SUS College of Engineering
&
Technology

Department of Computer Science & Engineering

“ONLINE BOOK RECOMMENDER SYSTEM”
USING
“COLLABORATIVE FILTERING ALGORITHM”

A project report submitted in partial fulfilment of requirements
for the award of the degree

Bachelor of Technology
In
Computer Science Engineering

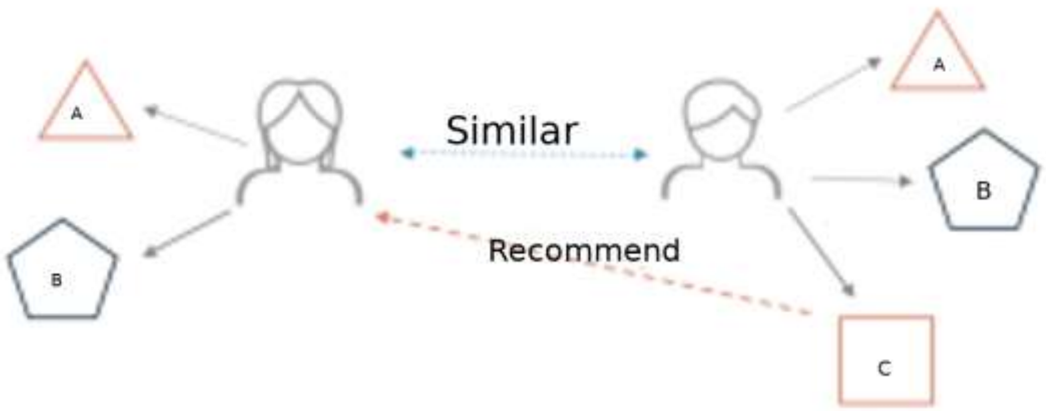
Submitted To: -	**Submitted By: -**
Er. Deepinder Kaur	**Vishal Kumar(1909088)**
	Branch:- B.Tech(CSE)

PROJECT REPORT

What is a recommender system?

A recommender system is a simple algorithm whose aim is to provide the most relevant information to a user by discovering patterns in a dataset. The algorithm rates the items and shows the user the items that they would rate highly. An example of recommendation in action is when you visit Amazon and you notice that some items are being recommended to you or when Netflix recommends certain book to you. They are also used by books streaming applications such as Spotify and Deezer to recommend study that you might like.

Below is a very simple illustration of how recommender systems work in the context of an e-commerce site.



Two users buy the same items A and B from an e-commerce store. When this happens the similarity index of these two users is computed. Depending on the score the system can recommend item C to the other user because it detects that those two users are similar in terms of the items they purchase.

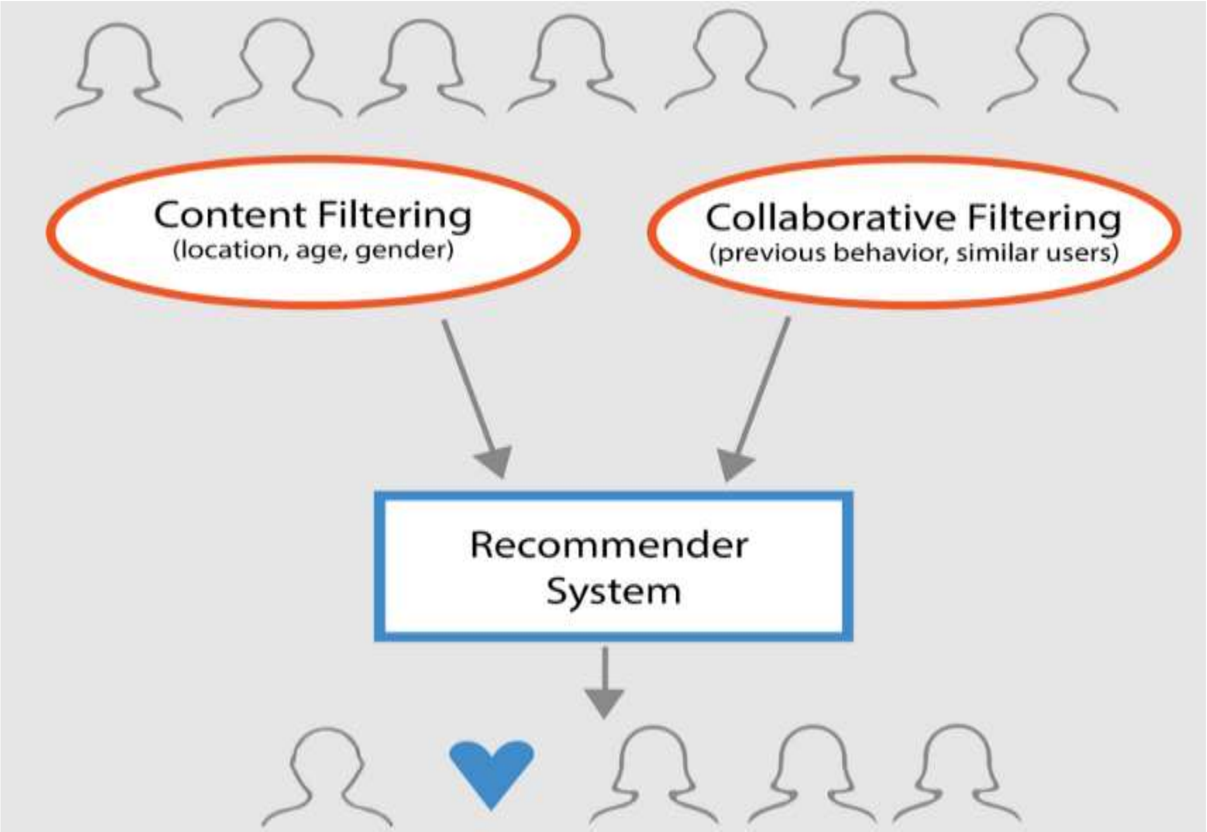
Different types of recommendation

The most common types of recommendation systems are **content-based** and **collaborative filtering** recommender systems. In collaborative filtering, the behavior of a group of users is used to make recommendations to other users. The recommendation is based on the preference of other users. A simple example would be recommending a movie to a user based on the fact that their friend liked the movie. There are two types of collaborative models **Memory-based** methods and **Model-based** methods. The advantage of memory-based techniques is that they are simple to implement and the resulting recommendations are often easy to explain. They are divided into two:

- **User-based collaborative filtering**: In this model, products are recommended to a user based on the fact that the products have been liked by users similar to the user. For example, if Derrick and Dennis like the same books and a new book come out that Derick like, then we can recommend that movie to Dennis because Derrick and Dennis seem to like the same books.
- **Item-based collaborative filtering:** These systems identify similar items based on users' previous ratings. For example, if users A, B, and C gave a 5-star rating to books X and Y then when a user D buys book Y they also get a recommendation to purchase book X because the system identifies book X and Y as similar based on the ratings of users A, B, and C.

Model-based methods are based on Matrix Factorization and are better at dealing with sparsity. They are developed using data mining, machine learning algorithms to predict users' rating of unrated items. In this approach techniques such as dimensionality reduction are used to improve accuracy. Examples of such model-based methods include Decision trees, Rule-based Model, Bayesian Model, and latent factor models.

Content-based systems use metadata such as author, writer to items say book s . Such a recommendation would be for instance recommending Infinity recommend War that featured Vin Diesel because someone studied and liked The Fate of the Furious. Similarly, you can get recommendations from certain book because you liked their book content . Content-based systems are based on the idea that if you liked a certain item you are most likely to like something that is similar to it.



- **Introduction** :-

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.

- **Introduction to Machine Learning :-**

  Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. It gives the computer that makes it more similar to humans i.e. ability to learn. Machine learning is used in many streams than anyone would accept. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.In order to enable the software to independently generate solutions, the prior action of people is necessary. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning:
- Finding, extracting and summarizing relevant data
- Making predictions based on the analysis data
- Calculating probabilities for specific results

Basically, algorithms play an important role in Machine Learning: On the one hand, they are responsible for recognizing patterns and on the other hand, they can generate solutions. Algorithms can be divided into different categories:

- **Supervised learning:**

In the course of monitored learning, example models are defined in advance. In order to ensure an adequate allocation of the information to the respective model groups of the algorithms, these then have to be specified. In other words, the system learns on the basis of given input and output pairs. In the course of monitored learning, a programmer, who acts as a kind of teacher, provides the appropriate values for a particular input. The aim is to train the system in the context of successive calculations with different inputs and outputs to establish connections. Supervised learning is where you have input variables (X) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. Y = f(X) The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Techniques of Supervised Machine Learning algorithms include linear and logistic regression, multi-class classification, Decision Tree and Support Vector Machine. Supervised Learning problems can be further grouped into Regression and Classification problems. The difference between these two is that the dependent attribute is numerical for regression and categorical for classification:

- **Regression:** Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.
- **Classification:** Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories. In short classification either predicts categorical class labels or classification data based on the training set and the values(class labels) in classifying attributes and uses it in classifying new data. There are number of

classification models. Classification models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree, One-vs.-One and Naïve Bayes.

- **Unsupervised learning:**

In unsupervised learning, artificial intelligence learns without predefined target values and without rewards. It is mainly used for learning segmentation (clustering). The machine tries to structure and sort the data entered according to certain characteristics. For example, a machine could (very simply) learn that coins of different colors can be sorted according to the characteristic "color" in order to structure them. Unsupervised Machine Learning algorithms are used when the information used to train is neither classified nor labeled. The system does not figure out the right output but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data. Unsupervised Learning is the training of Machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Unsupervised Learning is classified into two categories of algorithms:

- **Clustering:** A clustering problem is where you want to discover the inherent grouping in the data such as grouping customers by purchasing behavior.
- **Association:** An Association rule learning problem is where you want to discover rules that describe large portions of your data such as people that buy X also tend to buy Y.

Applications of Machine Learning:

Virtual Personal Assistants: Siri, Alexa, Google Now are some of the popular examples of virtual personal assistants. As the name suggests, they assist in finding information, when asked over voice. Machine learning is an important part of these personal assistants as they collect and refine the information on the basis of your previous involvement with them. Later, this set of data is utilized to render results that are tailored to your preferences.

Virtual Assistants are integrated to a variety of platforms. For example:

- Smart Speakers: Amazon Echo and Google Home

- Smartphone: Samsung Bixby on Samsung S8

- Mobile Apps: Google Allo

Videos Surveillance:

• Imagine a single person monitoring multiple video cameras! Certainly, a difficult job to do and boring as well. This is why the idea of training computers to do this job makessense.

• The video surveillance system, nowadays are powered by AI that makes it possible to detect crime before they happen. They track unusual behaviour of people like standing motionless for a long time, stumbling, or napping on benches etc. The system can thus give an alert to human attendants, which can ultimately help to avoid mishaps. And when such activities are reported and counted to be true, they help to improve the surveillance services. This happens with machine learning doing its job at the backend.

Social Media Services:

From personalizing your news feed to better ads targeting, social media platforms are utilizing machine learning for their own and user benefits.

- People You May Know
- Face Recognition

Search Engine Result Refining:

Google and other search engines use machine learning to improve the search results for you. Every time you execute a search, the algorithms at the backend keep a watch at how you respond to the results. If you open the top results and stay on the web page for long, the search engine assumes that the results it displayed were in accordance to the query. Similarly, if you reach the second or third page of the search results but do not open any of the results, the search engine estimates that the results served did not match requirement. This way, the algorithms working at the backend improve the search results.

- **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.

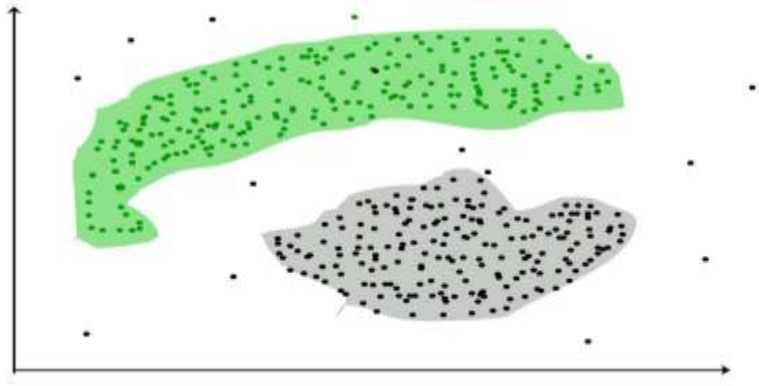


Fig :- Clustering of data points

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, CLIQ(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.

Matrix Factorization

Matrix factorization is a way to generate latent features when multiplying two different kinds of entities. Collaborative filtering is the application of matrix factorization to identify the relationship between items and users entities. With the input of users' ratings on the shop items, we would like to predict how the users would rate the items so the users can get the recommendation based on the prediction. Matrix Factorization is a technique to discover the latent factors from the ratings matrix and to map the items and the users against those factors. Consider a ratings matrix R with ratings by n users for m items. The ratings matrix R will have n × m rows and columns.

Matrix Factorization is a significant approach in many applications. Curse of dimensionality is a phenomenon which occurs in high dimensional space that hardly occur in lower dimensional space. Due to higher number of dimension model gets sparse. Higher dimensional space causes problem in clustering (becomes very difficult to separate one cluster data from another), search space increases, complexity of model increases. We can reduce the dimension by following two ways:

- **Feature selection:** Selecting important features which are relevant to model (it avoids the curse of dimensionality)
- **Feature extraction:** Transformation of high dimensional space into lower dimensional space by using various methods such as PCA, TSVD, T- SNE etc.

In this paper, we used Feature extraction for reducing the features and used method i.e. Truncated SVD for dimensionality reduction.

Silhouette Score

Silhouette score or silhouette coefficient is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. It refers to a method of interpretation and validation of consistency within clusters of data. Silhouette Score is a metric used to calculate the goodness of a clustering technique. The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

Silhouette Score = (b-a)/max (a, b)

Where,

- a = average intra-cluster distance i.e. the average distance between each point within a cluster.
- b = average inter-cluster distance i.e. the average distance between all clusters.

This value ranges from -1 to 1. Positive value indicates that mean clusters are well apart from each other and clearly distinguished so we require a << b. 0 indicates that mean clusters are indifferent, or we can say that the distance between clusters is not significant. Negative value indicates that mean clusters are assigned in the wrong way. We can also increase the likelihood of the silhouette being maximized at the correct number of clusters by re-scaling the data using feature weights that are cluster specific.

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.

1. Data Acquisition
2. Data Pre-processing
3. Feature Extraction
4. Training Methods
5. Testing Data

In Step-1, Dataset was collected from Good Reads Website in which three datasets are present i.e. Books Dataset, Ratings Dataset, Users Dataset. In Step-2, Datasets were pre-processed to make suitable for developing the Recommendation system. In Step-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-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-5, Testing of model with test data is performed.

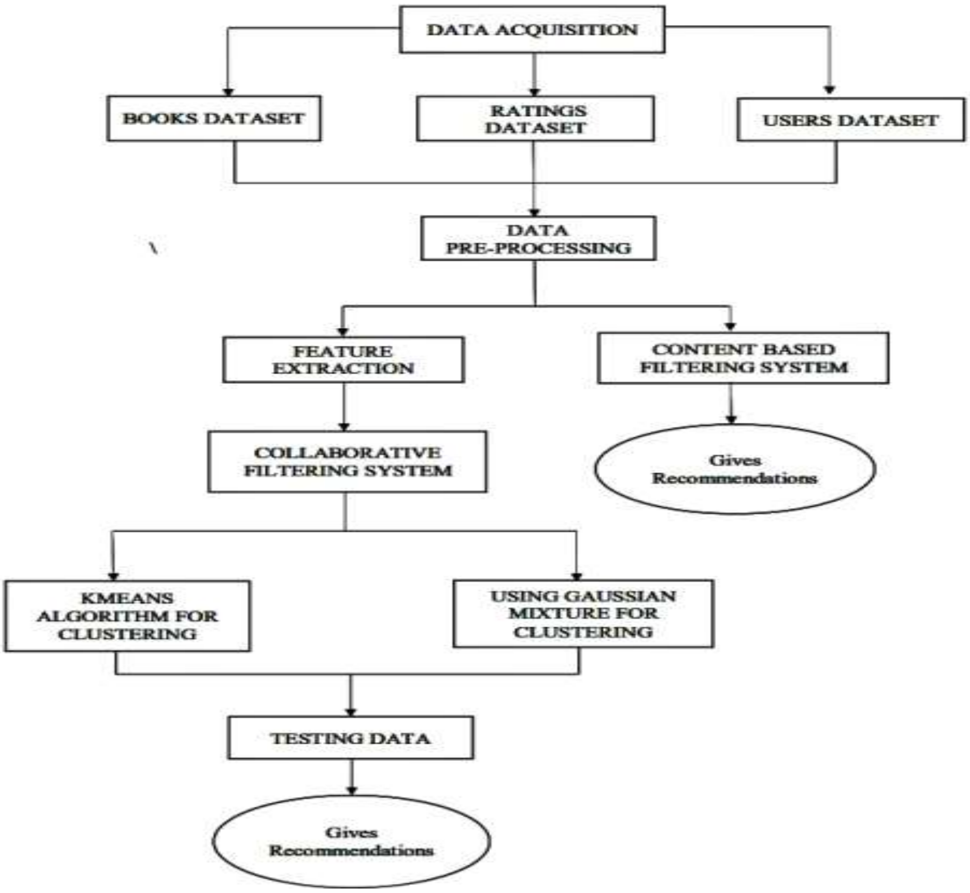


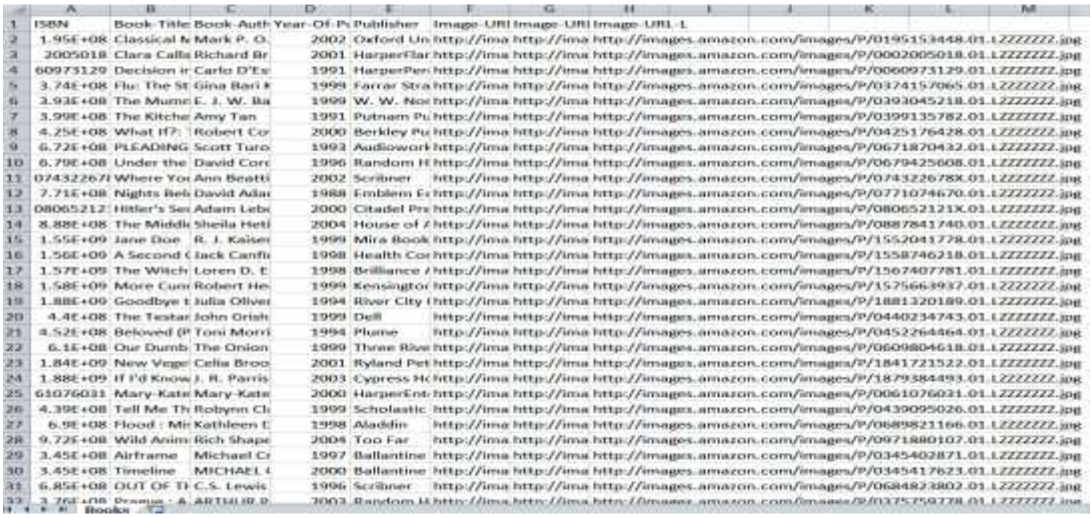
Fig- System Architecture

Modules Division

Let us discuss about the various modules in our proposed system and what each module contributes in achieving our goal.

- **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 an d the quality of the data in the collected dataset will determine the efficiency of the model. The Books dataset is collected from the **kaggle** website.



ISBN	Book-Title	Book-Auth	Year-Of-P	Publisher	Image-URI	Image-URI	Image-URL-L
1.95E+08	Classical M	Mark P. O.	2002	Oxford Un	http://ima	http://ima	http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg
2005018	Clara Calla	Richard Br	2001	HarperFlar	http://ima	http://ima	http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg
60973129	Decision ir	Carlo D'Es	1991	HarperPer	http://ima	http://ima	http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg
3.74E+08	Flu: The St	Gina Bari K	1999	Farrar Stra	http://ima	http://ima	http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg
3.93E+08	The Mumm	E. J. W. Ba	1999	W. W. Nor	http://ima	http://ima	http://images.amazon.com/images/P/0393045218.01.LZZZZZZZ.jpg
3.99E+08	The Kitche	Amy Tan	1991	Putnam Pu	http://ima	http://ima	http://images.amazon.com/images/P/0399135782.01.LZZZZZZZ.jpg
4.25E+08	What If?:	Robert Co	2000	Berkley Pu	http://ima	http://ima	http://images.amazon.com/images/P/0425176428.01.LZZZZZZZ.jpg
6.72E+08	PLEADING	Scott Turo	1993	Audiowork	http://ima	http://ima	http://images.amazon.com/images/P/0671870432.01.LZZZZZZZ.jpg
6.79E+08	Under the	David Corc	1996	Random H	http://ima	http://ima	http://images.amazon.com/images/P/0679425608.01.LZZZZZZZ.jpg
07432267I	Where You	Ann Beatti	2002	Scribner	http://ima	http://ima	http://images.amazon.com/images/P/074322678X.01.LZZZZZZZ.jpg
7.71E+08	Nights Bel	David Adai	1988	Emblem Ec	http://ima	http://ima	http://images.amazon.com/images/P/0771074670.01.LZZZZZZZ.jpg
08065212	Hitler's Sec	Adam Lebo	2000	Citadel Pre	http://ima	http://ima	http://images.amazon.com/images/P/080652121X.01.LZZZZZZZ.jpg
8.88E+08	The Middle	Sheila Heti	2004	House of A	http://ima	http://ima	http://images.amazon.com/images/P/0887841740.01.LZZZZZZZ.jpg
1.55E+09	Jane Doe	R. J. Kaiser	1999	Mira Book	http://ima	http://ima	http://images.amazon.com/images/P/1552041778.01.LZZZZZZZ.jpg
1.56E+09	A Second C	Jack Canfi	1998	Health Cor	http://ima	http://ima	http://images.amazon.com/images/P/1558746218.01.LZZZZZZZ.jpg
1.57E+09	The Witch	Loren D. E	1998	Brilliance A	http://ima	http://ima	http://images.amazon.com/images/P/1567407781.01.LZZZZZZZ.jpg
1.58E+09	More Cunn	Robert He	1999	Kensingtor	http://ima	http://ima	http://images.amazon.com/images/P/1575663937.01.LZZZZZZZ.jpg
1.88E+09	Goodbye t	Julia Oliver	1994	River City I	http://ima	http://ima	http://images.amazon.com/images/P/1881320189.01.LZZZZZZZ.jpg
4.4E+08	The Testar	John Grish	1999	Dell	http://ima	http://ima	http://images.amazon.com/images/P/0440234743.01.LZZZZZZZ.jpg
4.52E+08	Beloved (P	Toni Morri	1994	Plume	http://ima	http://ima	http://images.amazon.com/images/P/0452264464.01.LZZZZZZZ.jpg
6.1E+08	Our Dumb	The Onion	1999	Three Rive	http://ima	http://ima	http://images.amazon.com/images/P/0609804618.01.LZZZZZZZ.jpg
1.84E+09	New Vege	Celia Broo	2001	Ryland Pet	http://ima	http://ima	http://images.amazon.com/images/P/1841721522.01.LZZZZZZZ.jpg
1.88E+09	If I'd Know	J. R. Parris	2003	Cypress Hc	http://ima	http://ima	http://images.amazon.com/images/P/1879384493.01.LZZZZZZZ.jpg
61076031	Mary-Kate	Mary-Kate	2000	HarperEnt	http://ima	http://ima	http://images.amazon.com/images/P/0061076031.01.LZZZZZZZ.jpg
4.39E+08	Tell Me Th	Robynn Cl	1999	Scholastic	http://ima	http://ima	http://images.amazon.com/images/P/0439095026.01.LZZZZZZZ.jpg
6.9E+08	Flood : Mi	Kathleen D	1998	Aladdin	http://ima	http://ima	http://images.amazon.com/images/P/0689821166.01.LZZZZZZZ.jpg
9.72E+08	Wild Anim	Rich Shape	2004	Too Far	http://ima	http://ima	http://images.amazon.com/images/P/0971880107.01.LZZZZZZZ.jpg
3.45E+08	Airframe	Michael Cr	1997	Ballantine	http://ima	http://ima	http://images.amazon.com/images/P/0345402871.01.LZZZZZZZ.jpg
3.45E+08	Timeline	MICHAEL C	2000	Ballantine	http://ima	http://ima	http://images.amazon.com/images/P/0345417623.01.LZZZZZZZ.jpg
6.85E+08	OUT OF TH	C.S. Lewis	1996	Scribner	http://ima	http://ima	http://images.amazon.com/images/P/0684823802.01.LZZZZZZZ.jpg
3.76E+08	Dramas : A	ARTHUR P	2003	Random H	http://ima	http://ima	http://images.amazon.com/images/P/0375759778.01.LZZZZZZZ.jpg

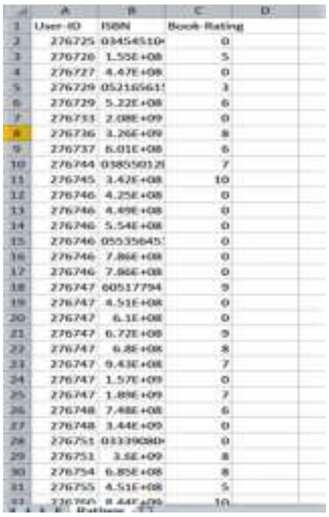
Books

Fig- Sample of acquired books dataset from kaggle Website

In the above Fig, we can see a sample of the dataset we have collected. This acquired dataset has around 2,00000 books and has 500 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 website.



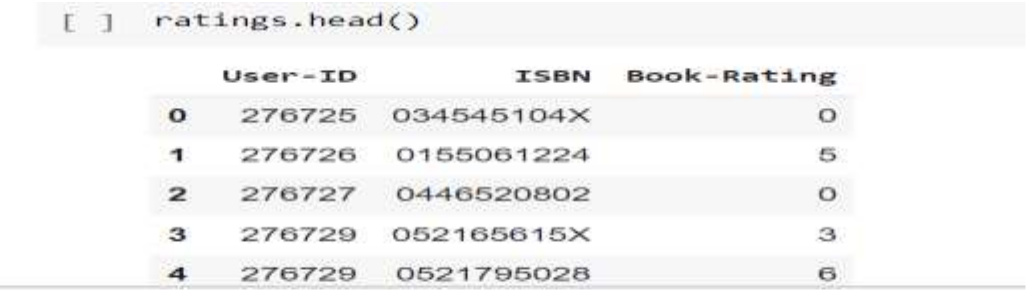
User-ID	ISBN	Book-Rating
276725	03454510	0
276726	1.55E+08	5
276727	4.47E+08	0
276729	05216561	3
276729	5.22E+08	6
276733	2.08E+09	0
276736	3.26E+09	8
276737	6.01E+08	6
276744	03855012	7
276745	3.42E+08	10
276746	4.25E+08	0
276746	4.49E+08	0
276746	5.54E+08	0
276746	05535645	0
276746	7.86E+08	0
276746	7.86E+08	0
276747	60517794	9
276747	4.51E+08	0
276747	6.1E+08	0
276747	6.72E+08	9
276747	6.8E+08	8
276747	9.43E+08	7
276747	1.57E+09	0
276747	1.89E+09	7
276748	7.48E+08	6
276748	3.44E+09	0
276751	03339080	0
276751	3.6E+09	8
276754	6.85E+08	8
276755	4.51E+08	5
276760	8.44E+09	10

Ratings

Fig-Sample of acquired ratings dataset from Good reads Website

In the above Fig, we can see a sample of the dataset we have collected. This acquired dataset has around 1000000 ratings and has 3 different features.

- User-ID
- ISBN
- BOOK-Rating



```
[ ]  ratings.head()
```

	User-ID	ISBN	Book-Rating
0	276725	034545104X	0
1	276726	0155061224	5
2	276727	0446520802	0
3	276729	052165615X	3
4	276729	0521795028	6

Fig-Reading the dataset from CSV file into python notebook

After acquiring the data our next step is to read the data from the csv file into python notebook. we have read data from csv file using the inbuilt python functions that are part of pandas library.

- ## Sample code:-

```
import numpy as np
import pandas as pd

books = pd.read_csv('books.csv')
users = pd.read_csv('users.csv')
ratings = pd.read_csv('ratings.csv')

books['Image-URL-M'][1]

users.head()
ratings.head()

print(books.shape)
print(ratings.shape)
print(users.shape)

books.isnull().sum()

users.isnull().sum()

ratings.isnull().sum()

books.duplicated().sum()
```

```
ratings.duplicated().sum()

users.duplicated().sum()
```

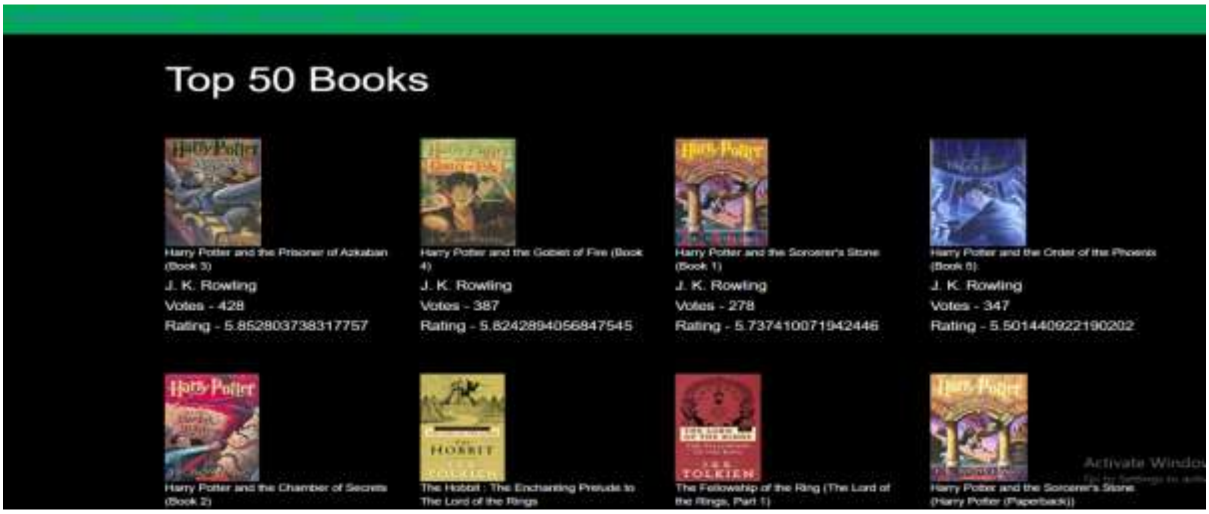
Popularity Based Recomender System

```
ratings_with_name = ratings.merge(books,on='ISBN')

num_rating_df=ratings_with_name.groupby('BookTitle').count()['BookRating'].reset_index()
num_rating_df.rename(columns={'Book-Rating':'num_ratings'},inplace=True)
num_rating_df

avg_rating_df=ratings_with_name.groupby('Book-Title').mean()['Book-Rating'].reset_index()
avg_rating_df.rename(columns={'Book-Rating':'avg_rating'},inplace=True)
avg_rating_df

popular_df = num_rating_df.merge(avg_rating_df,on='Book-Title')
popular_df

popular_df=popular_df[popular_df['num_ratings']>=250].sort_values('avg_rating',ascending=False).head(50)

popular_df = popular_df.merge(books,on='Book-Title').drop_duplicates('Book-Title')[['Book-Title','Book-Author','Image-URL-M','num_ratings','avg_rating']]

popular_df['Image-URL-M'][0]
```

Collaborative Filtering Based recommender System

```
x = ratings_with_name.groupby('User-ID').count()['Book-Rating'] > 200
padhe_likhe_users = x[x].index

filtered_rating = ratings_with_name[ratings_with_name['User-ID'].isin(padhe_likhe_users)]

y = filtered_rating.groupby('Book-Title').count()['Book-Rating']>=50
famous_books = y[y].index

final_ratings = filtered_rating[filtered_rating['Book-Title'].isin(famous_books)]

pt = final_ratings.pivot_table(index='Book-Title',columns='User-ID',values='Book-Rating')
```

```
pt.fillna(0,inplace=True)

pt

from sklearn.metrics.pairwise import cosine_similarity

similarity_scores = cosine_similarity(pt)

similarity_scores.shape

def recommend(book_name):
    # index fetch
    index = np.where(pt.index==book_name)[0][0]
    similar_items = sorted(list(enumerate(similarity_scores[index])),key=lambda
x:x[1],reverse=True)[1:5]

    data = []
    for i in similar_items:
        item = []
        temp_df = books[books['Book-Title'] == pt.index[i[0]]]
        item.extend(list(temp_df.drop_duplicates('Book-Title')['Book-Title'].values))
        item.extend(list(temp_df.drop_duplicates('Book-Title')['Book-Author'].values))
        item.extend(list(temp_df.drop_duplicates('Book-Title')['Image-URL-M'].values))

        data.append(item)

    return data

recommend('1984')

pt.index[545]

import pickle
pickle.dump(popular_df,open('popular.pkl','wb'))

books.drop_duplicates('Book-Title')

pickle.dump(pt,open('pt.pkl','wb'))
pickle.dump(books,open('books.pkl','wb'))
pickle.dump(similarity_scores,open('similarity_scores.pkl','wb'))
```



Top 50 Books

Harry Potter and the Prisoner of Azkaban (Book 3)
J. K. Rowling
Votes - 428
Rating - 5.852803738317757

Harry Potter and the Goblet of Fire (Book 4)
J. K. Rowling
Votes - 387
Rating - 5.8242894056847545

Harry Potter and the Sorcerer's Stone (Book 1)
J. K. Rowling
Votes - 278
Rating - 5.737410071942446

Harry Potter and the Order of the Phoenix (Book 5)
J. K. Rowling
Votes - 347
Rating - 5.501440922190202

Harry Potter and the Chamber of Secrets (Book 2)

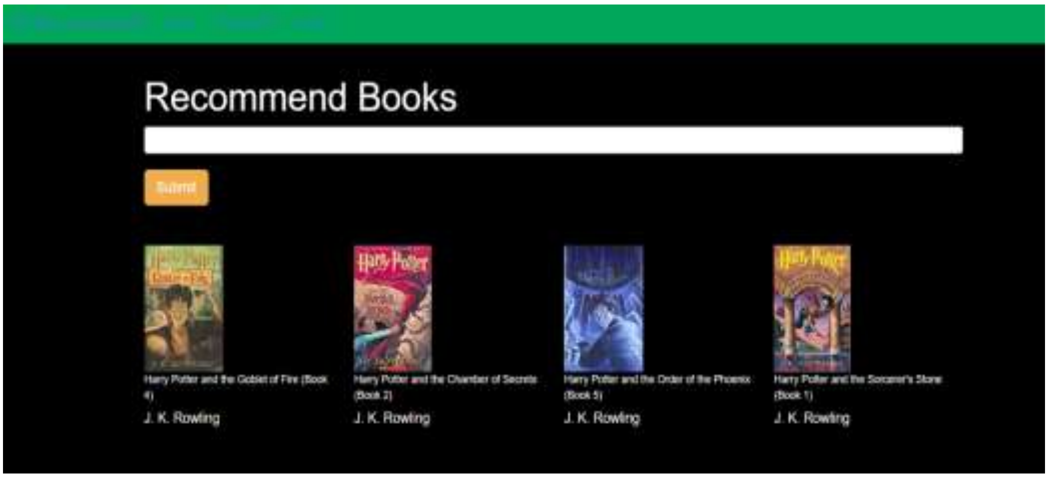
The Hobbit : The Enchanting Prelude to The Lord of the Rings

The Fellowship of the Ring (The Lord of the Rings, Part 1)

Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))

Fig- Recommended books in Home Page

Under the Recommended books, there are top 50 books which are recommended by the Content Based Filtering system are also shown in the Home page, which are named as popular books. These books were shown to the new users also irrespective of the user data. We can see the books in above Fig.



Recommend Books

Harry Potter and the Goblet of Fire (Book 4)
J. K. Rowling

Harry Potter and the Chamber of Secrets (Book 2)
J. K. Rowling

Harry Potter and the Order of the Phoenix (Book 5)
J. K. Rowling

Harry Potter and the Sorcerer's Stone (Book 1)
J. K. Rowling

Fig-Searching book with book name

We want clicking on the recommend button.After clicking on the button, we are redirected to another page where we can give the book-name in the surch bar and clicking on the submit button. after clicking on the Show rated books.

Walkthrough of building a recommender system

We are going to use the movie lens to build a simple item similarity-based recommender system. The first thing we need to do is to import pandas and numpy.

```
import pandas as pd
import numpy as np
```

Next, we load in the data set using pandas `read_csv()` utility. The dataset is tab separated so we pass in `\t` to the `sep` parameter. We then pass in the column names using the `names` parameter.

```
books = pd.read_csv(r'C:\Users\vishalkumar\Desktop\Book_recommendetion_system\Books.csv.zip')
```

Now let's check the head of the data to see the data we are dealing with.

```
books['Image-URL-M'][1]
```

It would be nice if we can see the titles of the book instead of just dealing with the IDs. Let's load in the book titles and merge it with this dataset.

```
rating= pd.read_csv(' rating ')
rating.head()
```

Since the `item_id` columns are the same we can merge these datasets on this column.

```
ratings_with_name = ratings.merge(books,on='ISBN')

popular_df = num_rating_df.merge(avg_rating_df,on='Book-Title')

popular_df

filtered_rating = ratings_with_name[ratings_with_name['User-ID'].isin(padhe_likhe_users)]

popular_df = popular_df.merge(books,on='Book-Title').drop_duplicates('Book-Title')[['Book-Title','Book-
Author','Image-URL-M','num_ratings','avg_rating
```

How to improve the recommendation system

This system can be improved by building a Memory-Based Collaborative Filtering based system. In this case, we'd divide the data into a training set and a test set. We'd then use techniques such as cosine similarity to compute the similarity between the book s. An alternative is to build a Model-based Collaborative Filtering system. This is based on matrix factorization. Matrix factorization is good at dealing with scalability and sparsity than the former. You can then evaluate your model using techniques such as Root Mean Squared Error (RMSE).

Tools and Libraries used

- Python –
- Pandas
- Numpy
- Streamlit
- ML
- AI

•Python – 3.x

•Pandas – 1.2.4

•Matplotlib – 3.3.4

•TensorFlow – 2.4.1

•NLP

•ML

•AI