Capstone Project

Book Recommendation System

Team
Arman Alam
Abdul Rahman

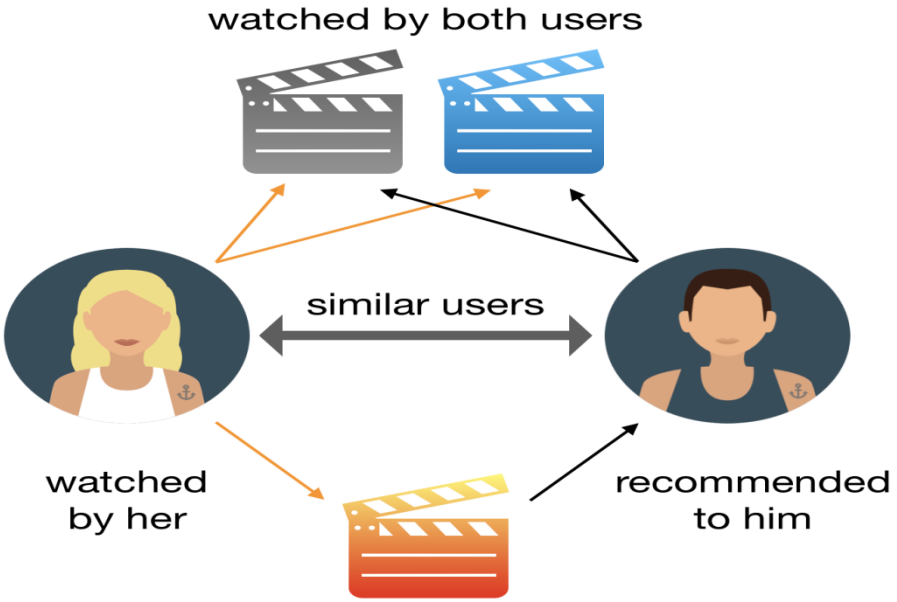
AlmaBetter

Table of contents:

Introduction

Recommender systems are machine learning systems that help users discover new product and services.

A recommendation system helps an organization to create loyal customers and build trust by them desired products and services for which they came on your site. A book recommendation system is a type of recommendation system where we have to recommend similar books to the reader based on his/her interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, goodReads, etc.



Problem Statement

During the last few decades, with the rise of Youtube, Amazon, Netflix, and many other such web services, recommender systems have taken more and more place in our lives. From e-commerce (suggest to buyers articles that could interest them) to online advertisement (suggest to users the right contents, matching their preferences), recommender systems are today unavoidable in our daily online journeys.

In a very general way, recommender systems are algorithms aimed at suggesting relevant items to users (items being movies to watch, text to read, products to buy, or anything else depending on industries).

Recommender systems are really critical in some industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors. The main objective is to create a book recommendation system for users.

Data Overview

Understanding datasets better :

The Book-Crossing dataset comprises 3 files.

Users

Contains the users. Note that user IDs (User-ID) have been anonymized and map to integers. Demographic data is provided (Location, Age) if available. Otherwise, these fields contain NULL values.

Books

Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given (Book-Title,Book-Author, Year-Of-Publication, Publisher), obtained from Amazon Web Services. Note that in the case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavors (Image-URL-S, Image-URL-M, Image-URL-L), i.e., small, medium, large. These URLs point to the Amazon website.

Ratings

Contains the book rating information. Ratings (Book-Rating) are either explicit, *expressed* on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

Continued...

Books Data :

- Data set Contains 'ISBN', 'Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher', 'Image-URL-S', 'Image-URL-M', 'Image-URL-L' information's.
- The Data set Contains 271354 Rows

Users Data :

- Data set Contains 'User-ID' , 'Location', 'Age' information's.
- The Data set Contains 278858 Rows

Ratings Data :

- Data set Contains 'User-ID' , 'ISBN', 'Book-Ratings' information's.
- The Data set Contains 1149780 Rows

Data Preparation for k-Nearest Neighbor Model

* I will take only those users who rate on minimum 200 books.
* I will consider only those books which are rated by at least 50 users.

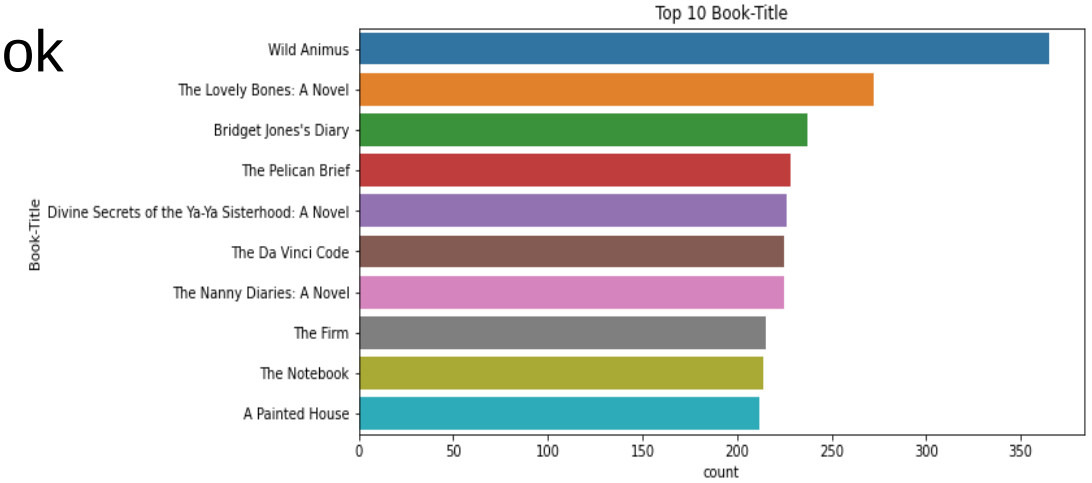
Observation :

* We are getting observation from here that 278858 people had rated on 1149780 Books but out of this only 905 people have given rating on 527556 Books
* This tells us that most people do not rate
* That's why we removed them from the data
* Our final Data Frame contains only 194280 rows

Exploratory Data Analysis

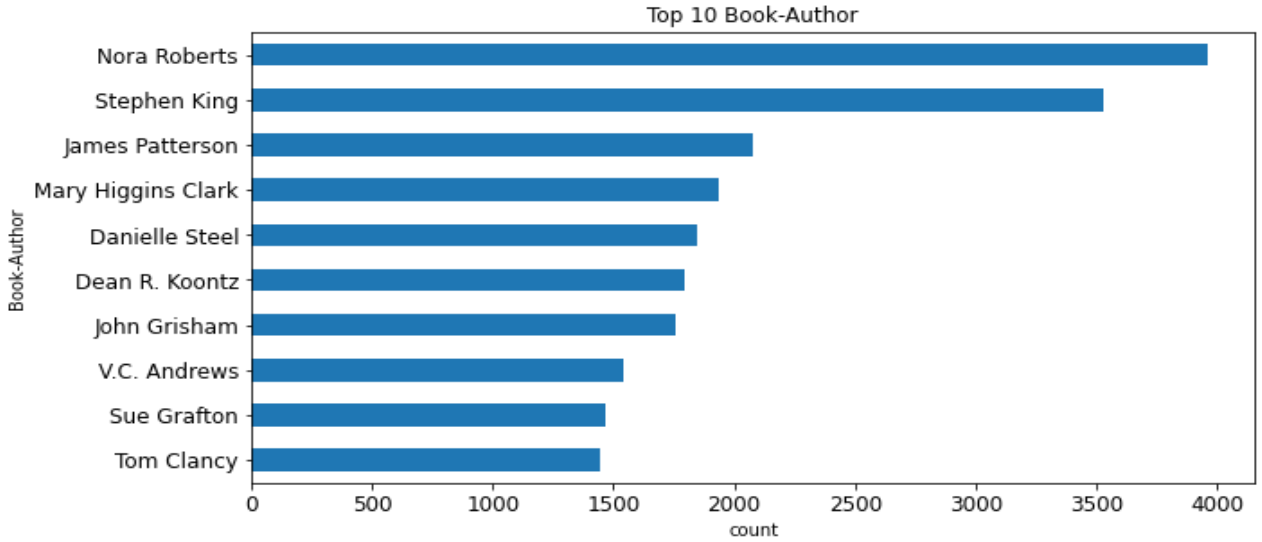
Books

- We have 8061 unique Book Title.
- The top most book title found is Wild Animus.
- **Wild Animus** is present in 365 rows.



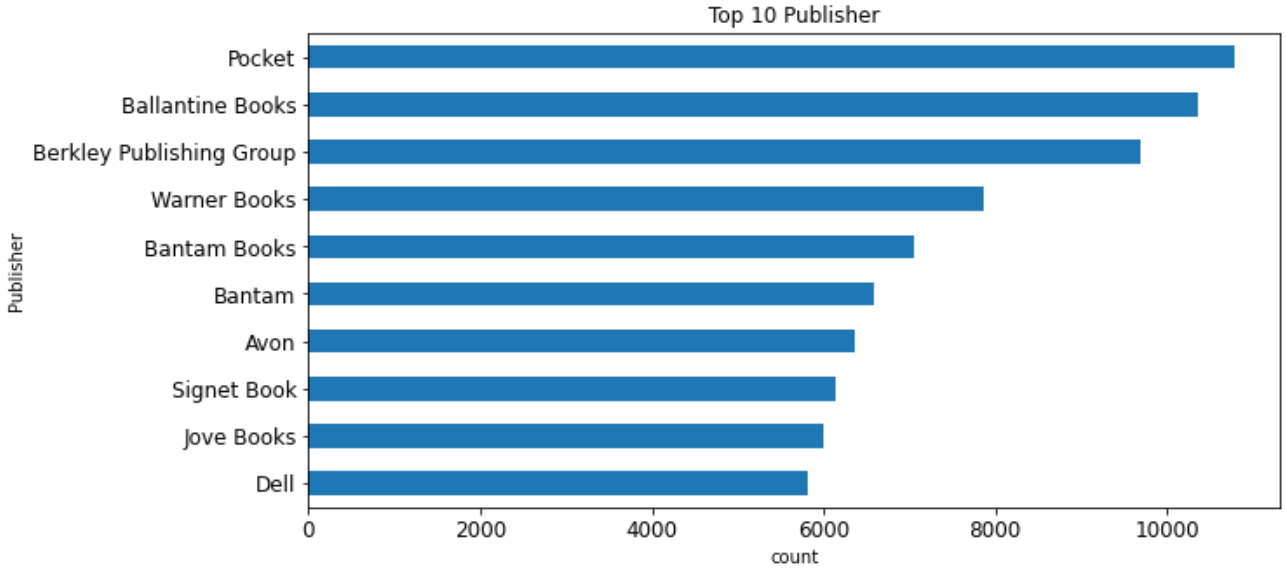
Book Authors

- **We have 4249 unique Authors.**
- **The top most author found is Nora Roberts.**
- **3958 Books written By Nora Roberts**

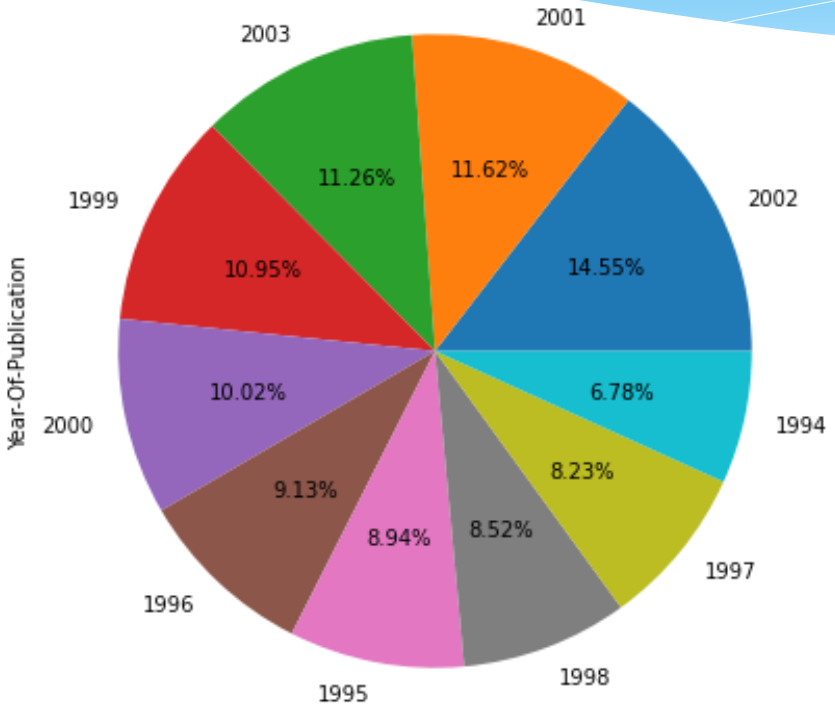


Publisher

- **We have 1075 unique Authors.**
- **The top most Publishers found is Pocket..**
- **10786 Books Published By Pocket.**



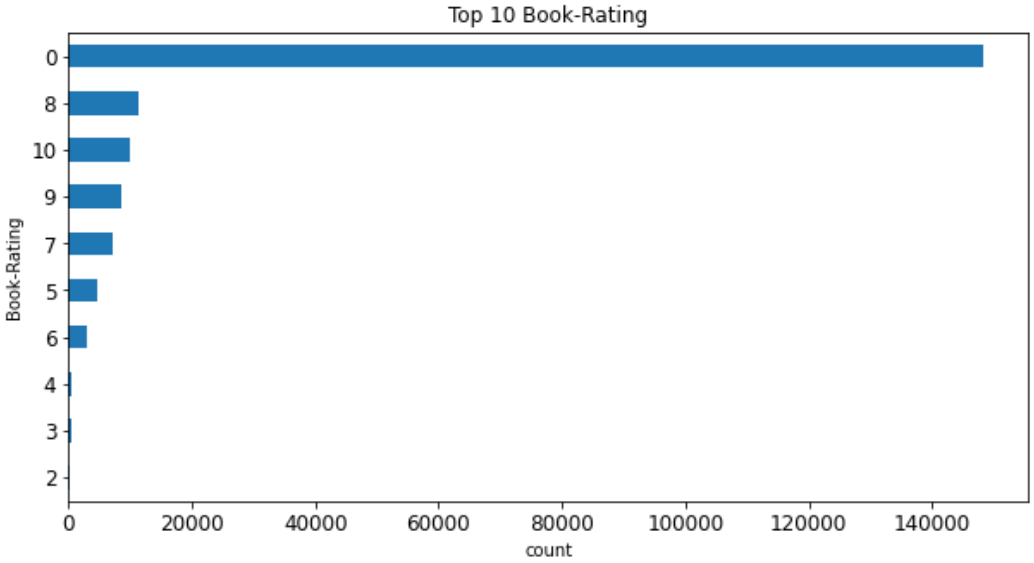
Year of Publication



Most books were published in the year of 2002.

Book Rating

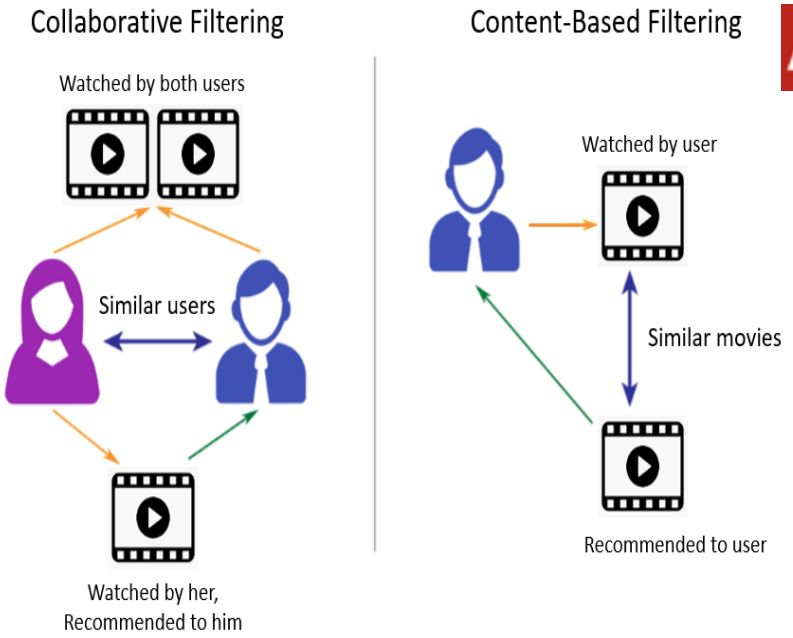
- **Highest rating is 8 out of 10.**
- **Lowest rating is 1 out of 10.**
- **Most of the people haven't given the ratings.**

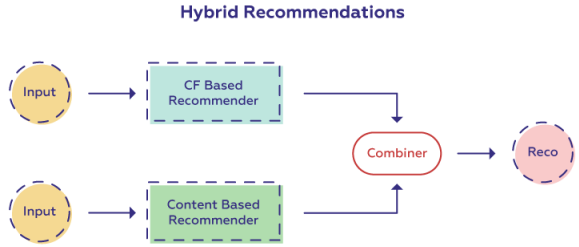


Model Creation

Recommendation system is usually classified on rating estimation:

- **Collaborative Filtering system**
- **Content based system**
- **Hybrid based system**



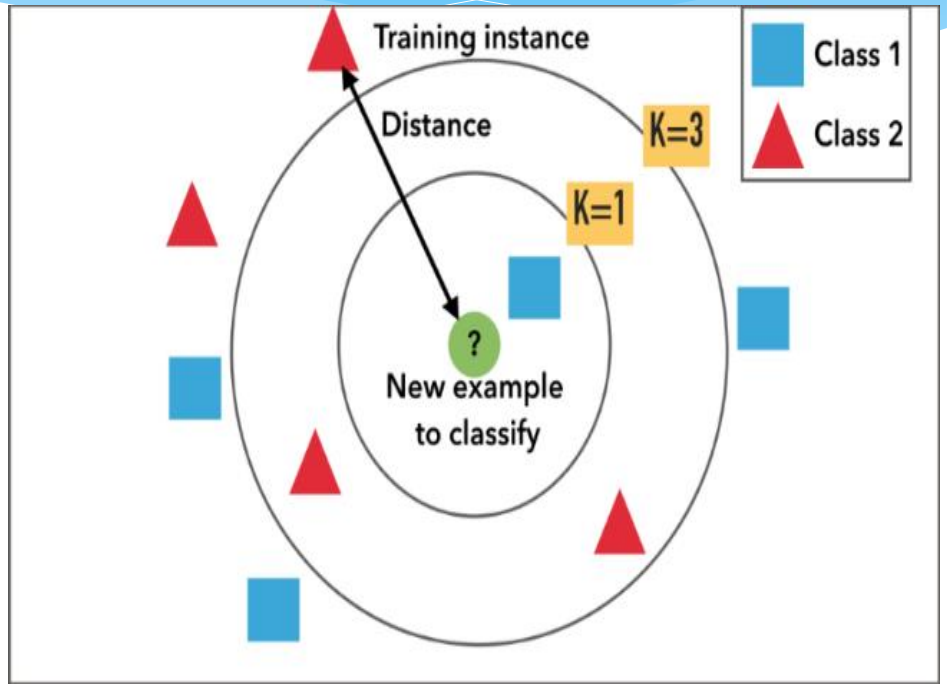


sciforce

Collaborative Filtering Using k-Nearest Neighbors (kNN)

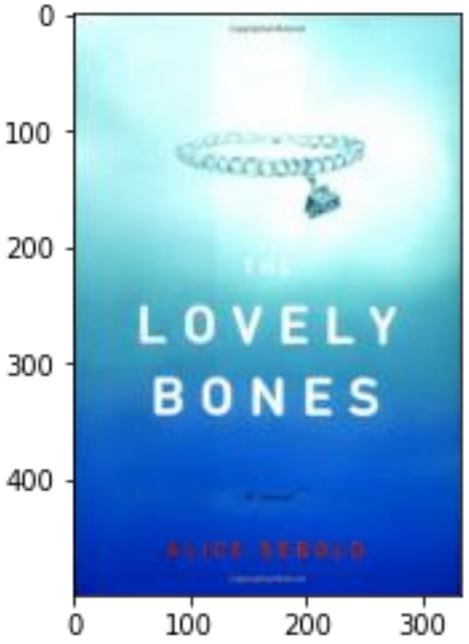
AI

kNN is a machine learning algorithm to find clusters of similar users based on common book ratings, and make predictions using the average rating of top-k nearest neighbors.

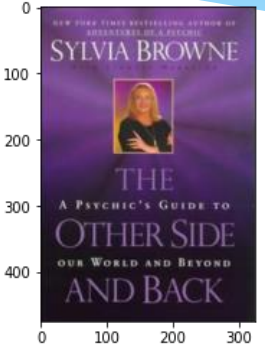


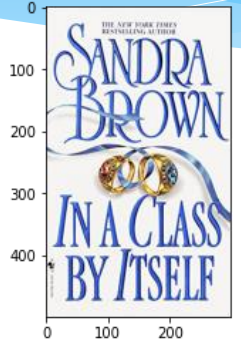
Result of K-Nearest Neighbour

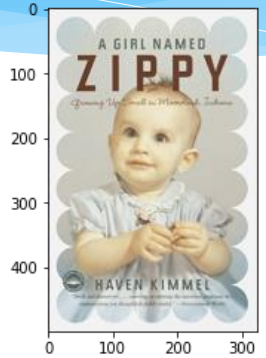
Target Book

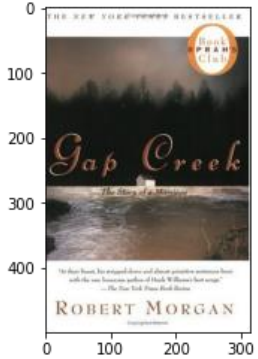


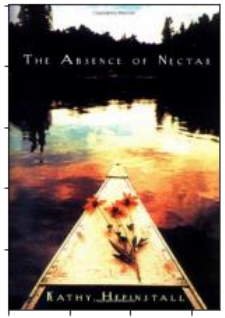
Recommendations











Data Preparation for SVD Model

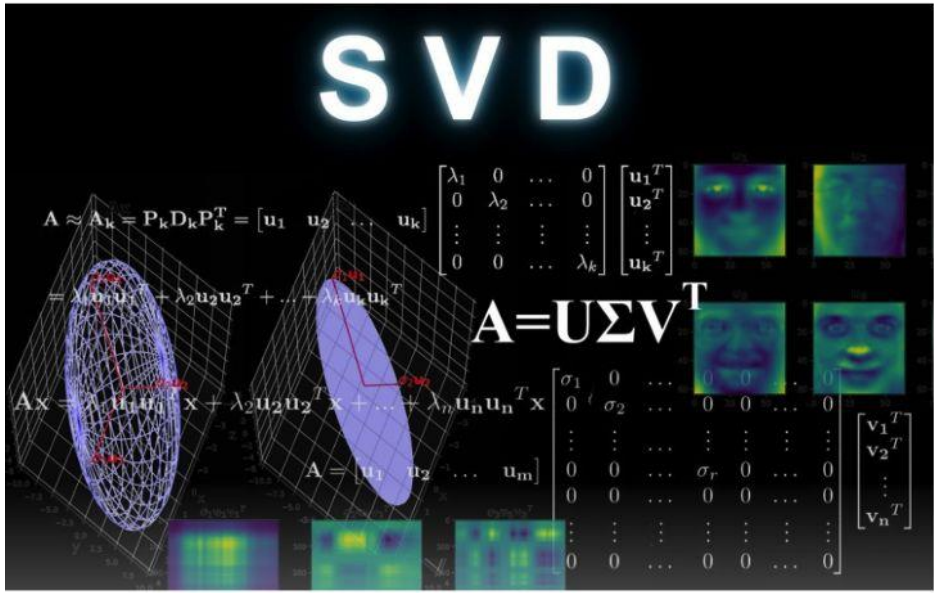
* I am Considering only those ratings that are not equal to 0
* Filter Users with at least 20 interaction
* Filter Books(ISBN) with at least 10 Ratings

Observation :

* After Dropping 0 Ratings, remaining rows are 433671.
* After Filtering out Users with at least 20 interaction & Books(ISBN) with at least 10 Ratings, remaining rows are 72385

Collaborative Filtering Using Singular Value Decomposition (SVD)

The Singular-Value Decomposition, is a matrix decomposition method for reducing a matrix to its constituent parts in order to make certain subsequent matrix calculations simpler. It provides another way to factorize a matrix, into singular vectors and singular values.



Evaluation

In Recommender Systems, there are a set metrics commonly used for evaluation. We choose to work with Top-N accuracy metrics, which evaluates the accuracy of the top recommendations provided to a user, comparing to the items the user has actually interacted in test set. This evaluation method works as follows:

1. **For each user**
2. **For each item the user has interacted in test set**
3. **Sample 100 other items the user has never interacted.**
4. **Ask the recommender model to produce a ranked list of recommended items, from a set composed of one interacted item and the 100 non-interacted items**
5. **Compute the Top-N accuracy metrics for this user and interacted item from the recommendations ranked list**
6. **Aggregate the global Top-N accuracy metrics**

Evaluation Contd.

```
Evaluating Collaborative Filtering (SVD Matrix Factorization) model...
3274 users processed

Global metrics:
{'modelName': 'Collaborative Filtering', 'recall@5': 0.9626994543068316, 'recall@10': 0.977481522414865, 'recall@15': 0.977481522414865}
```

	hits@5_count	hits@10_count	hits@15_count	interacted_count	recall@5	recall@10	recall@15	User-ID
59	380	384	384	394	0.964467	0.974619	0.974619	11676
97	67	68	68	68	0.985294	1.000000	1.000000	16795
194	63	63	63	63	1.000000	1.000000	1.000000	98391
48	60	60	60	60	1.000000	1.000000	1.000000	153662
252	55	57	57	57	0.964912	1.000000	1.000000	95359
258	52	54	54	54	0.962963	1.000000	1.000000	104636
962	53	53	53	53	1.000000	1.000000	1.000000	114368
63	41	41	41	41	1.000000	1.000000	1.000000	158295
228	41	41	41	41	1.000000	1.000000	1.000000	123883
152	39	40	40	41	0.951220	0.975610	0.975610	60244

Challenges

- **High Volume of Data.**
- **Elevating evaluation score for the models.**
- **Choosing optimal number of books and ratings.**
- **Crashing of session due to large pivot matrix.**
- **Choosing optimal number of Factors in SVD.**

Conclusion

- **A recommendation system helps an organization to create loyal customers .**
- **The recommendation system today are very powerful that they can handle the new customer too who has visited the site for the first time.**
- **They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profit to the company.**
- **A book recommendation system is a type of recommendation system where we have to recommend similar type of books to the reader based on his interest. The books recommendation system is used by online websites which provide ebooks like google play books, open library, good Read's, etc.**
- **As we can see, after implementing Collaborative Filtering and evaluating it using SVD matrix , We got a recall rate of around 97 for hit@15 is pretty good .**

THANK YOU!!