Barah (C0860531)

Catharin Jose (C0860087)

Danny Jose (C0864600)

Sri Bindu Chintakayala (C0857498)

AIMT DEPARTMENT, LAMBTON COLLEGE
2022F AML 2203 1 Advanced Python AI and ML Tools
PROFESSOR VAHID HADAVI
NOVEMBER 26, 2022

Abstract

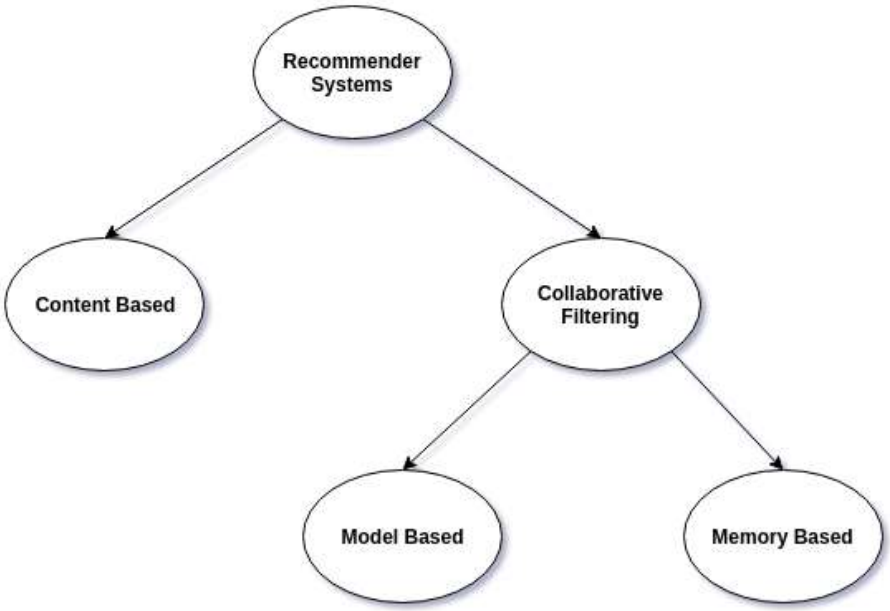
A recommendation engine is a form of machine learning that provides the customer with relevant suggestions. It is one of the most popular data science applications. A recommendation system uses data gathered from interactions with people and products to understand their preferences, previous decisions, and characteristics. These include impressions, clicks, likes, and purchases. Content providers and product providers like recommender systems because they can predict consumer interests and desires on a highly personalized level. Any product or service consumers can find interesting can be marketed through them, such as books, videos, health classes, or clothing. Prior to the recommendation system, the most common way to buy was to take a friend's recommendation. According to your search, watch, or purchase history, Google now knows what kind of news you're likely to read, and YouTube knows what type of videos you're likely to watch. We are developing a 'Book Recommendation System' in order to perform machine learning tasks to gain insights from the interactions between users and items and provide the best recommendations.

Introduction

A recommendation System (RS) is software that proposes comparable products to a customer based on previous purchases or preferences. Vast amounts of data on products are examined by RS, who then develops a list of those products that would satisfy the buyer's needs. Here, we have focused on the Book Recommendation system. In a book recommendation system, we must suggest related books to the reader depending on his interests. Online stores like Google Play Books, Open Library, Good Reads, and others that sell eBooks employ the book recommendation algorithm.

There are several approaches to using a recommendation system. Content-based and collaborative filtering systems are the two main categories. Content-based systems are created using knowledge of the attributes of the products themselves. When "other items like this" are shown to you, that is what you see. They are highly helpful since consumers are probably going to appreciate another book or movie that is comparable to the one they are looking for. Conversely, Collaborative Filtering systems rely on product ratings from users. A compilation of user evaluations of various products is used to compute them. The theory behind it is that users who are similar to one another would have similar tastes and that people tend to appreciate things that are similar to one another.

It has been found that models can generally predict the best recommendations for data according to their RMSE scores. As for Collaborative Filtering, we’ll be focusing on RMSE. Different RMSE scores generated from different methods can be used to reveal the best recommendation.



(https://www.kdnuggets.com/2019/09/machine-learning-recommender-systems.html)

Here, in this project, three different datasets were used

- Book data – (ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image URL-M, Image-URL-L)
- Users' data - (User-ID, Location, Age)
- Rating data - (User-ID, ISBN, Book-Rating)

We aim to develop a predictive recommendation model that could help companies gain insights from user-item interactions and provide users with the best recommendations.

Data Preparation (Data Cleaning and Feature Engineering)

Null values treatment

We got null values in the book dataset for the columns book-author and publisher. In the users, the dataset null value was observed for the Age column. The rating dataset doesn't contain any null values.

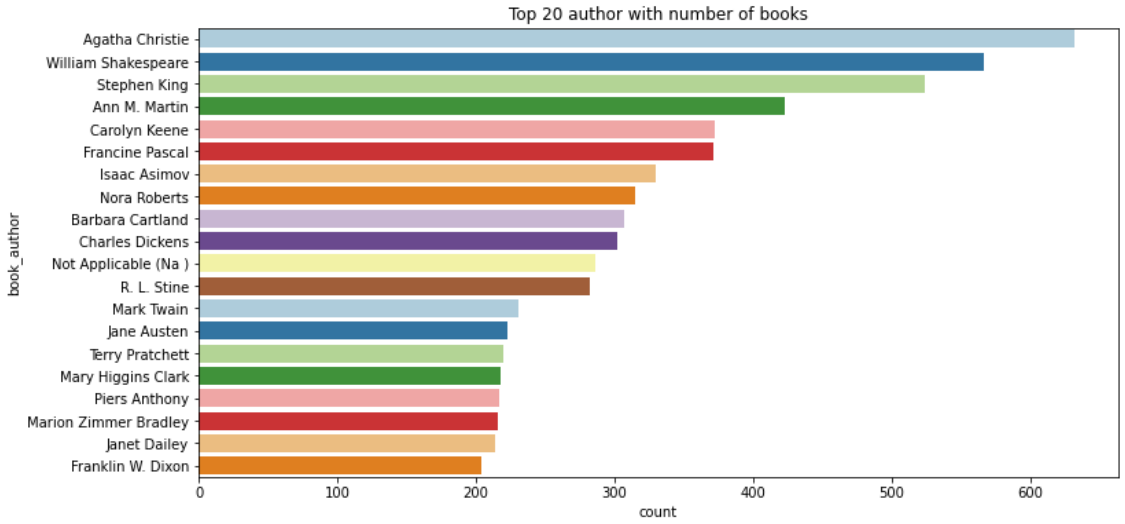
Dropping and replacing data

Moving forward with data cleaning and feature selection is critical. We dropped features like image URLs. To create a space between words, we replaced the feature with lowercase and '-.' We replaced some of the null values (Age) in feature data with the mean of that particular

feature. Handled mismatched features such as book title, book author, year of publication, and publisher. We analyzed and recommended users' data based solely on their age range of 5-90.

Exploratory Data Analysis

The top Author with the number of books.

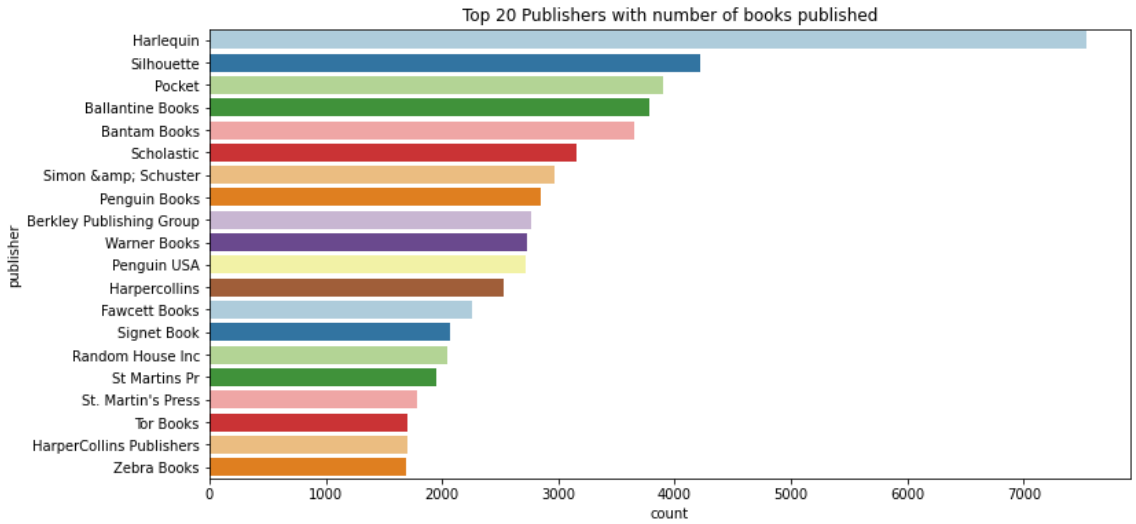


Agatha Christie is leading at top with more than 600 counts, followed by William Shakespeare.

We can draw some hypotheses based on it:-

There is a possibility that Agatha Christie is not the best Author, even though Agatha Christie has published the most books compared to others. William Shakespeare is one of the most famous authors in the world. Still he doesn't have the most books.Among other Authors, it might happen that a few of them have some works that are best sellers and have sold millions of copies.

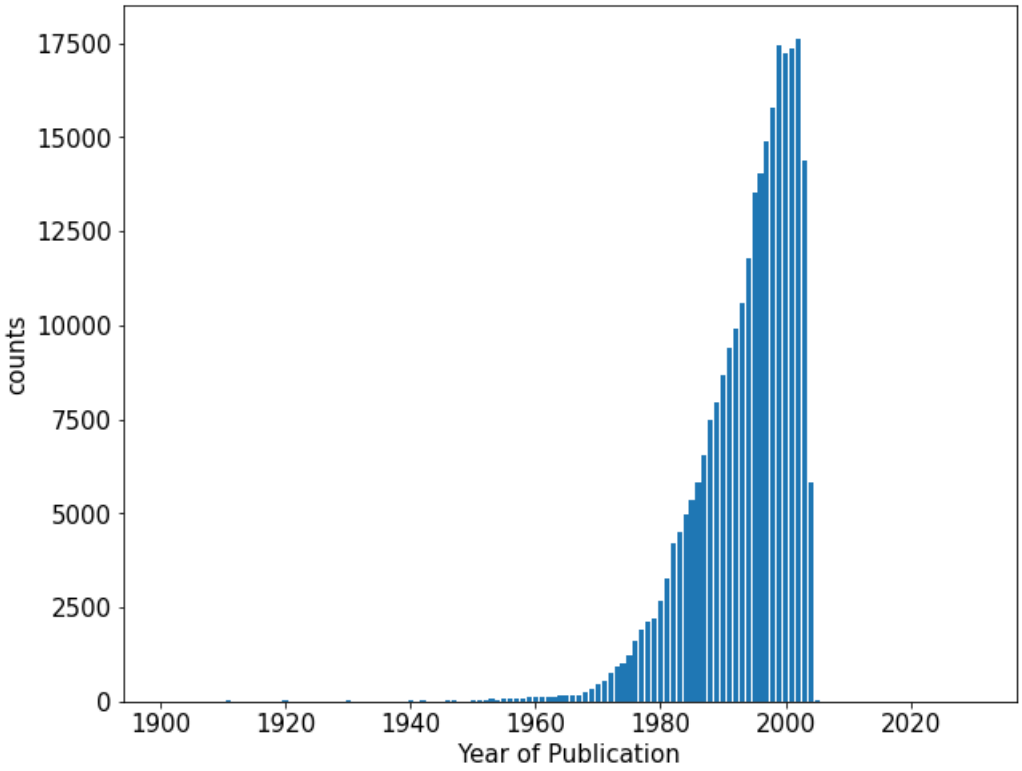
Top publishers with published books



Harlequin has the greatest number of books published, followed by Silhouette.

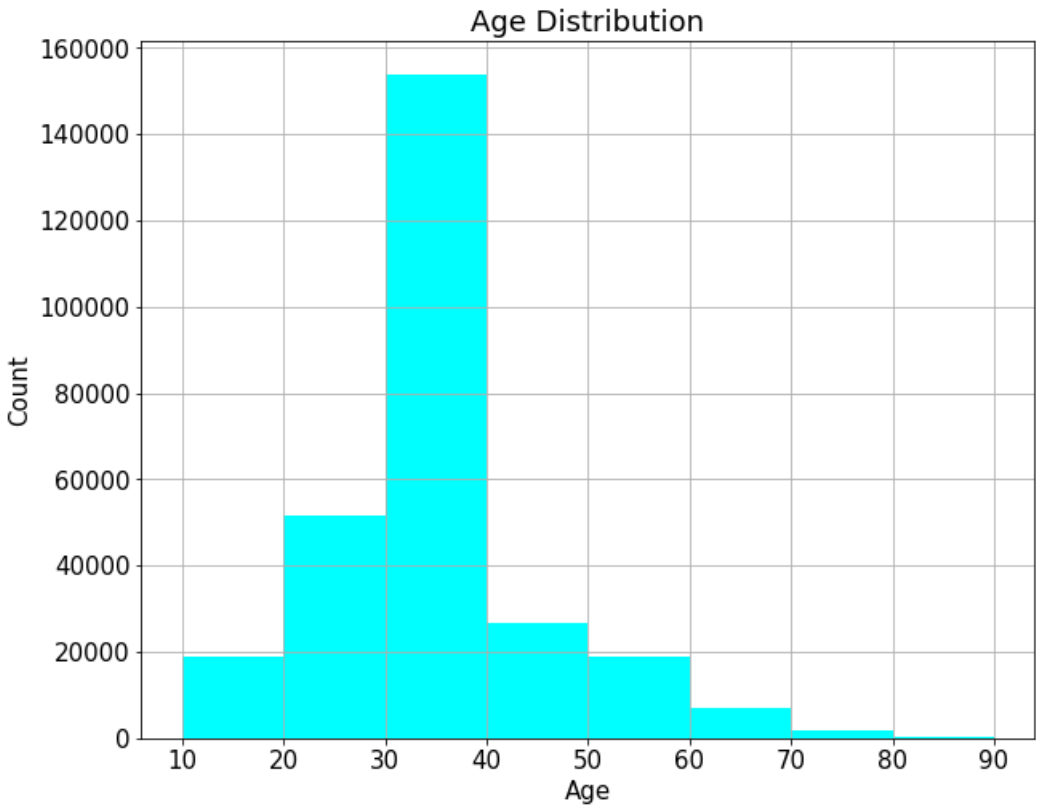
- Some of the top Author's had published their books from Harlequin.
- We can observe Harlequin publisher’s marking better performance than any other publishers.
- Penguin Books, Warner Books, Penguin USA, Berkely Publishing Group and many more are among popular publisher's remarking competition with Harlequin.
- Though Penguin Books Publisher has less number of books published but it might happen that only top Author are approaching towards Penguin Books Publisher.

Number of Books published in yearly.



So we can see that publication year are somewhat between 1950 - 2005 here. The publication of books got vital when it starts emerging from 1950.

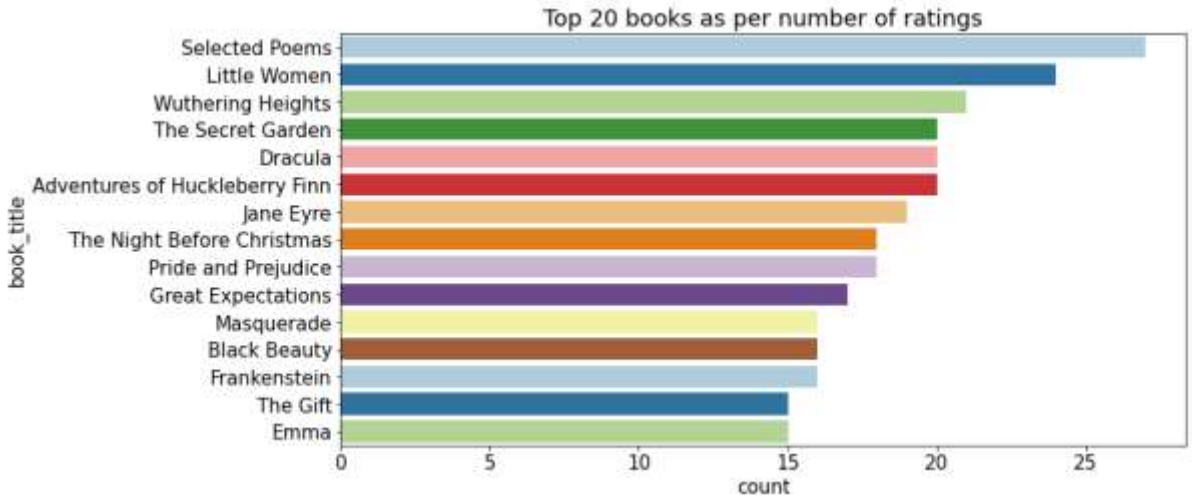
Age distributions of users data



Looking towards the users age between 30-40 prefer more books and followed by users age between 20-30.

- It is obvious that most of the user books are from Age 30 to 40.
- The age group between 20-30 are immensely attracted to read books published by Author.
- We can observe same pitch for Age group between 10-20 and 50-60. There are can be lot of different reasons.

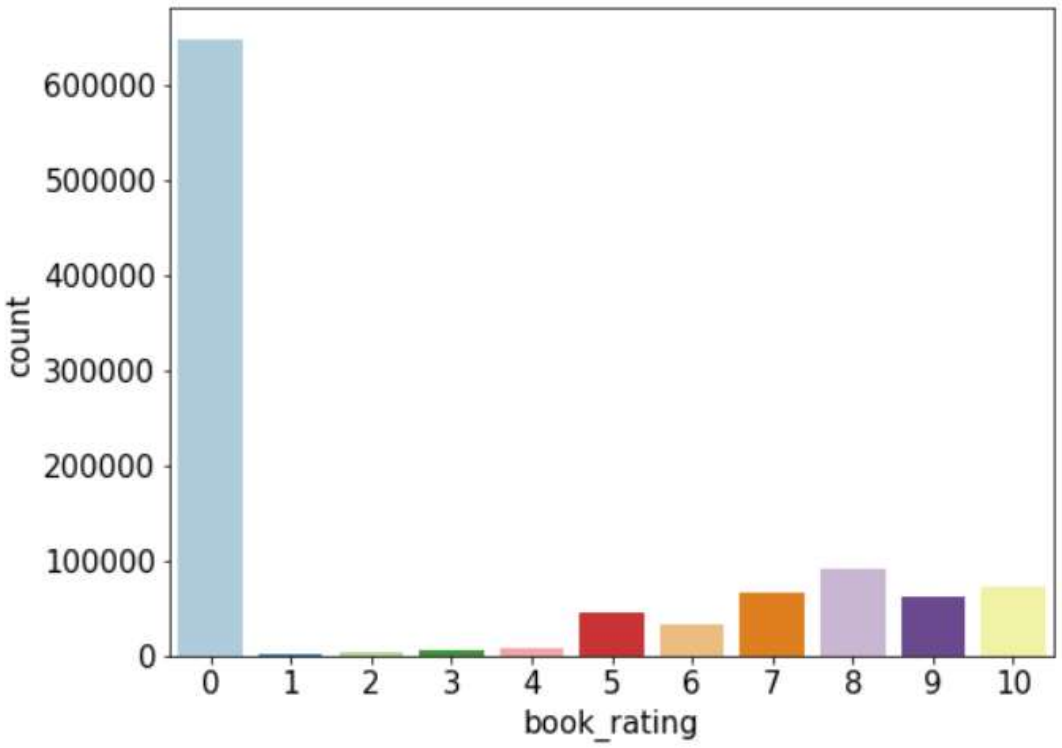
The top 20 books as per number of ratings



As per ratings "Selected Poems" has been rated most followed by "Little Women".

* Selected Poems are most favourable to users as per ratings. Three of the books, 'The Secret Garden, 'Dracula,' and 'Adventures of Huckleberry Finn' are having almost similar ratings. We can observe a similar trend in 'Masquerade', 'Black Beauty', and Frankenstein'.

Unique ratings from the 'ratings data' and 'books data' datasets.



Firstly the above ratings are unique ratings from 'ratings_data' and 'books_data' dataset. We have to separate the explicit ratings represented by 1–10 and implicit ratings represented by 0.

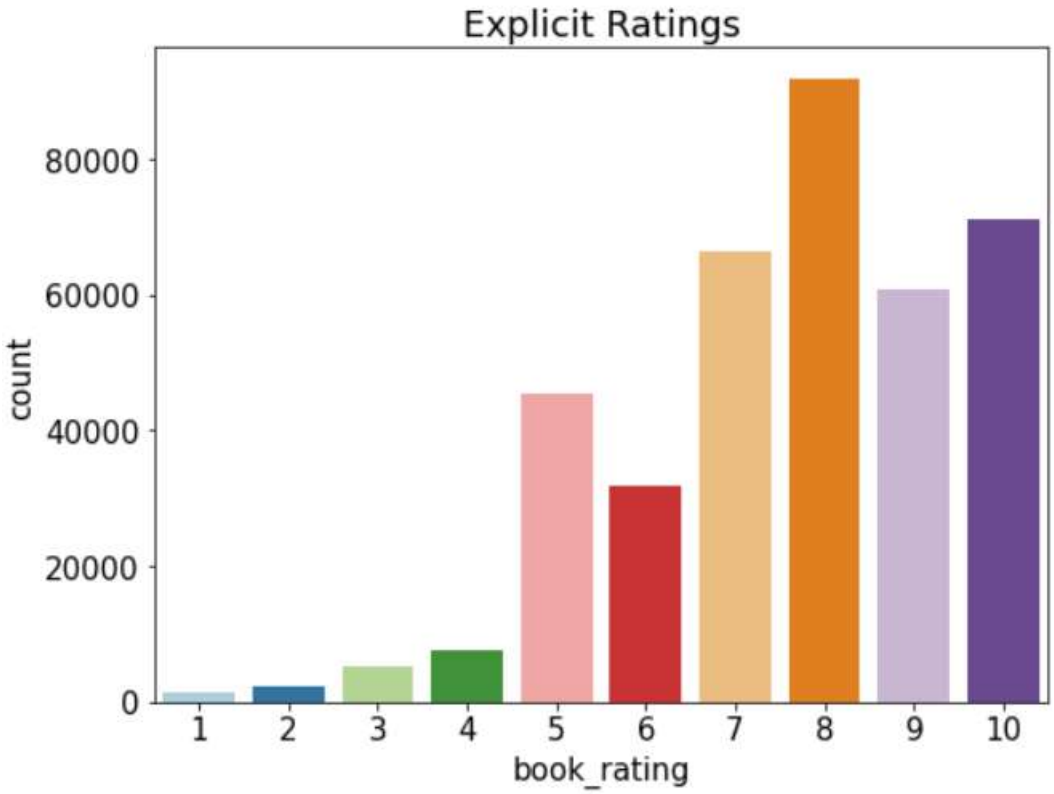
We can draw some hypotheses based on it:-

* This count plot shows users have rated 0 the most, which means they haven't rated books at all.

* Still, we can see a pattern to recognize in ratings from 1-10.

* Mostly, the users have rated 8 ratings out of 10 as per books. It might happen that the feedback is positive but not extremely positive as 10 ratings (i.e best books ever).

Explicit Ratings



Now this count plot of book ratings indicates that higher ratings are more common amongst users, and rating 8 has been rated the highest number of times.

There can be many assumptions based on ratings of users :-

* Let's take the rating group from 1-4. This can be a negative impact on books being published if they have ratings from 1 to 4.

* For 5 ratings, the users might not sure about book ratings whether it's positive or negative impact.

* Let's take the rating group from 6-10. This is positive feedback. 6 ratings are very low among other ratings. As we can aspects 7 and 8 are average and have more ratings from users. 9 and 10 ratings are the top best ratings based on Author's, Publisher's and Books published.

Top 10 recommendation books based on Explicit Ratings

book_title	book_rating
The Lovely Bones: A Novel	707
Wild Animus	581
The Da Vinci Code	494
The Secret Life of Bees	406
The Nanny Diaries: A Novel	393
The Red Tent (Bestselling Backlist)	383
Bridget Jones's Diary	377
A Painted House	366
Life of Pi	336
Harry Potter and the Chamber of Secrets (Book 2)	326

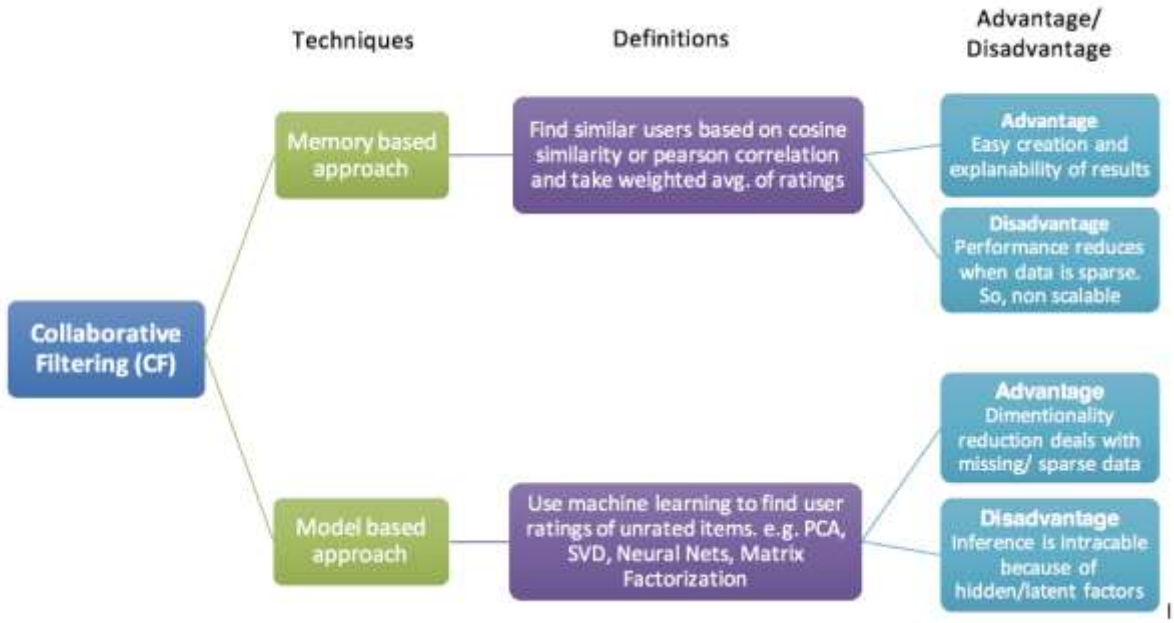
The above are the top 10 books recommendation as per ratings. But this is not based on any recommendation system. They are top 10 books as per explicit ratings.

Methods

Collaborative filtering:

Collaborative filtering separates information based on user interactions and data gathered by the system from other users. It is predicated on the notion that people are more likely to agree again in the future if they previously agreed on how they rated specific goods. The link between users and objects is the main emphasis of collaborative filtering systems. The similarity of the ratings given to the objects by the people who have rated both things establishes how comparable they are.

(https://builtin.com/data-science/collaborative-filtering-recommender-system)



(https://en.wikipedia.org/wiki/Collaborative_filtering)

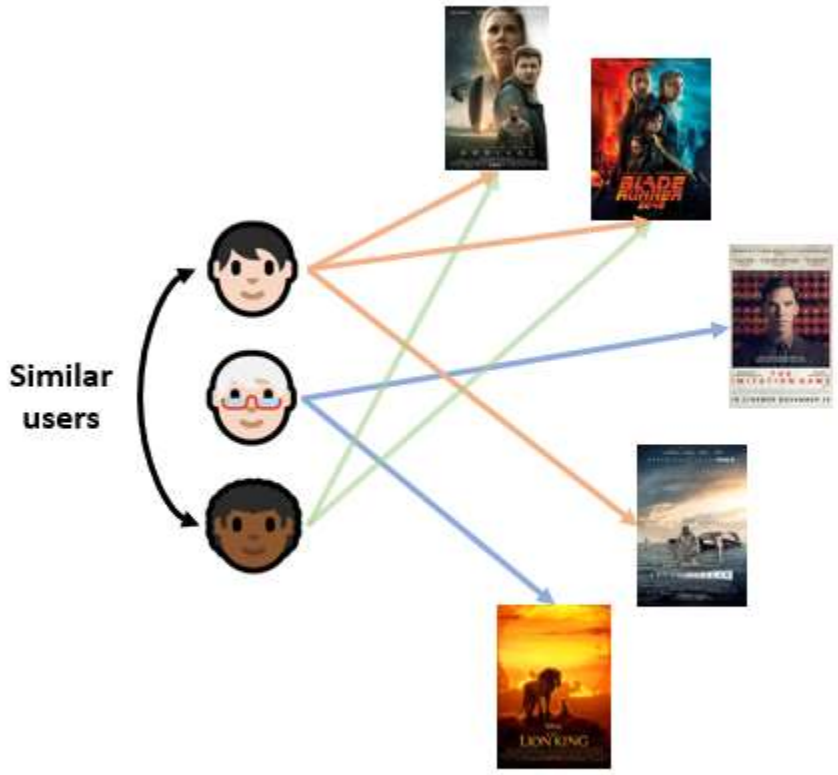
Memory-based CF:

Memory-based techniques calculate the similarity between users or products using past user ratings data. These techniques work by defining a similarity metric between users or objects, then identifying the most comparable to suggest undiscovered stuff. User-Based and Item-Based are the two primary categories of memory-based collaborative filtering algorithms.

User-based:

To determine the k most comparable users to an active user, the user-based top-N recommendation method employs a similarity-based vector model. The user-item matrices for the k most comparable users are then combined to determine the set of things that should be suggested. Locality-sensitive hashing, which employs the closest neighbour approach in linear time, is a well-liked technique for locating other users who have common characteristics.

The benefits of this technique include the findings' explainability, which is a crucial component of recommendation systems; ease of construction and use; simplicity in facilitating the introduction of new data; the recommendations' content independence; and effective scalability with co-rated items. (https://en.wikipedia.org/wiki/Collaborative_filtering)



(https://towardsdatascience.com/how-does-collaborative-filtering-work-da56ea94e331)

Model-based CF:

This method uses machine learning methods to create CF models that forecast user ratings of unrated objects. In the fields of data science and machine learning, Singular Value Decomposition (SVD), a traditional linear algebraic approach, is becoming more and more well-liked. This popularity derives from its usage in creating recommender systems. Many online user-centric apps, such as video players, music players, e-commerce applications, etc., suggest additional content for users to interact with.

Model Evaluation metrics

Model evaluation metrics are important to distinguish the best collaborative filtering – either by memory-based or model-based approach. RMSE score is the best way to model performance for a recommendation system. Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data.

- Memory-Based approach - Cosine Similarity

The memory based approach – Cosine Similarity shows RMSE score for item-based CF is 7.95, and for user based CF it shows 7.95. We can make improvements in this score by using the Single Value Decomposition model (SVD) model.

- Model-Based approach – Singular Value Decomposition (SVD)

Model-based collaborative filtering made it a better score with the Latent Factor Model called SVD. The score improved to 1.63 for both SVD RMSE and accuracy scores.

Results

Testing Results

To test the result, a user id 276744 and ISBN 038550120X have been taken. The estimated rating for the book with ISBN code 038550120X from user #276744 is 7.48. The actual rating given for this was 7.00.

Getting top recommendations for books and ratings

First, we map the predictions to each user. Then sort the predictions for each user and retrieve the k highest ones. We took a random User id 32440 for getting a recommendation. We have the top recommendation of books and ratings respective to them.

```
84 Charing Cross Road: 8.65474449552575
The Magician's Nephew (rack) (Narnia): 8.507017449752833
The Secret Life of Bees: 8.506174365880865
Girl with a Pearl Earring: 8.26748648765635
Life of Pi: 8.203716165725332
Middlesex: A Novel: 8.18050252105318
The Thorn Birds: 8.135046954949175
The Handmaid's Tale : A Novel: 8.065235093028042
A Walk in the Woods: Rediscovering America on the Appalachian Trail (Official Guides to the Appalachian Trail): 7.9429278824185
925
Anne of Green Gables (Children's Classics): 7.908995747944084
```

Conclusion

- Among the top 20 Authors, the highest number of books has been held by Agatha Christie. Agatha Christie is leading at the top with more than 600 counts, followed by William Shakespeare.
- Harlequin has the greatest number of books published, followed by Silhouette.
- Number of Books published yearly between 1950 - 2005.

- Most of the users between 30-40 prefer more books, and somewhat we can also view between 20-30.
- As per ratings, "Selected Poems" has been rated the most, followed by "Little Women.” The counterplot shows users have rated 0 the most, which means they haven't rated books at all.
- The top 10 books recommendation as per ratings with top "The Lovely Bones: A Novel" with 707 book ratings. But this is not based on some recommendation system. They are top 10 books as per ratings.
- As we perform by cosine similarity in the recommendation system, it gives a 7.95 RMSE score, and SVD improved score to 1.63 RSME score by Singular Value Decomposition model (SVD).
- The evaluation metrics for SVD are the best RMSE score for all datasets.
- As model-based approach was best to signify, and at last, we got the top 10 recommended books and ratings, respectively.

Refrences

(https://towardsdatascience.com/what-should-i-read-next-f02a16bec832)