

High Level Design(HLD)

BOOK RECOMMENDATION SYSTEM(BRS)

CONTENTS

ABSTRACT

CHAPTER 1 INTRODUCTION

1.1 Introduction

1.1.1 Why this High-Level Design Document

1.1.2 Scope

1.1.3 Definitions

CHAPTER 2 General Descriptions

2.1 Product Perspective

2.2 Problem Statements

2.3 Proposed Solution

2.4 Further improvements

2.5 Data Requirements

2.7 Tools Used

2.7.1 Software requirements

2.7.2 Hardware requirements

2.8 Constraints

2.9 Assumption

CHAPTER 3 Design Details

3.1 Process flow

3.1.1 Model Training and Evaluation

3.1.2 Deployment Process

3.2 Event log

3.3 Error handling

3.4 Performance

3.5 Reusability

3.6 Application compatibility

3.7 Resource utilization

3.8 Deployment

CHAPTER 4 Dash Board

4.1 KPIs (Key Performance Indicator)

CHAPTER 5 Conclusion

ABSTRACT

Now-a-days, everyone depending on reviews by others in many things such as selecting a movie to watch, buying products, reading a book. Recommender systems are used for that purpose only. A recommender system is a kind of filtering system that predicts a user's rating of an item. Recommender systems recommend items to users by filtering through a large database of information using a ranked list of predicted ratings of items. Online Book recommender system is a recommender system for ones who love books. When selecting a book to read, individuals read and rely on the book ratings and reviews that previous users have written. In this paper, Hybrid Recommender system is used in which Collaborative Filtering and Content-Based Filtering techniques are used. The author used Collaborative techniques such as Clustering in which data-points are grouped into clusters. Algorithms such as K-means clustering and Gaussian mixture are used for clustering. The better algorithm was selected with the help of silhouette score and used for clustering. Matrix Factorization technique such as Truncated-SVD which takes sparse matrix as input is used for reducing the features of a dataset. Content Based Filtering System used TF-IDF vectorizer which took statements as input and return a matrix of vectors. RMSE (Root Mean Square Error) is used for finding the deviation of an absolute value from an obtained value and that value is used for finding the fundamental accuracy.

CHAPTER 1 Introduction

1.1 Why this High-Level Design Document

The purpose of this High-level Design Document is to add the necessary details to the current project description to represent suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

1. Present all of the design aspects and define them in details
2. Describe the user interface being implemented
3. Describe the hardware and software interface
4. Describe the performance requirements
5. Include design features and the architectural of the project
6. List and describe the non-functional attribute like

1.1.2 Future Scope:-

The System has adequate scope for modification in future if it is necessary. Development and launching of Mobile app and refining existing services and adding more service, System security, data security and reliability are the main features which can be done in future. The API for the shopping and payment gateway can be added so that we can also buy a book at the moment. In the existing system there are only some selected categories, so as an extension to the site we can add more categories as compared to existing site. Also we can add admin side with some functionalities like books management, User management etc.

1.1.3 Definitions:-

Now-a-days, online rating and reviews are playing an important role in books sales. Readers were buying books depend on the reviews and ratings by the others. Recommender system focuses on the reviews and ratings by the others and filters books. In this paper, Hybrid recommender system is used to boost our recommendations. The technique used by recommender systems is Collaborative filtering. This technique filters information by collecting data from other users. Collaborative filtering systems apply the similarity index-based technique. The ratings of those items by the users who have rated both items determine the similarity of the items. The similarity of users is determined by the similarity of the ratings given by the users to an item. Content-based filtering uses the description of the items and gives recommendations which are similar to the description of the items. With these two filtering systems, books are recommended not only based on the user's behaviour but also with the content of the books. So, our recommendation system recommends books to the new users also. In this recommender system, books are recommended based on

collaborative filtering technique and similar books are shown using content based filtering.

The required dataset for the training and testing of our model is downloaded from Good-Reads website. Matrix Factorization technique such as Truncated-SVD which takes sparse matrix of dataset is used for reduction of features. The reduced dataset is used for clustering to build a recommendation system. Clustering is a collaborative filtering technique that is used to build our recommendation system in which data points are grouped into clusters. . In this paper, we used two methods i.e., K-means and Gaussian mixture for clustering the users. The better model is selected based on the silhouette score and used for clustering. Silhouette score or silhouette coefficient is used to calculate how good the clustering is done. Negative value shows that clustering is imperfect whereas positive value shows that clustering was done perfectly. Difference between the mean rating before clustering and after clustering is calculated. Root Mean square Error is used to measure the error between the absolute 2 values and obtained values. That RMSE value is used to find the fundamental accuracy.

CHAPTER 2 General Description

2.1 Product Perspective

In the present world, all products are buying based on the reviews and ratings by the others. There are so many products with high rating and reviews but we only put our interest in some products which we like. Recommender system works on this principle only i.e. it recommends products based on the interest of the users. Our idea is to create recommender system that recommends books based on the user's interest i.e. we recommends books which are similar to the books that user already liked. It can also recommend books which are liked by similar

users. Similar users are those who liked the books which are liked by the current user.

We also add another feature i.e. we recommends books which are independent of the users interest. With this feature, we can recommend books to the new users also. Book recommendation sites that were available online now a days shows the books which are recommended by the system. Here, we are also recommending books based on the description of the book. We will get books which are similar to the book we selected in this system. For that purpose only, we built Hybrid recommender system. Hybrid recommender system is a combination of Collaborative Filtering system and Content Based Filtering system.

2.2 Problem Statements:-

Recommending books using Machine learning algorithm is the main goal of this project. Books are recommended by the clustering model and we are going to train and build using various features such as user's rating, book description, book titles etc. The system groups users into clusters so that each data point within cluster is similar and dissimilar to the data point in the other cluster. The system we would like to develop will also be able to find an average rating for each cluster and it is going to find top rated books of users from each cluster. All these books shortlisted by our system will be used for training our model in future. The prediction model needs to be trained so as to produce better results.

2.3 Proposed Solution

2.3.1:-System Architecture

System Architecture describes “the overall structure of the system and the ways in which the structure provides conceptual integrity”. The system architecture to build a recommendation system involves the following five major steps.

2.3.1 Data Acquisition

The goal of this step is to find and acquire all the related datasets or data sources. In this step, the main aim is to identify various available data sources, as data are often collected from various online sources like databases and files. The size and the quality of the data in the collected dataset will determine the efficiency of the model. The Books dataset is collected from the Kaggle Dataset website.

Fig-3.2 Sample of acquired books dataset from Kaggle Dataset Website

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ISBN	Book-Title	Book-Autl	Year-Of-P	Publisher	Image-URL-S	Image-URL-M	Image-URL-L					
2	195153448	Classical Mythology	Mark P. O.	2002	Oxford University Press	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg					
3	2005018	Clara Callan	Richard Br	2001	HarperFlamingo Canada	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg					
4	60973129	Decision in Normand	Carlo D'Es	1991	HarperPerennial	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg					
5	374157065	Flu: The Story of the	Gina Bari	1999	Farrar Straus Giroux	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg					
6	393045218	The Mummies of Uru	E. J. W. Ba	1999	W. W. Norton & Compa	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0393045218.01.LZZZZZZZ.jpg					
7	399135782	The Kitchen God's Wl	Amy Tan	1991	Putnam Pub Group	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0399135782.01.LZZZZZZZ.jpg					
8	425176428	What If?: The World's	Robert Co	2000	Berkley Publishing Group	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0425176428.01.LZZZZZZZ.jpg					
9	671870432	PLEADING GUILTY	Scott Turo	1993	Audioworks	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0671870432.01.LZZZZZZZ.jpg					
10	679425608	Under the Black Flag	David Cori	1996	Random House	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0679425608.01.LZZZZZZZ.jpg					
11	074322678X	Where You'll Find Me	Ann Beatt	2002	Scribner	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/074322678X.01.LZZZZZZZ.jpg					
12	771074670	Nights Below Station	David Ada	1988	Emblem Editions	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0771074670.01.LZZZZZZZ.jpg					
13	080652121X	Hitler's Secret Banke	Adam Leb	2000	Citadel Press	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/080652121X.01.LZZZZZZZ.jpg					
14	887841740	The Middle Stories	Sheila Het	2004	House of Anansi Press	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0887841740.01.LZZZZZZZ.jpg					
15	1552041778	Jane Doe	R. J. Kaise	1999	Mira Books	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1552041778.01.LZZZZZZZ.jpg					
16	1558746218	A Second Chicken So	Jack Canfi	1998	Health Communications	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1558746218.01.LZZZZZZZ.jpg					
17	1567407781	The Witchfinder (Am	Loren D. E	1998	Brilliance Audio - Trade	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1567407781.01.LZZZZZZZ.jpg					
18	1575663937	More Cunning Than A	Robert He	1999	Kensington Publishing Corp.	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1575663937.01.LZZZZZZZ.jpg					
19	1881320189	Goodbye to the Butt	Julia Olive	1994	River City Pub	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1881320189.01.LZZZZZZZ.jpg					
20	440234743	The Testament	John Grish	1999	Dell	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0440234743.01.LZZZZZZZ.jpg					
21	452264464	Beloved (Plume Con	Toni Morr	1994	Plume	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0452264464.01.LZZZZZZZ.jpg					
22	609804618	Our Dumb Century: T	The Onior	1999	Three Rivers Press	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0609804618.01.LZZZZZZZ.jpg					
23	1841721522	New Vegetarian: Bol	Celia Broc	2001	Ryland Peters & Small	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1841721522.01.LZZZZZZZ.jpg					
24	1879384493	If I'd Known Then W	J. R. Parris	2003	Cypress House	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/1879384493.01.LZZZZZZZ.jpg					
25	61076031	Mary-Kate & Asl	Mary-Kate	2000	HarperEntertainment	http://images.amaz	http://images.ama	http://images.amazon.com/images/P/0061076031.01.LZZZZZZZ.jpg					

In the above Fig, we can see a sample of the dataset we have collected. This acquired dataset has around 3,000 books and has 8 different features. The features are listed below:

- ISBN

- Book-Title
- Book-Author
- Year-Of-Publication
- Publisher
- Image-URL-S
- Image-URL-M
- Image-URL-L

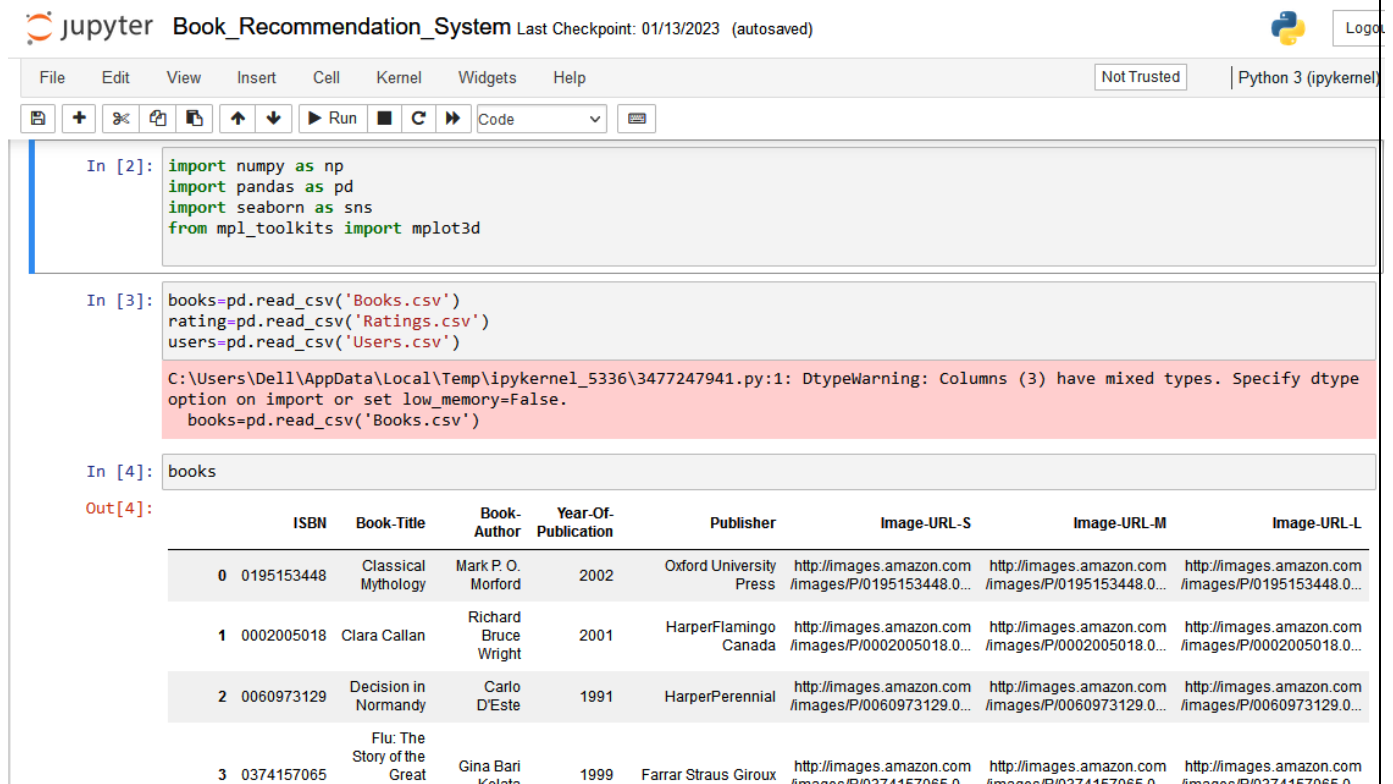
One more dataset i.e. ratings dataset was also collected from Kaggle Dataset website.

	A	B	C	D
1	User-ID	ISBN	Book-Rating	
2	276725	034545104X	0	
3	276726	155061224	5	
4	276727	446520802	0	
5	276729	052165615X	3	
6	276729	521795028	6	
7	276733	2080674722	0	
8	276736	3257224281	8	
9	276737	600570967	6	
10	276744	038550120X	7	
11	276745	342310538	10	
12	276746	425115801	0	
13	276746	449006522	0	
14	276746	553561618	0	
15	276746	055356451X	0	
16	276746	786013990	0	
17	276746	786014512	0	
18	276747	60517794	9	
19	276747	451192001	0	
20	276747	609801279	0	
21	276747	671537458	9	
22	276747	679776818	8	
23	276747	943066433	7	
24	276747	1570231028	0	
25	276747	1885408226	7	

Above Fig Sample of acquired ratings dataset from Kaggle Dataset Website
In the above Fig we can see a sample of the dataset we have collected. This acquired dataset has around 400000 ratings and has 3 different features.

- User-ID
- ISBN
- Book-Rating

After acquiring the data our next step is to read the data from the csv file into python notebook. Python notebook is used in our project for data pre-processing, features selection and for model comparison. In the fig-we have read data from csv file using the inbuilt python functions that are part of pandas library.



The screenshot shows a Jupyter Notebook titled "Book_Recommendation_System" with a last checkpoint of "01/13/2023 (autosaved)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running code, and viewing output. The notebook contains three code cells:

```
In [2]: import numpy as np
import pandas as pd
import seaborn as sns
from mpl_toolkits import mplot3d
```

```
In [3]: books=pd.read_csv('Books.csv')
rating=pd.read_csv('Ratings.csv')
users=pd.read_csv('Users.csv')
```

A warning message is displayed: "C:\Users\Dell\AppData\Local\Temp\ipykernel_5336\3477247941.py:1: DtypeWarning: Columns (3) have mixed types. Specify dtype option on import or set low_memory=False." followed by the code: `books=pd.read_csv('Books.csv')`

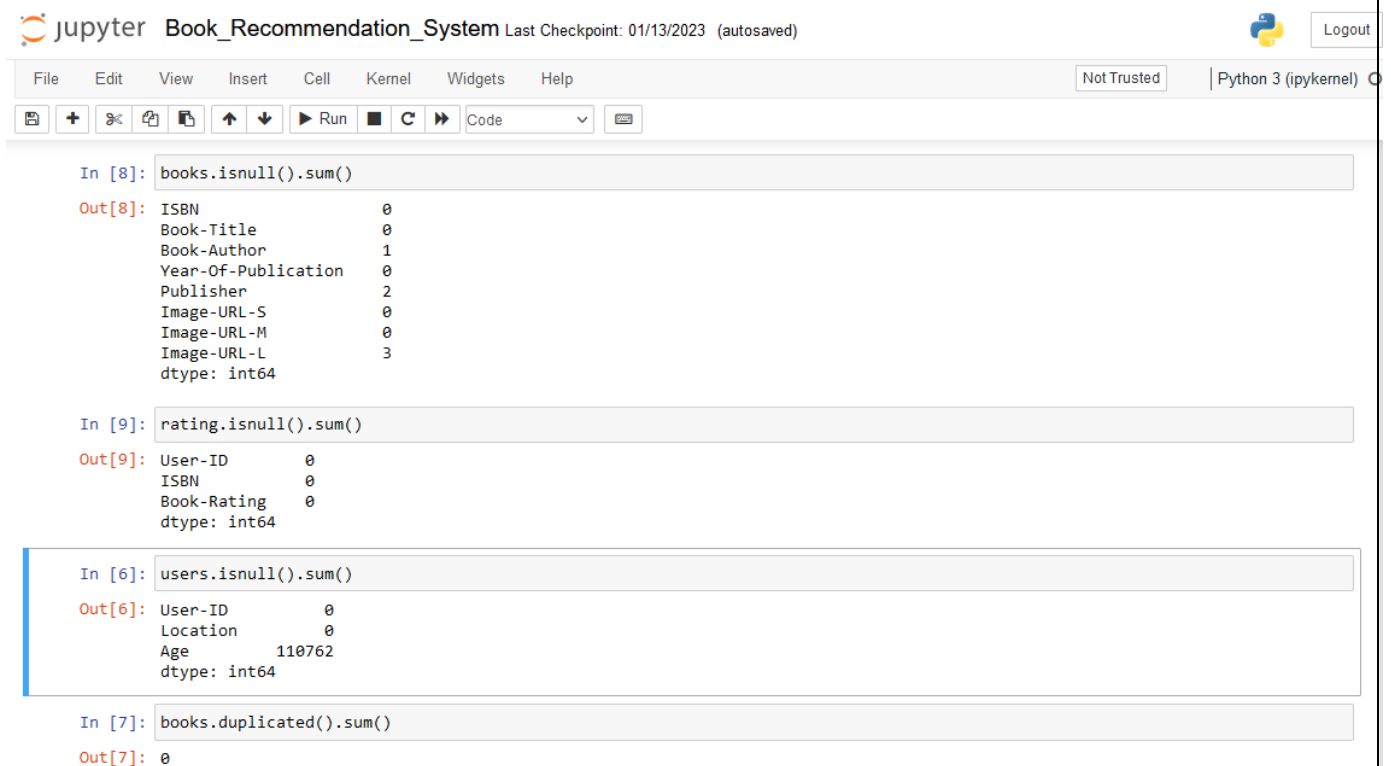
```
In [4]: books
```

The output of the fourth cell is a preview of the 'Books' dataset, showing the first four rows of a table with the following columns: ISBN, Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M, and Image-URL-L.

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S	Image-URL-M	Image-URL-L
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/0195153448.0...
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/0002005018.0...
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...	http://images.amazon.com/images/P/0060973129.0...
3	0374157065	Flu: The Story of the Great	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...	http://images.amazon.com/images/P/0374157065.0...

2.3.2 Data Pre-processing

The goal of this step is to study and understand the nature of data that was acquired in the previous step and also to know the quality of data. In this step, we will check for any null values and remove them as they may affect the efficiency. Identifying duplicates in the dataset and removing them is also done in this step.



```
In [8]: books.isnull().sum()
Out[8]: ISBN                0
Book-Title                0
Book-Author              1
Year-Of-Publication       0
Publisher                 2
Image-URL-S              0
Image-URL-M              0
Image-URL-L              3
dtype: int64

In [9]: rating.isnull().sum()
Out[9]: User-ID           0
ISBN                 0
Book-Rating          0
dtype: int64

In [6]: users.isnull().sum()
Out[6]: User-ID           0
Location            0
Age              110762
dtype: int64

In [7]: books.duplicated().sum()
Out[7]: 0
```

2.3.3 Feature Extraction

After pre-processing the acquired data, the next step is to reduce the features i.e. Dimensionality reduction. The reduced features should be able to give high efficiency.

```
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)
In [30]: popular_df=popular_df[popular_df['num_rating']>=250].sort_values('avg_rating',ascending=False).head(50)
In [31]: popular_df
Out[31]:
```

	Book-Title	num_rating	avg_rating
80434	Harry Potter and the Prisoner of Azkaban (Book 3)	428	5.852804
80422	Harry Potter and the Goblet of Fire (Book 4)	387	5.824289
80441	Harry Potter and the Sorcerer's Stone (Book 1)	278	5.737410
80426	Harry Potter and the Order of the Phoenix (Boo...	347	5.501441
80414	Harry Potter and the Chamber of Secrets (Book 2)	556	5.183453
191612	The Hobbit: The Enchanting Prelude to The Lor...	281	5.007117
187377	The Fellowship of the Ring (The Lord of the RI...	368	4.948370
80445	Harry Potter and the Sorcerer's Stone (Harry P...	575	4.895652
211384	The Two Towers (The Lord of the Rings, Part 2)	260	4.880769
219741	To Kill a Mockingbird	510	4.700000
183573	The Da Vinci Code	898	4.642539
187880	The Five People You Meet in Heaven	430	4.551163
180556	The Catcher in the Rye	449	4.545657
196326	The Lovely Bones: A Novel	1295	4.468726
764	1984	284	4.454225
444465	Prodigal Summer: A Novel	252	4.450502

2.3.4 Training Methods

Now, we have our training and testing data. The next step is to identify the possible training methods and train our models. We have used two different clustering methods for training models. After that based on the silhouette score of each model, we would decide on which model to use finally.

2.3.5 Testing Data

In Step 2.3.1, Dataset was collected from Kaggle Dataset Website in which three datasets are present i.e. Books Dataset, Ratings Dataset, Users Dataset. In Step 2.3.2, Datasets were pre-processed to make suitable for developing the Recommendation system. In Step 2.3.3, Feature extraction is performed in which Truncated-SVD is used to reduce the features of the dataset and Data splitting is done in which training dataset and testing dataset are divided into 80:20 ratio. In Step 2.3.4, Content Based Filtering System is developed in which book description is taken as an input and Collaborative Filtering System is developed by building a model using K-Means Algorithm over Gaussian

Mixture after comparing with Silhouette scores. In step 2.3.5, Testing of model with test data is performed.

2.4 Further improvements

- Given more information regarding the books dataset, namely features like Genre, Description etc., we could implement a content-filtering based recommendation system and compare the results with the existing collaborative-filtering based system.
- We would like to explore various clustering approaches for clustering the users based on Age, Location etc., and then implement voting algorithms to recommend items to the user depending on the cluster into which it belongs.

2.5 Data Requirements

The goal of this step is to find and acquire all the related datasets or data sources. In this step, the main aim is to identify various available data sources, as data are often collected from various online sources like databases and files. The size and the quality of the data in the collected dataset will determine the efficiency of the model. The Books dataset is collected from the Kaggle Dataset website.

Sample of acquired dataset from Kaggle Dataset Website it contain three csv file like:-

Books.csv

Users.csv

Rating.csv

2.7 Tools Used

2.7.1 **Software Requirements**

1. Pycharm used as IDE
2. For visualization of plots, Matplotlib, Seaborn and plotly are used.
3. Azure is used for deployment of the model
4. MongoDB is used to retrieve, insert, delete and update database
5. Frontend development is done using HTML/CSS
6. Python flask is used for backend development
7. Github used for version control

2.7.2 **Hardware Requirements**

1. RAM: 4 GB or above
2. Storage: 30 to 50 GB
3. Processor: Any Processor above 500MHz

2.8 Constraints

Recommender systems **provide valuable support for users who are searching for products and services in e-commerce environments**. Research in the field long focused on algorithms supporting the recommendation of quality taste products such as news, books, or movies.

2.9 Assumption

A recommendation engine is a class of machine learning which offers relevant suggestions to the customer. Before the recommendation system, the major tendency to buy was to take a suggestion from friends. But Now Google knows

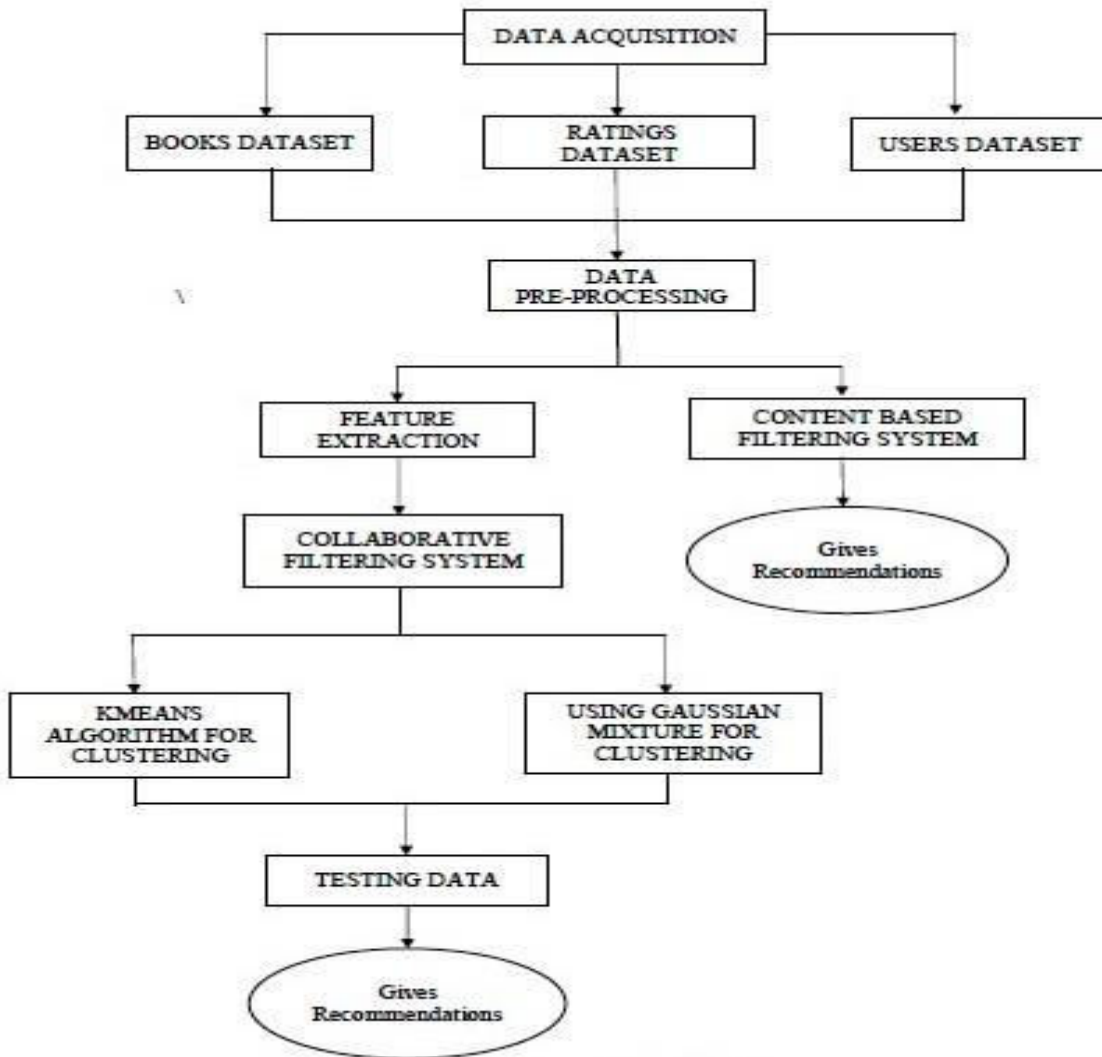
what news you will read, Youtube knows what type of videos you will watch based on your search history, watch history, or purchase history.

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. The recommendation system today are so 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.

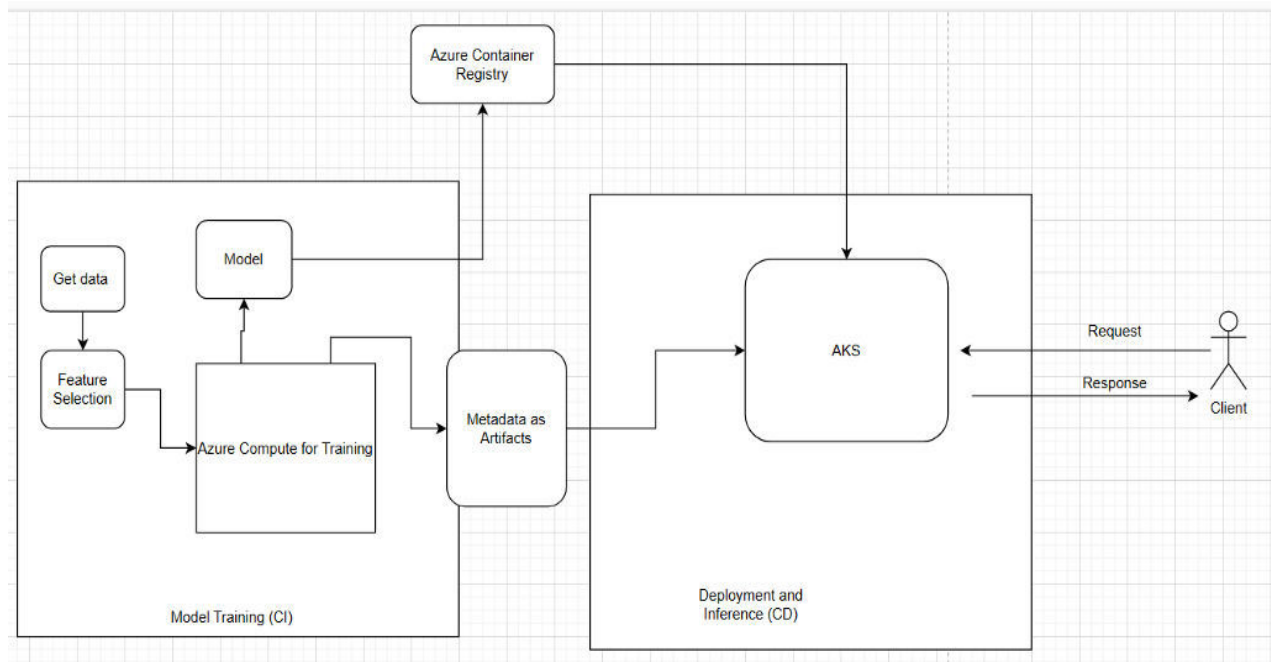
CHAPTER 3 Design Details

3.1 Process flow

3.1.1 Model Training and Evaluation



3.1.2 Deployment Process



3.2 Event log

The system should log every event so that the user will know what will know what process is running internally.

Initial Step-By-Step Description

- 1.The System identifies at what step logging required.
- 2.The System should be able to log each and every system flow.
- 3.Developer can choose logging method. You can choose database logging/File logging as well.
- 4.System should not hang even after using so many loggings. Logging just because we can easily debug issue so logging issues so logging mandatory to do.

3.3 Error handling.

Should error be encountered an explanation will be displayed as to what went wrong. An error will be defined as anything that falls outside the normal and intended usage.

3.4 Performance

3.5 Reusability

The code written and the components used should have the ability to be reused with no problems.

3.6 Application compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform and it is the job of the python to ensure proper transfer of information

3.7 Resource utilization

When any task is performed, it will likely use all the processing power available until that fucation is finished.

3.8 Deployment









CHAPTER 4 Dash Board

4.1 KPIs (Key Performance Indicator)

Book Recommendation System

Home Recommendation contact

Top 50 Books

 <p>Harry Potter and the Prisoner of Azkaban (Book 3)</p> <p>J. K. Rowling</p> <p>Vote - 428</p> <p>Rating- 5.852803738317757</p>	 <p>Harry Potter and the Goblet of Fire (Book 4)</p> <p>J. K. Rowling</p> <p>Vote - 387</p> <p>Rating- 5.8242894056847545</p>	 <p>Harry Potter and the Sorcerer's Stone (Book 1)</p> <p>J. K. Rowling</p> <p>Vote - 278</p> <p>Rating- 5.737410071942446</p>	 <p>Harry Potter and the Order of the Phoenix (Book 5)</p> <p>J. K. Rowling</p> <p>Vote - 347</p> <p>Rating- 5.501440922190202</p>	 <p>Harry Potter and the Chamber of Secrets (Book 2)</p> <p>J. K. Rowling</p> <p>Vote - 556</p> <p>Rating- 5.183453237410072</p>	 <p>The Hobbit: The Enchanting Prelude to The Lord of the Rings</p> <p>J.R.R. TOLKIEN</p> <p>Vote - 281</p> <p>Rating- 5.00711743772242</p>
--	--	---	---	---	--

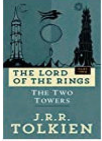
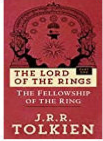


Book Recommendation System

Home Recommendation contact

Recommend Book

The Hobbit: The Enchanting Prelude to The Lord of the Rings

Submit Query

 <p>"J.R.R. TOLKIEN"</p> <p>"The Two Towers (The Lord of the Rings, Part 2)"</p>	 <p>"J.R.R. TOLKIEN"</p> <p>"The Fellowship of the Ring (The Lord of the Rings, Part 1)"</p>	 <p>"Wilson Rawls"</p> <p>"Where the Red Fern Grows"</p>	 <p>"Janet Evanovich"</p> <p>"One for the Money (A Stephanie Plum Novel)"</p>
---	---	--	--

CHAPTER 5 Conclusion

In this project, we have recommended the books for a user using the model trained using K-Means Clustering which is a Collaborative Filtering Technique. We have also compared different models built using different methods and identified the best model and justifies why it has chosen that model. We have used the books dataset that is available in the Kaggle Dataset website which consists of more than 3000 books. The models are built using the reduced features which is done by Truncated SVD. Based on those features the author built a model that gives a positive Silhouette score. The model that is suggested by this paper is useful for book readers. The system we have developed can make recommendations for new users also.

LOW LEVEL DOCUMENT (BRS)

Contents

1. Introduction

1.1. What is Low-Level design document?

1.2. Scope

2. Architecture

3. Architecture Description

3.1. Data Description

3.2. Web Scrapping

3.3. Data Transformation

3.4. Data Insertion into Database

3.5. Export Data from Database

3.6. Data Pre-processing

3.7. Data Clustering

3.10. Model Building

3.11. Data from User

3.12. Data Validation

3.13. User Data Inserting into Database

3.14. Data Clustering

3.15. Model Call for Specific Cluster

3.16. Deployment

4. Unit Test Cases

1. Introduction

1.1. What is Low-Level design document?

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code for Book Recommendation System. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

1.2. Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work

2. Architecture

3. Architecture Description

3.1. Data Description

We have 3 files in our dataset which is extracted from some books selling websites.

- Books – first are about books which contain all the information related to books like an author, title, publication year, etc.

- Users – The second file contains registered user's information like user id, location.
- Ratings – Ratings contain information like which user has given how much rating to which book.

So based on all these three files we can build a powerful collaborative filtering model. let's get started.

3.2. Web Scrapping

Web scraping (or data scraping) is a technique used to collect content and data from the internet. This data is usually saved in a local file so that it can be manipulated and analyzed as needed. If you've ever copied and pasted content from a website into an Excel spreadsheet, this is essentially what web scraping is, but on a very small scale.

all web scraping bots follow three basic principles:

- Step 1: Making an HTTP request to a server
- Step 2: Extracting and parsing (or breaking down) the website's code
- Step 3: Saving the relevant data locally

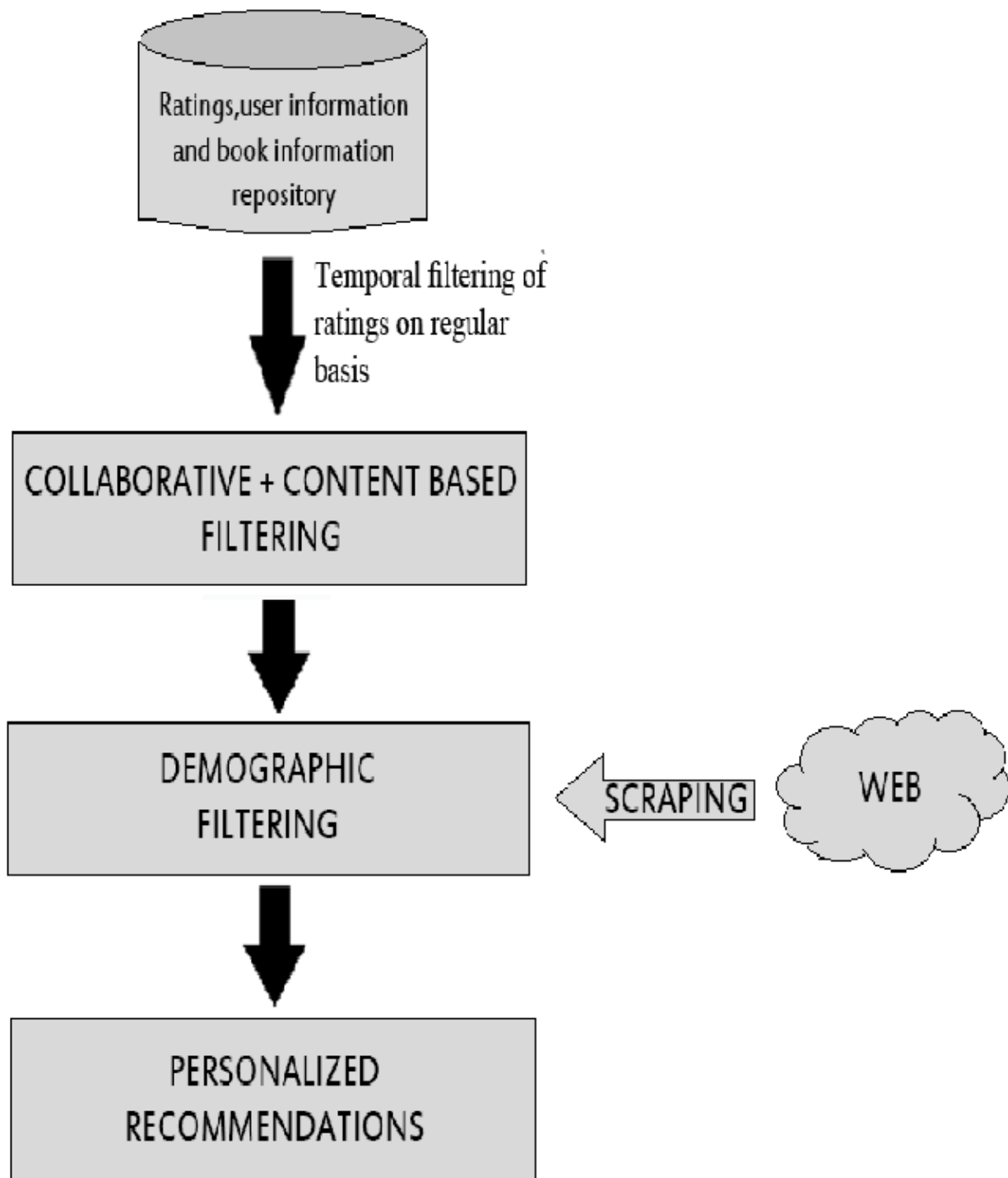


Fig. 2. Workflow of the Recommendation Process

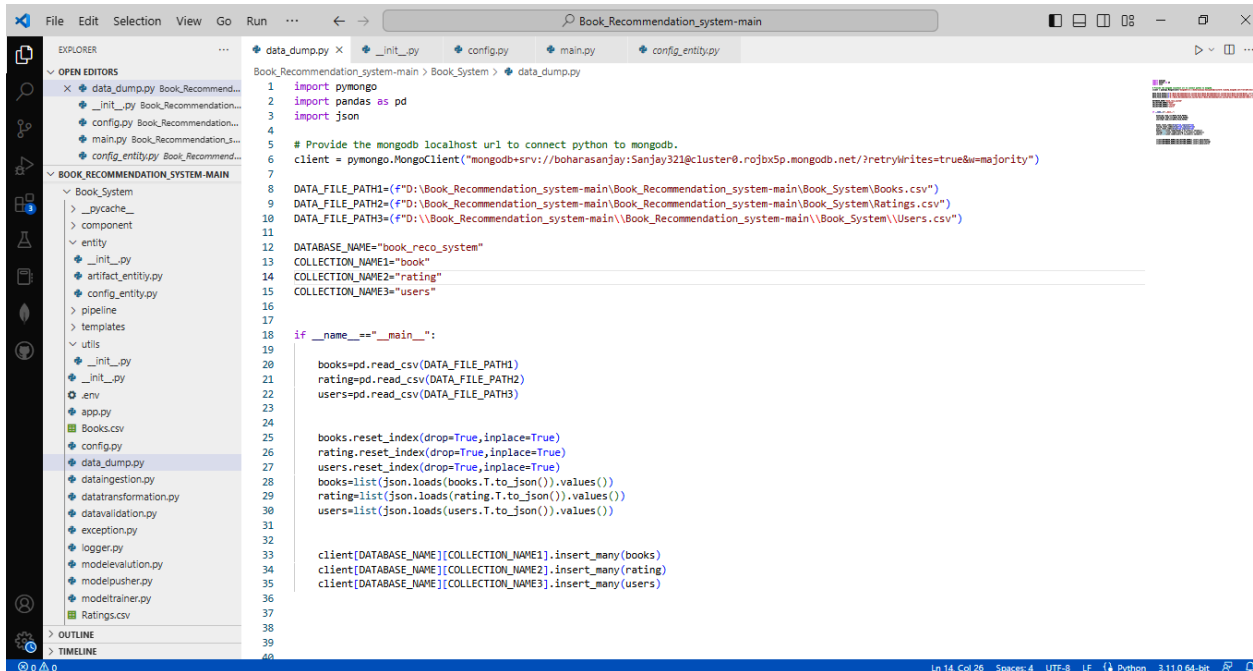
3.3. Data Transformation

Let us start while importing libraries and load datasets. While loading the file we have some problems like.

- The values in the CSV file are separated by semicolons, not by a comma.
- There are some lines which not work like we cannot import it with pandas and It throws an error because python is Interpreted language.
- Encoding of a file is in Latin

So while loading data we have to handle these exceptions and after running the below code you will get some warning and it will show which lines have an error that we have skipped while loading.

3.4. Data Insertion into Database



```
1 import pymongo
2 import pandas as pd
3 import json
4
5 # Provide the mongodb localhost url to connect python to mongodb.
6 client = pymongo.MongoClient("mongodb+srv://boharasanjay:Sanjay321@cluster0.rojbx5p.mongodb.net/?retryWrites=true&majority")
7
8 DATA_FILE_PATH1=(f"D:\\Book_Recommendation_system-main\\Book_Recommendation_system-main\\Book_System\\Books.csv")
9 DATA_FILE_PATH2=(f"D:\\Book_Recommendation_system-main\\Book_Recommendation_system-main\\Book_System\\Ratings.csv")
10 DATA_FILE_PATH3=(f"D:\\Book_Recommendation_system-main\\Book_Recommendation_system-main\\Book_System\\Users.csv")
11
12 DATABASE_NAME="book_reco_system"
13 COLLECTION_NAME1="book"
14 COLLECTION_NAME2="rating"
15 COLLECTION_NAME3="users"
16
17
18 if __name__=="__main__":
19
20     books=pd.read_csv(DATA_FILE_PATH1)
21     rating=pd.read_csv(DATA_FILE_PATH2)
22     users=pd.read_csv(DATA_FILE_PATH3)
23
24
25     books.reset_index(drop=True,inplace=True)
26     rating.reset_index(drop=True,inplace=True)
27     users.reset_index(drop=True,inplace=True)
28     books=list(json.loads(books.T.to_json()).values())
29     rating=list(json.loads(rating.T.to_json()).values())
30     users=list(json.loads(users.T.to_json()).values())
31
32
33     client[DATABASE_NAME][COLLECTION_NAME1].insert_many(books)
34     client[DATABASE_NAME][COLLECTION_NAME2].insert_many(rating)
35     client[DATABASE_NAME][COLLECTION_NAME3].insert_many(users)
36
37
38
39
40
```

3.5. Export Data from Database

Data Export from Database - The data in a stored database is exported as a CSV file to be used for Data Pre-processing and Model Training.

3.6. Data Pre-processing

Now in the books file, we have some extra columns which are not required for our task like image URLs. And we will rename the columns of each file as the name of the column contains space, and uppercase letters so we will correct as to make it easy to use.

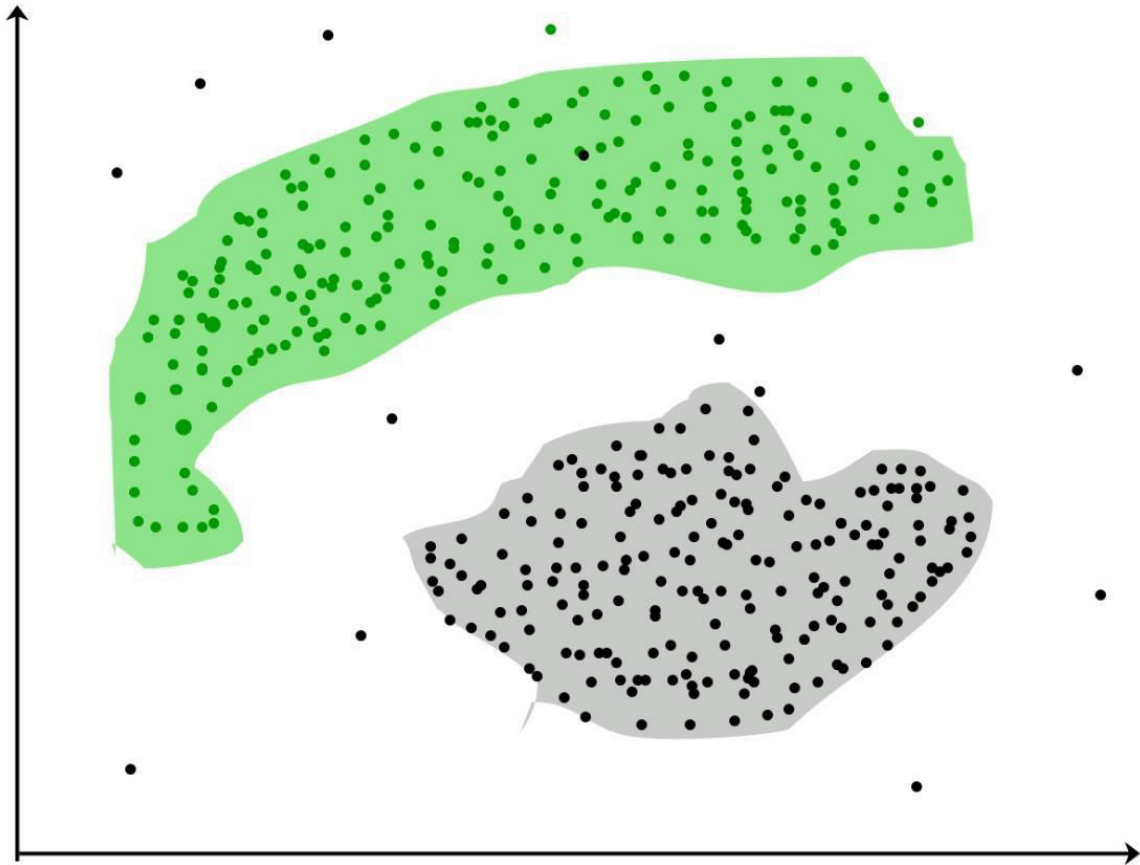
```
books = books[['ISBN', 'Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher']]
```

```
books.rename(columns = {'Book-Title':'title', 'Book-Author':'author',  
'Year-Of-Publication':'year', 'Publisher':'publisher'}, inplace=True)  
users.rename(columns = {'User-ID':'user_id', 'Location':'location',  
'Age':'age'}, inplace=True)  
ratings.rename(columns = {'User-ID':'user_id', 'Book-Rating':'rating'},  
inplace=True)
```

3.7. Data Clustering

Clustering is an unsupervised learning method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them. Clustering is very important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.



3.10. Model Building

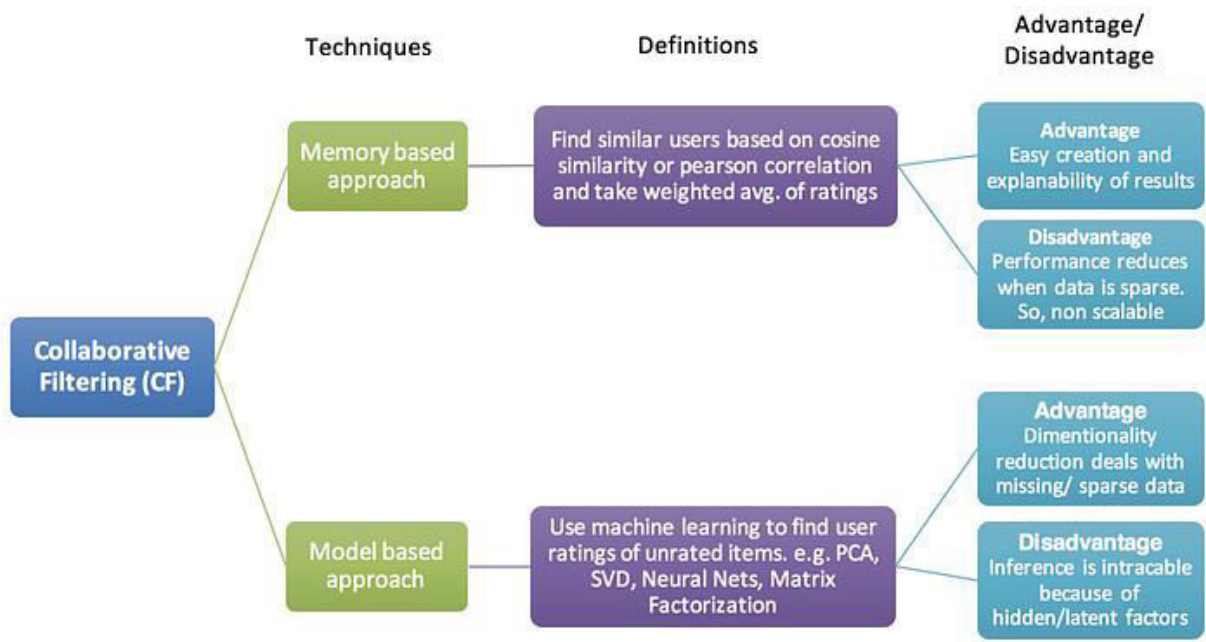
The obvious step in building a recommendation engine is finding the best-rated books that are a must-read for everyone. To do it, you have to get all book ratings and calculate the average rating score of every book in the dataset. After each book has a rating, the best books are easily extracted

3.11. Data from User

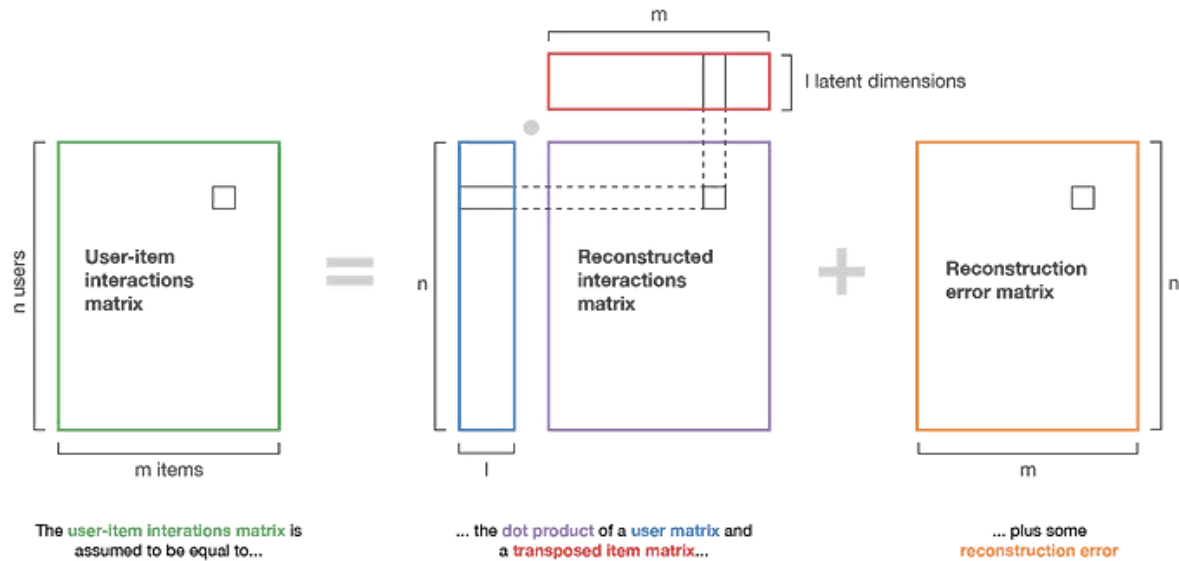
We have built a machine learning model for recommending books and now we will need to create a function using Python. When this function is called, we will have to pass the name of the book to it. The model will try to find books based on the features. We'll store those book names that the system recommends in a list and return them at the end.

3.12. Data Validation

We do not want to find a similarity between users or books. we want to do that If there is user A who has read and liked x and y books, And user B has also liked this two books and now user A has read and liked some z book which is not read by B so we have to recommend z book to user B. This is what collaborative filtering is.



So this is achieved using Matrix Factorization, we will create one matrix where columns will be users and indexes will be books and value will be rating. Like we have to create a Pivot table.



3.13. User Data Inserting into Database

1. Database Creation and connection - Create a database with name passed. If the database is already created, open the connection to the database.
2. Table creation in the database.
3. Insertion of files in the table

3.14. Data Clustering

Clustering Methods:

□ **Density-Based Methods:** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters.

□ Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure) etc.

□ **Hierarchical Based Methods:** The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two categories:

o **Agglomerative** (*bottom up approach*)

o **Divisive** (*top down approach*)

□ Examples: CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) etc.

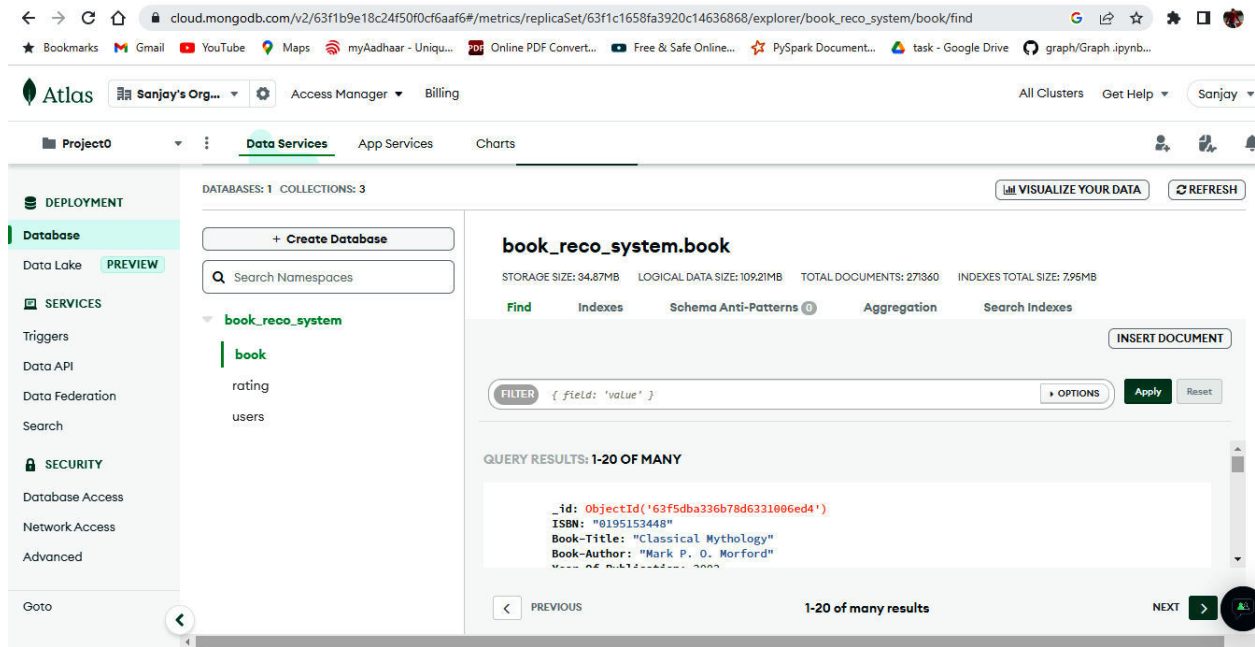
□ **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter

example K-means, CLARANS (Clustering Large Applications based upon Randomized Search) etc.

□ **Grid-based Methods:** In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (Clustering In Quest) etc.

In this paper, partitioning method of clustering is used. We used Clustering algorithm which is simplest unsupervised learning algorithm in this paper and it partition n observations into k clusters where each observation belongs to the cluster.

3.15. Model Call for Specific Cluster

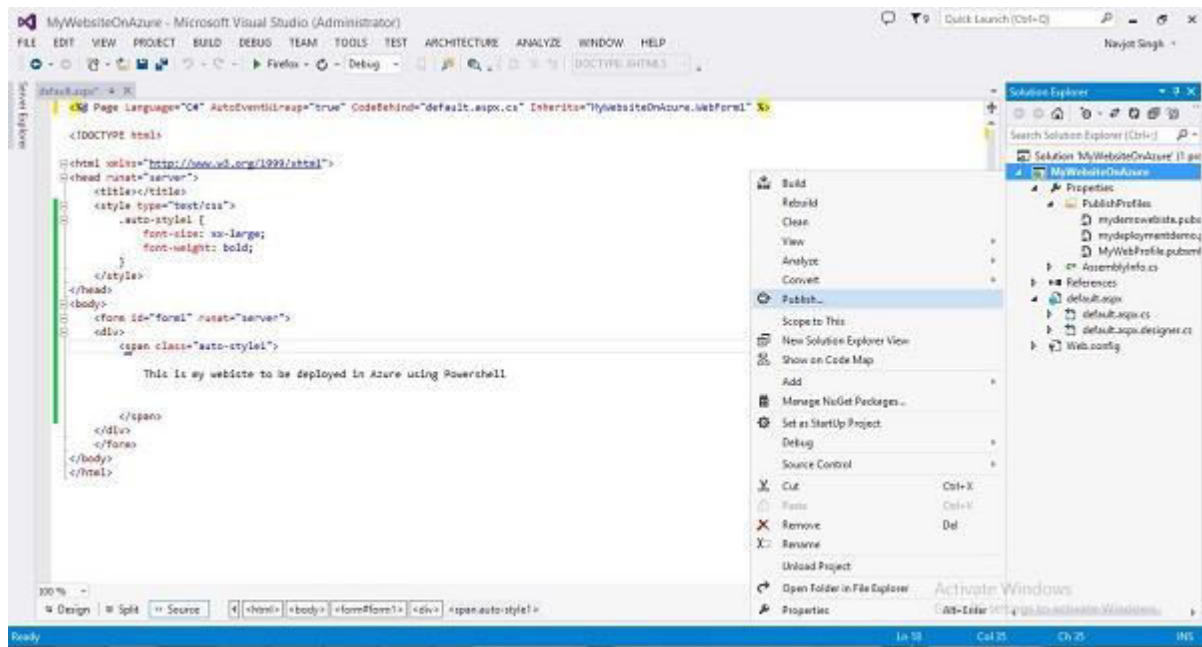


3.16. Deployment

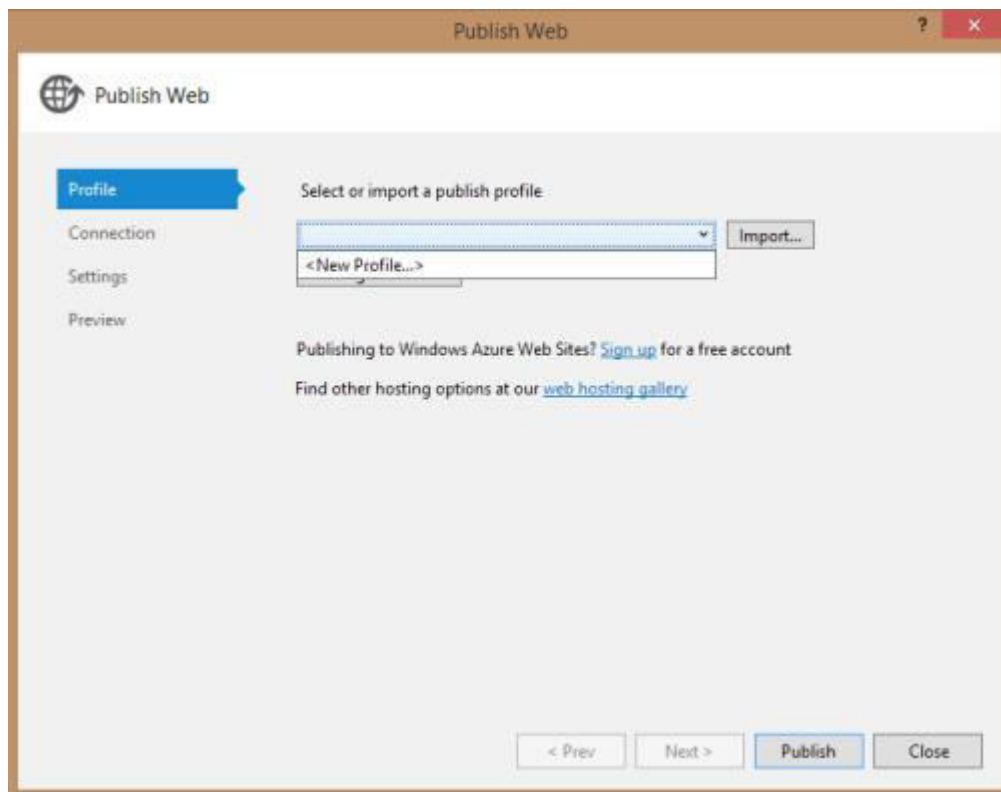
Create a Deployment Package

Step 1 – Go to your website in Visual Studio.

Step 2 – Right-click on the name of the application in the solution explorer. Select 'Publish'.



Step 3 – Create a new profile by selecting ‘New Profile’ from the dropdown. Enter the name of the profile. There might be different options in dropdown depending on if the websites are published before from the same computer.

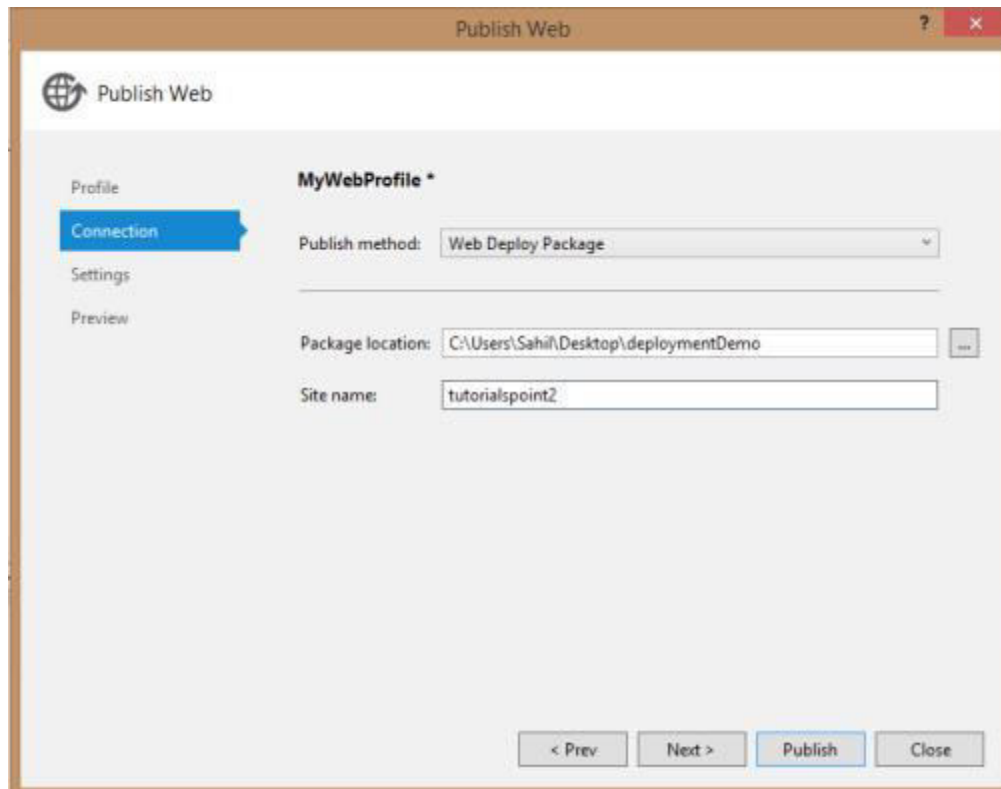


Step 4 – On the next screen, choose ‘Web Deploy Package’ in Publish Method.

The screenshot shows the 'Publish Web' dialog box with the following fields and options:

- Profile:** MyWebProfile *
- Connection:** Selected tab.
- Publish method:** Web Deploy (selected), Web Deploy Package (highlighted), FTP, File System.
- Server:** (empty field)
- Site name:** e.g. www.contoso.com or Default Web Site/MyApp (placeholder text)
- User name:** (empty field)
- Password:** (empty field)
- Save password:** ☐
- Destination URL:** e.g. http://www.contoso.com (placeholder text)
- Buttons:** < Prev, Next >, Publish (highlighted), Close.

Step 5 – Choose a path to store the deployment package. Enter the name of site and click Next.



Step 6 – On the next screen, leave the defaults on and select ‘publish’.

After it’s done, inside the folder in your chosen location, you will find a zip file which is what you need during deployment.

4. Unit Test Cases

Test Case Description	Pre-Requisite	Expected Result
Verify whether the Application URL is accessible to the user	Application URL should be defined	Application URL should be accessible to the user
Verify whether the Application loads completely for the user when the URL is accessed	1.Application URL is accessible 2.Application is deployed	The Application should load completely for the user when the URL is accessed
Verify whether user gets Submit button to submit the inputs	Application is accessible	User should get Submit button to submit the inputs
Verify whether user is presented with recommended results on clicking submit	1. Application is accessible 2. User is signed up to the application 3. User is logged in to the application	User should be presented with recommended results on clicking submit
Verify whether the recommended results are in accordance to the selections user made	1. Application is accessible	The recommended results should be in accordance to the selections user made