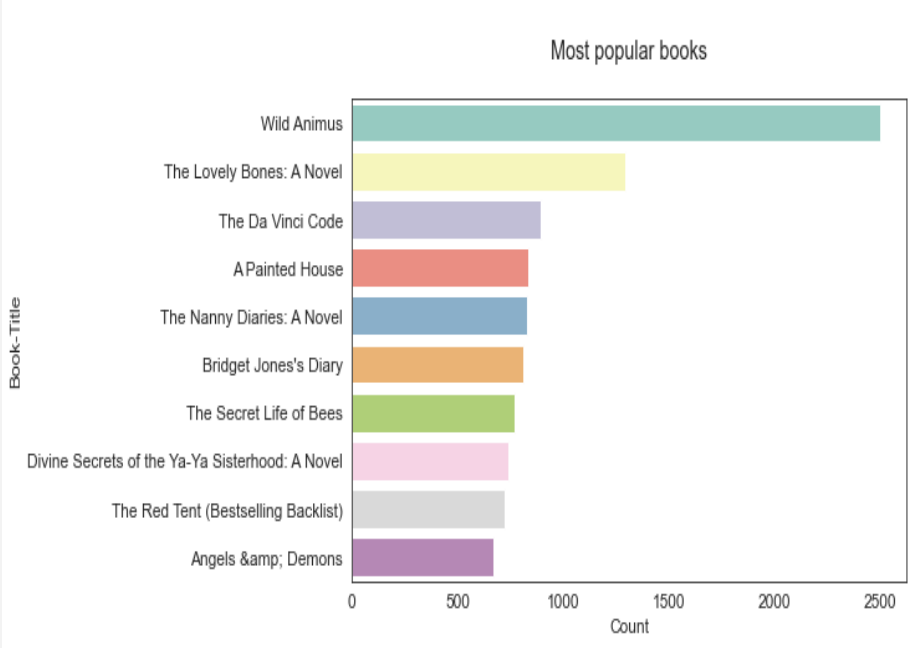
People who rate more number of books tend to give good ratings.

How is the distribution of age?



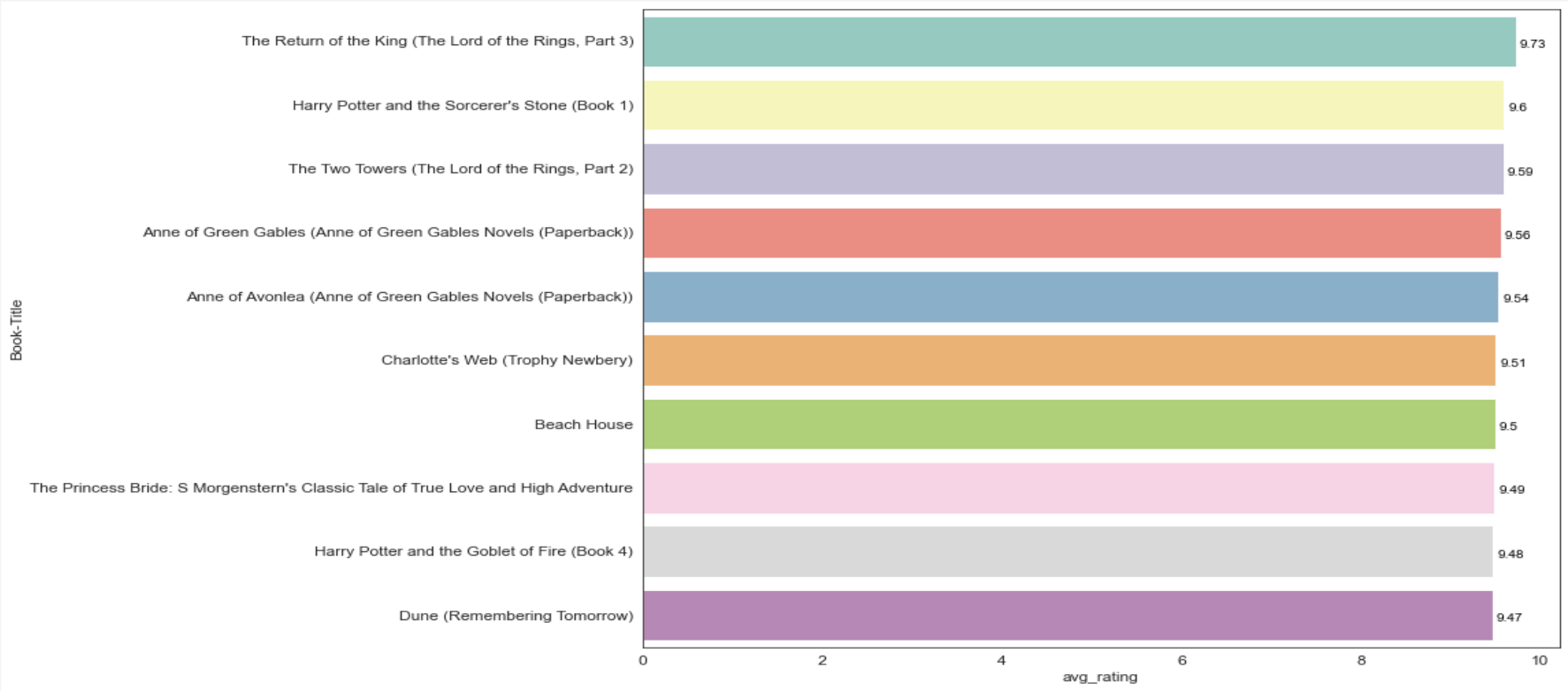
- The majority of readers are between the ages of 25 and 40.
- Readers who are 80 to 100 years old make up a tiny minority.

Most Popular Book among Users



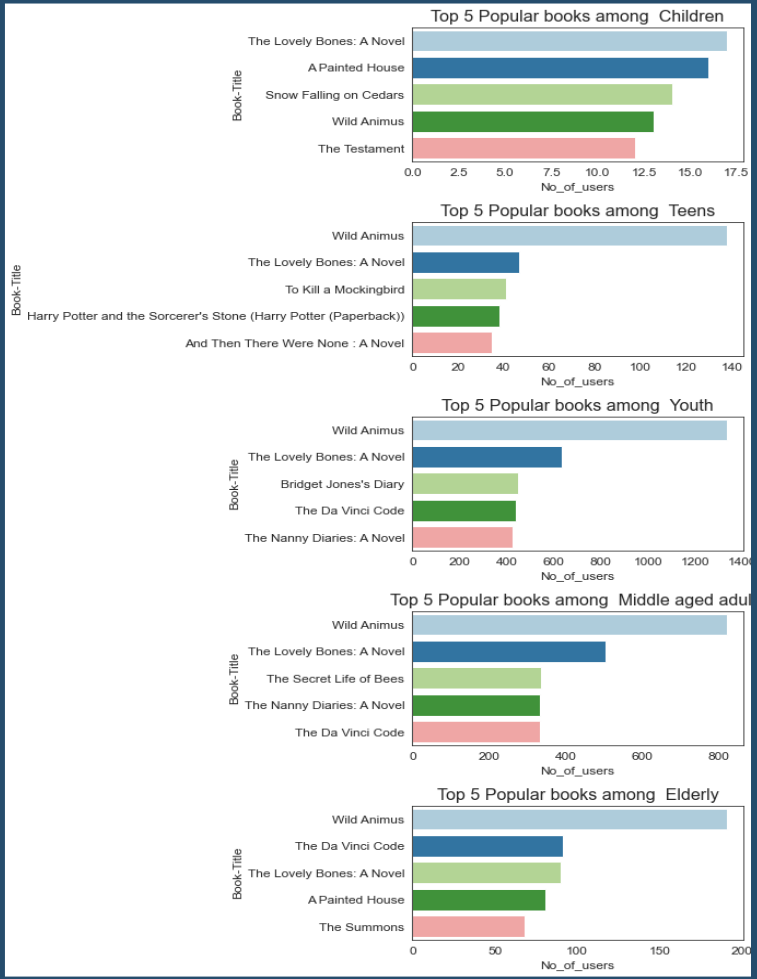
- The book which has been rated by most of the users ‘Wild Animus’.

Book with Highest Average Rating



The book with highest average rating is The Return of the King (The Lord of the Rings,Part 3)

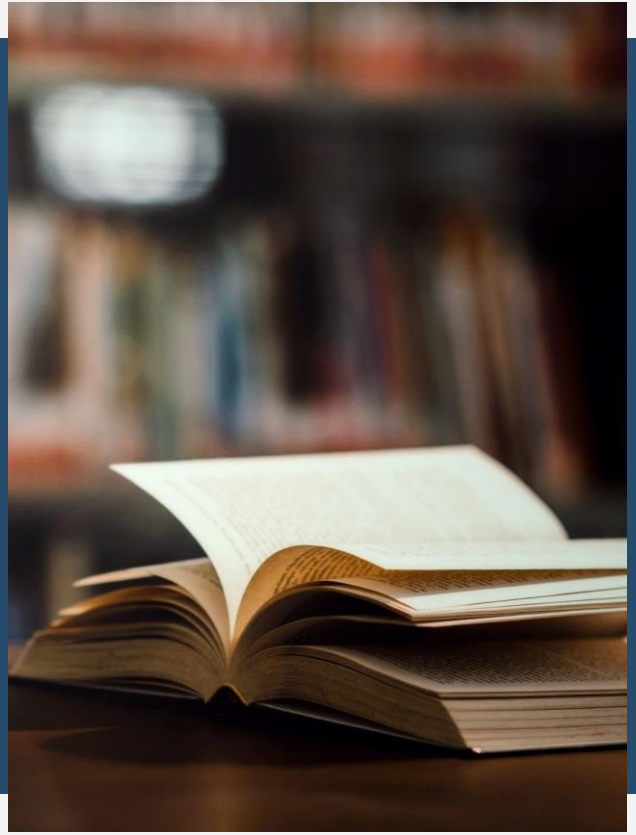
Most Popular Books among different Age Groups



Top Rated Books among different Age Groups

RECOMMENDER SYSTEMS

1. Popularity Based Recommender System
2. Collaborative Filtering
3. Content Based Recommendation System



1. POPULARITY BASED RECOMMENDER

Simple Popularity

Book-Title	Book-Author	Total Number of Users Rated
Wild Animus	RICH SHAPERO	2502
The Lovely Bones: A Novel	ALICE SEBOLD	1295
The Da Vinci Code	DAN BROWN	883
Divine Secrets of the Ya-Ya Sisterhood: A Novel	REBECCA WELLS	732
The Red Tent (Bestselling Backlist)	ANITA DIAMANT	723
A Painted House	JOHN GRISHAM	647

It was observed that the Wild Animus was the most popular Book.
Other than that, books like The Lovely Bones: A Novel, The Da Vinci Code and Divine Secrets of the Ya-Ya Sisterhood: A Novel were among the top books.

1. POPULARITY BASED RECOMMENDER

Country Wise

Book-Title	Book-Author	Total Number of Users Rated
Wild Animus	RICH SHAPERO	235
The Lovely Bones: A Novel	ALICE SEBOLD	118
The Pelican Brief	JOHN GRISHAM	88
She's Come Undone (Oprah's Book Club)	WALLY LAMB	88
The Joy Luck Club	AMY TAN	86

```
Enter the Name of the Country : CANADA
```

Most Popular Books in CANADA

Here the user searched for the country CANADA. It was observed that the Wild Animus was the most popular Book.
Other than that, books like The Lovely Bones: A Novel and The Pelican Brief were among the top books.

1. POPULARITY BASED RECOMMENDER

Weighted Average Rating Method

Book-Title	Book-Author	Weighted Average Rating
The Return of the King (The Lord of the Rings, Part 3)	J.R.R. TOLKIEN	9.6
The Two Towers (The Lord of the Rings, Part 2)	J.R.R. TOLKIEN	9.47
Roots	ALEX HALEY	9.43
Anne of the Island	LUCY MAUD MONTGOMERY	9.42
How Stella Got Her Groove Back	TERRY MCMILLAN	9.42

Most Popular Books Based on Weighted Average Rating

Weighted Rating :9.6 | Weighted Rating :9.47 | Weighted Rating :9.43 | Weighted Rating :9.42 | Weighted Rating :9.42

It was observed that the books from The Lord of the Rings series were the most popular based on weighted rating

1. POPULARITY BASED RECOMMENDER

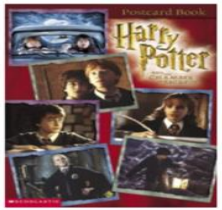
Author Based

Book-Title	Weighted Average Rating
Harry Potter and the Chamber of Secrets Postcard Book	9.41
Harry Potter and the Sorcerer's Stone (Book 1)	9.4
Harry Potter and the Goblet of Fire (Book 4)	9.38
Harry Potter and the Sorcerer's Stone	9.38
Harry Potter and the Prisoner of Azkaban (Book 3)	9.37

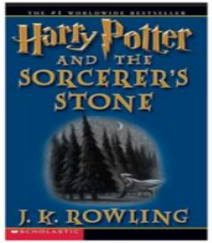
```
Enter the name of the book :Harry Potter and the Chamber of Secrets (Book 2)
The author of the book Harry Potter and the Chamber of Secrets (Book 2) is J. K. ROWLING

Here are the top 5 books from the same author
```

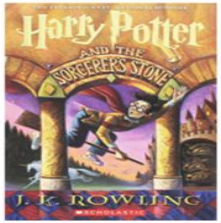
Most Popular Books written by J. K. ROWLING



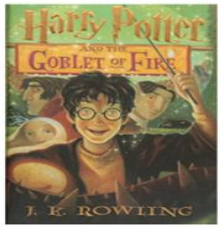
Weighted Rating :9.41



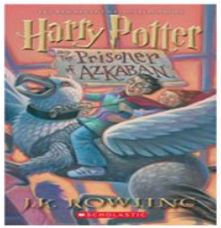
Weighted Rating :9.4



Weighted Rating :9.38



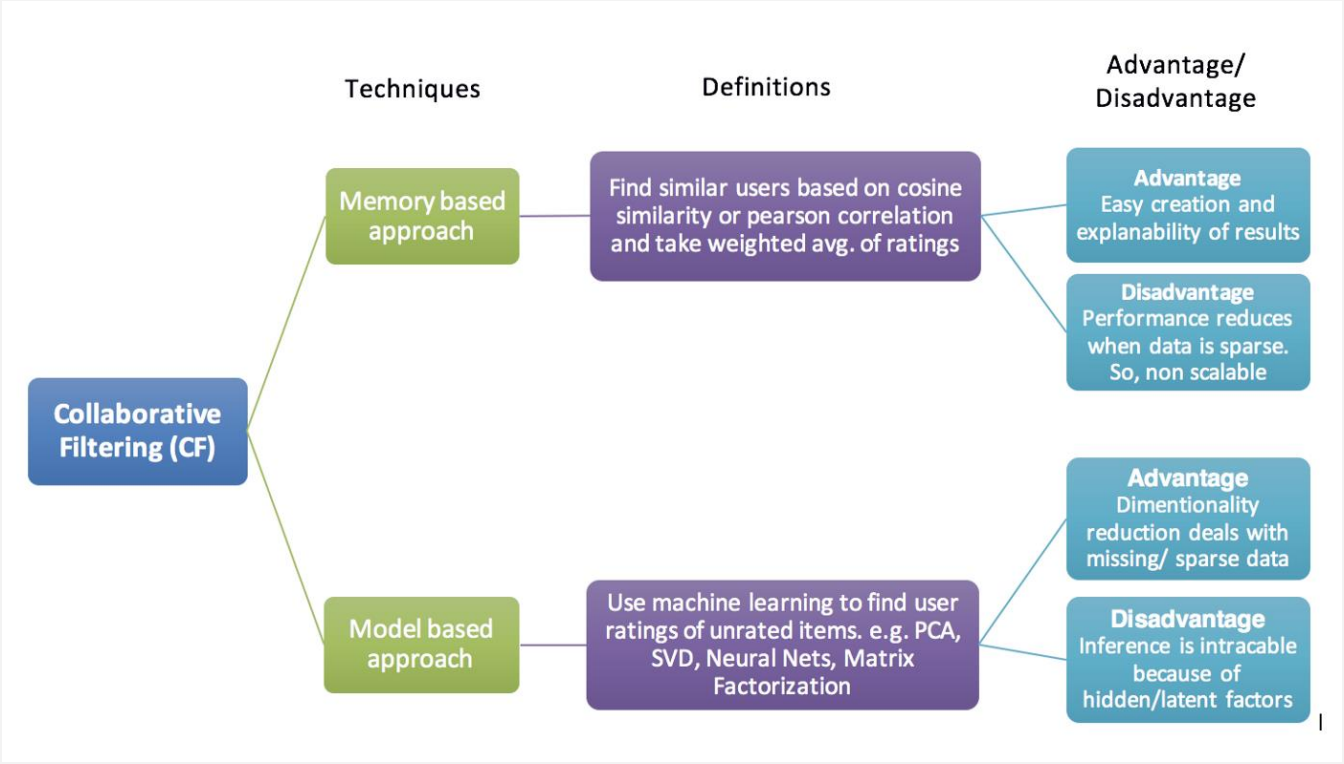
Weighted Rating :9.38



Weighted Rating :9.37

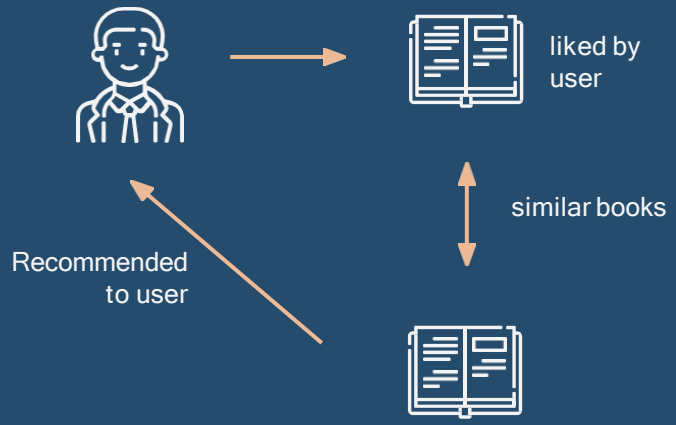
Here the user searched for the book "Harry Potter and the Chamber of Secrets (Book 2)".
The author of that Book is J. K. Rowling.
The top 5 books from the same author were displayed.

2. COLLABORATIVE FILTERING



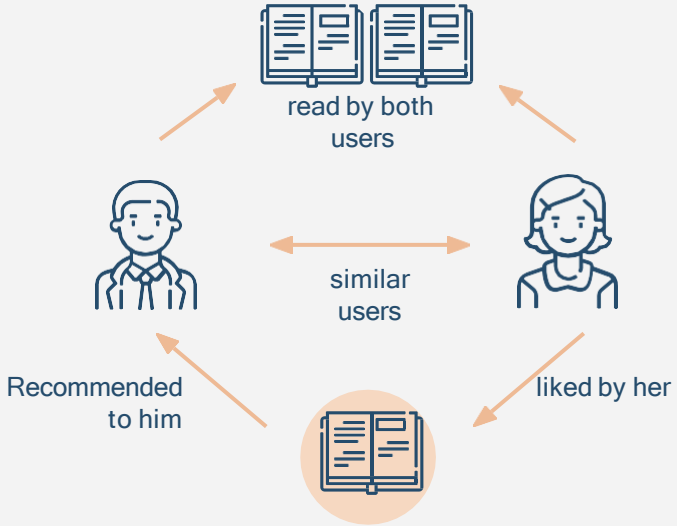
Collaborative Filtering Types

Item Based Collaborative Filtering



Measure Similarity between Books

User Based Collaborative Filtering



Measure Similarity between Users

2.1 Memory Based Collaborative Filtering

Item Based CF (Using Euclidean Distance)

Book-Title	Book-Author	Euclidean Distance
Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. ROWLING	158.590
Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	162.754
Harry Potter and the Sorcerer's Stone (Book 1)	J. K. ROWLING	184.997
Harry Potter and the Order of the Phoenix (Book 5)	J. K. ROWLING	192.678
A Cat on Jingle Bell Rock: An Alice Nestleton Mystery (Alice Nestleton Mysteries	LYDIA ADAMSON	193.134

Enter the name of the book :Harry Potter and the Chamber of Secrets (Book 2)

The top 5 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

Euclidean Distance :184.997 | Euclidean Distance :158.59 | Euclidean Distance :162.754 | Euclidean Distance :192.678 | Euclidean Distance :193.134

Here the user searched for the book "Harry Potter and the Chamber of Secrets (Book 2)".
The top recommendations for this book are "Harry Potter and the Prisoner of Azkaban", "Harry Potter and the Goblet of Fire", etc.

2.1 Memory Based Collaborative Filtering

Item Based CF (Using Cosine Similarity)

Book-Title	Book-Author	Similarity Score
Harry Potter and the Prisoner of Azkaban (Book 3)	J. K. ROWLING	0.36830
Harry Potter and the Goblet of Fire (Book 4)	J. K. ROWLING	0.40062
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))	J. K. ROWLING	0.54164
Harry Potter and the Order of the Phoenix (Book 5)	J. K. ROWLING	0.58850
Harry Potter and the Sorcerer's Stone (Book 1)	J. K. ROWLING	0.58913

Enter the name of the book :Harry Potter and the Chamber of Secrets (Book 2)

The top 5 Recommended books for Harry Potter and the Chamber of Secrets (Book 2) are:

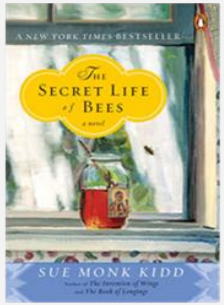
Here the user searched for the book 'Harry Potter and the Chamber of Secrets (Book 2)'. The top recommendations for this book are "Harry Potter and the Prisoner of Azkaban", "Harry Potter and the Goblet of Fire", etc.

2.1 Memory Based Collaborative Filtering

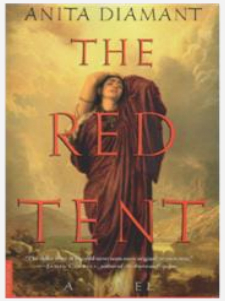
User Based CF

Book-Title	Recall-Strength
The Red Tent (Bestselling Backlist)	0.904644
Divine Secrets of the Ya-Ya Sisterhood: A Novel	0.853586
Where the Heart Is (Oprah's Book Club (Paperback)	0.774157
House of Sand and Fog	0.76866
The Secret Life of Bees	0.764796

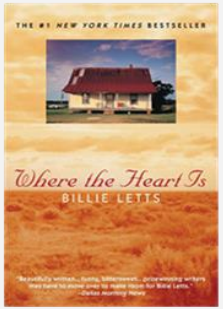
The top 5 Recommended books for User ID: 69078 are:



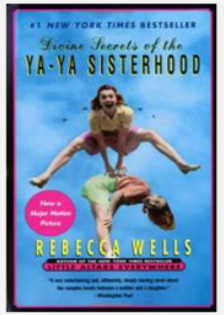
recStrength :0.7648



recStrength :0.90464



recStrength :0.77416



recStrength :0.85359

recStrength :0.76866

User 69078 has high recall capacity for certain books like The Red Tent, meaning there is a high chance of this book being recommended.
Similarly, lesser recall strength means that the corresponding book being recommended has a lower chance.

Overall Model Accuracy

Global recall@5 : 0.4512

Global recall@10: 0.5646

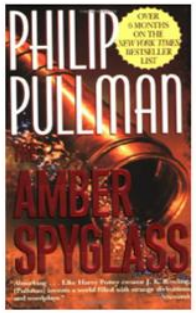
2.2 Model Based Collaborative Filtering

SVD Model

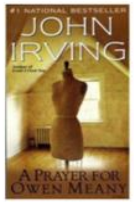
User Id	ISBN	Book Title	Actual Rating	Predicted Rating
193458	345413377	The Amber Spyglass (His Dark Materials, Book 3)	10	8.982
193458	345361792	A Prayer for Owen Meany	10	8.942
193458	1853260002	Pride & Prejudice (Wordsworth	10	8.498
193458	679429220	Midnight in the Garden of Good and Evil: A	9	7.777
193458	64471047	The Lion, the Witch, and the Wardrobe (The	9	8.893

Actual Top Rated books for User ID 193458 are:

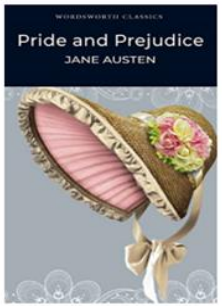
Actual Rating :9



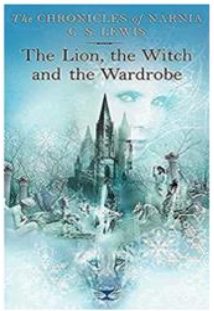
Actual Rating :10



Actual Rating :10



Actual Rating :10



Actual Rating :9

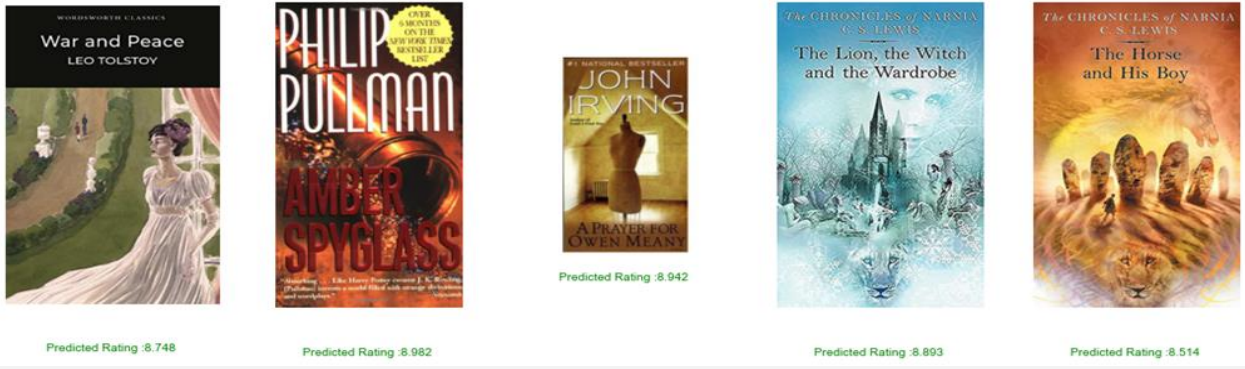
User Id 193458 was selected for analysis of predicted rating using SVD model.
Actual Top-Rated Books by the user 193458 is displayed on left.

2.2 Model Based Collaborative Filtering

SVD Model

User Id	ISBN	Book Title	Actual Rating	Predicted Rating
193458	345413377	The Amber Spyglass (His Dark Materials, Book 3)	10	8.982
193458	345361792	A Prayer for Owen Meany	10	8.942
193458	64471047	The Lion, the Witch, and the Wardrobe (The	9	8.893
193458	1853260622	War and Peace (Wordsworth Classics)	5	8.748
193458	64471063	The Horse and His Boy	9	8.514

Predicted Top Rated books for User ID 193458 are:



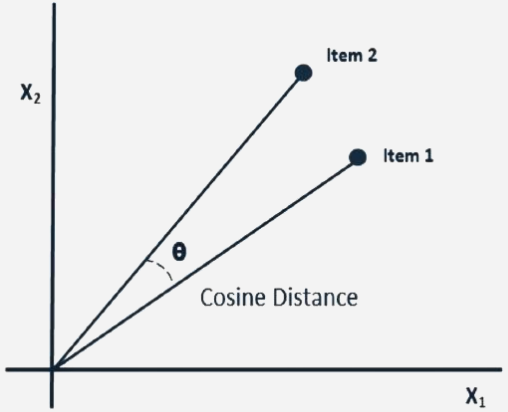
Predicted Top-Rated Books by the user 193458 is displayed on left.
On comparing Actual top rated and Predicted top rated Books we can observe that there is only a very little difference between the prediction and actual ratings.

Model Performance

test_rmse : 1.120441
test_mae : 0.754634
fit_time : 6.476631
test_time : 2.526748

3. CONTENT BASED FILTERING

This approach makes recommendations to users based on the features or characteristics of the books. Using item metadata, the computer will assess how similar the books are to one another and then recommend the books that are most like the one the user loved. One of the way is using cosine similarity.



Flow Content Based Filtering

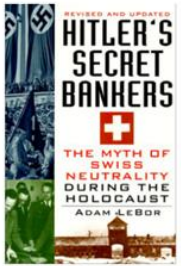
1. Collect feature (title, author, and publisher details) as a content
2. Convert content using `TfidfVectorizer`
3. Measure cosine similarity for all books
4. Input favorite book
5. Rank the cosine similarity of the books based on input
6. Recommend books

3. CONTENT BASED RECOMMENDER

BASED ON BOOK TITLE USING COUNT VECTORISER

Book-Title	Score	Words
Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust	1.000	myth , secret
SGML: The Billion Dollar Secret	0.707	secret
Callahan's Secret	0.707	secret
Till We Have Faces: A Myth Retold	0.707	myth
Sprite's Secret (Pixie Tricks, 1)	0.707	secret

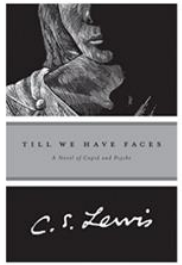
Recommendations for the book "Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust" are:



Similarity Score :1.0

Similarity Score :0.707

Similarity Score :0.707



Similarity Score :0.707

Similarity Score :0.707

Recommendations are provided for the same book as previous: Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust.

It will look for the cosine similarity between the title words in the inputted book and other available books.

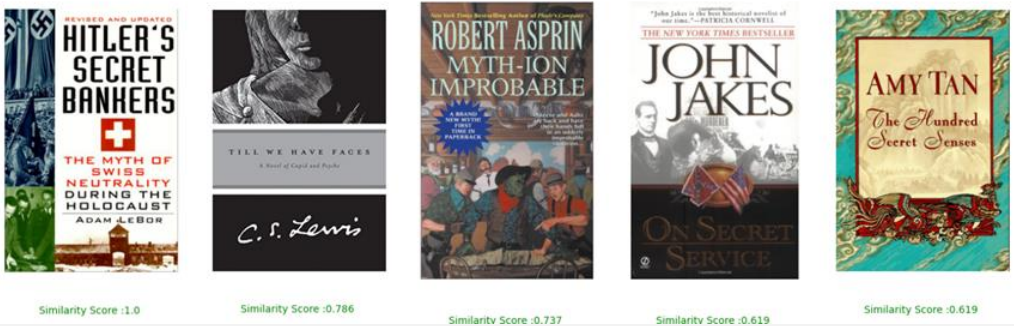
It is helpful for recommending the same book with different volumes or different editions.

3. CONTENT BASED RECOMMENDER

BASED ON BOOK TITLE USING TF-IDF VECTORISER

Book-Title	Score	Words
Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust	1.000	myth , secret
Till We Have Faces: A Myth Retold	0.786	myth
Myth-Ion Improbable (Myth Books)	0.737	books , myth
On Secret Service	0.619	secret
The Hundred Secret Senses	0.619	secret

Recommendations for the book "Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust" are:



Similarity Score :1.0 | Similarity Score :0.786 | Similarity Score :0.737 | Similarity Score :0.619 | Similarity Score :0.619

Recommendations are provided for the same book as previous: Hitler's Secret Bankers: The Myth of Swiss Neutrality During the Holocaust.

It looked for other similar books by looking at not only the word frequency, but also giving weightage based on term importance in book title.

3. CONTENT BASED RECOMMENDER

BASED ON BOOK AUTHOR USING TF-IDF VECTORISER

Book-Title	Book-Author	Similarity Score
Mortal Fear	GREG ILES	1.000
The Quiet Game	GREG ILES	1.000
Sleep No More	GREG ILES	1.000
The Footprints of God : A Novel	GREG ILES	1.000
Dead Sleep	GREG ILES	1.000

Recommendations for the book "The Quiet Game" by Greg Iles are:

Similarity Score :1.0 | Similarity Score :1.0 | Similarity Score :1.0 | Similarity Score :1.0 | Similarity Score :1.0

The recommendations are provided for the book The Quiet Game by Greg Iles. The books written by the same author, Greg Iles, are provided as recommendations.

This recommender is based on the idea that books written by the same author are likely to share commonalities in taste and style.

3. CONTENT BASED RECOMMENDER

BASED ON SEARCH KEYWORD USING TF-IDF VECTORISER

Book-Title	Book-Author	Publisher	Similarity Score
Fly Away Home	JUDITH KELMAN	Bantam Books	0.4104
Summer of Storms	JUDITH KELMAN	Jove Books	0.3639
And Then You Die	IRIS JOHANSEN	Bantam	0.3091
Till We Meet Again	JUDITH KRANTZ	Bantam Books	0.1662
Die upon a Kiss	BARBARA HAMBLY	Bantam Books	0.1498

Recommendations for the keyword "If I Should Die" are:

Recommendations provided for the book If I Should Die written by Judith Kelman and published by Bantam Books are given.

This recommender can provide more precise recommendations as it takes into account greater amount of information.

3. CONTENT BASED RECOMMENDER

BASED ON SEARCH KEYWORD VIA TAGS

```
Enter a Search Keyword: King
```

Book-Title	Book-Author	Publisher	Corpus	Tags
All the King's Men	ROBERT PENN WARREN	Harvest Books	ROBERT PENN WARREN All the King's Men Harvest Books	robert,penn,warren,king,men,harvest,book
The Girl Who Loved Tom Gordon	STEPHEN KING	Pocket	STEPHEN KING The Girl Who Loved Tom Gordon Pocket	stephen,king,girl,love,tom,gordon,pocket
The King of Torts	JOHN GRISHAM	Doubleday Books	JOHN GRISHAM The King of Torts Doubleday Books	john,grisham,king,tort,doubleday,book
Pet Sematary	STEPHEN KING	Signet Book	STEPHEN KING Pet Sematary Signet Book	stephen,king,pet,sematari,signet,book
The Winter King: A Novel of Arthur (The Warlord Chronicles: I)	BERNARD CORNWELL	St. Martin's Griffin	BERNARD CORNWELL The Winter King: A Novel of Arthur (The Warlord Chronicles: I) St. Martin's Griffin	bernard,cornwel,winter,king,novel,arthur,warlord,chronicl,st,martin,griffin

The King keyword was searched across the tags of all books to find a match and provide recommendation.

When the user don't know the precise name of book title, publisher name or book author name, this recommender system comes for help in identifying the required book.

This type of recommender is particularly useful for **search bars.**

Conclusion

Key Findings from the Study

Strength and Weakness of Each Model

METHOD	PROS	CONS
Popularity Based Recommender System	• suitable for new user • relatively easy	• does not provide user-specific recommendations • result is obvious
Collaborative Filtering	• more personalized • more diverse	• cold-start problem when there is new user/new item • need data about other users
Content Based Filtering	• personalized, recommend based on user interest • doesn't need data about other users	• result rely on what is written in the content

Key Findings

- Popularity based recommendation system can be implemented for new (cold-start) users for providing recommendation.
- Memory based Collaborative Filtering could capture user-user similarity and item-item similarity and provide relevant suggestions.
- For modelling, it was observed that for model based collaborative filtering, SVD technique outperformed NMF with lower Root Mean Square Error (RMSE).
- In the content-based recommender system, recommendations based on tags proved to be useful for searching for a book without knowing its exact name. Additionally, recommendations based on Count Vectorizer and TF-IDF Vectorizer were applied in this section.
- Recommender systems based on book titles proved to be useful for suggesting the same book but with different volumes or editions. This can help users discover other available versions or additional parts of the same book.
- Recommender systems based on author names proved to be beneficial for users as they can discover books with similar themes or genres. Books written by the same author are likely to share commonalities in taste and style, making it easier for users to find similar books.

Although we have performed offline-based evaluations for some of the methods, online-based evaluation techniques such as A/B testing will be necessary to compare and understand recommendation systems better. These techniques will be used after deploying the methods in an application or website.

References:

- https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset
- https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada
- https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f
- https://towardsdatascience.com/recommendation-system-matrix-factorization-svd-explained-c9a50d93e488
- https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning

Thank you