Information Retrieval
Baseline Results

Anupam Garg
anupam20555@iiitd.ac.in

Saharsh Dev
saharsh20572@iiitd.ac.in

Saksham Singhi
saksham20463@iiitd.ac.in

Vishal Kumar
vishal20154@iiitd.ac.in

Vanshika Goel
vanshika20413@iiitd.ac.in

Vanisha Singh
vanisha20347@iiitd.ac.in

Updated Problem Formulation and Literature Review:

Problem Statement:

With millions of books available, it may be overwhelming for people to select and choose what to read next, and sometimes, the contents of the books don't turn out what they were expected to be. A book recommendation system will personalize, increase efficiency and engagement and diversify the range of books.

Literature review:

[1] In 2021, Sarma, Mittra, Hossain proposed a book recommendation system using machine-learning algorithms and created a clustering-based system by applying the k-means algorithm and cosine similarity.

[2] Kurmashov et al. in their paper[1] tried to propose a book recommendation service by taking inputs of the preference like the most favorite genre and ratings on different books to further narrow down the space of search for recommendations while a user logs into the system.

[3] In 2022, Mishra, Asthana tried to outline the limitations of the content and the collaborative-based filtering and proposed an effective solution in this regard. While content filtering works on the traditional methods, hybrid filtering tries to combine them and employs a collaborative networking approach, which compares the results with a wider audience and produces accurate results.

Problem Formulation:

1. After studying various research papers, we formulated the idea of developing a book recommendation system using Collaborative Filtering by the Model-based approach. Collaborative filtering (CF) systems collect user feedback in the form of ratings for items in a given domain and use similarities in rating behavior across multiple users to determine how to recommend an item. It is built on the notion that people who have previously agreed in their assessments of certain items are more likely to agree in the future.

2. We will be using the Model-based approach, that is, using machine learning algorithms as we have reviewed in (Sarma et al., 2021), and modify the approach as applied in this study from K-mean clustering to k-Nearest Neighbors (k-NN), which is a non-parametric, supervised learning classifier that uses proximity to make predictions about the grouping of an individual data point.

3. k-NN does not make assumptions and relies on the item feature similarity. If it needs to make an inference then it will calculate the “distance” between the target book and every other book in its database, and returns the top k-nearest books as recommendations. The reason behind the update is that k-NN clustering may be a better choice than K-means clustering in scenarios where the data is non-linearly separable, the data is noisy, or the number of clusters is unknown.

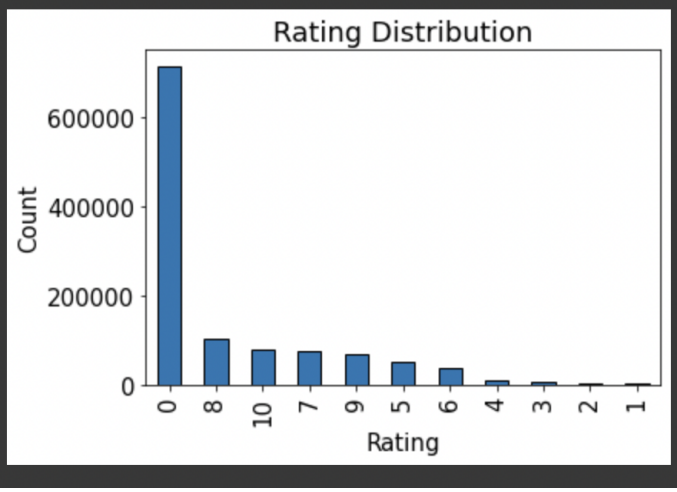
4. To further enhance the working of our recommendation system, we will incorporate the Content-based filtering method as we have studied in (Benkoussas et al., 2015), and (Mishra et al., 2022) to develop a hybrid model which combines the features and overcomes the shortcomings of content and collaborative filtering when applied individually.

Baseline Results:

1. Firstly we read all the datasets of books, users, and ratings.

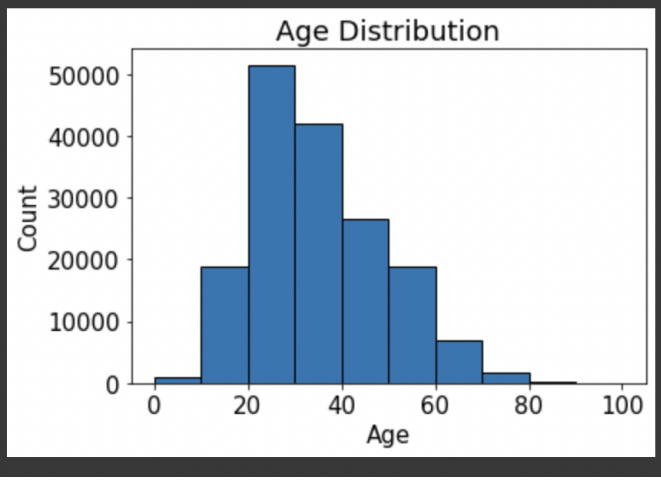
2. In order to study the given dataset, prior processing is required. We first plotted a graph to study how many ratings are given to how many number of books.

```
[ ] book_rating.bookRating.value_counts(sort = True).plot(kind = 'bar', edgecolor = "black")
    plt.title('Rating Distribution'), plt.xlabel('Rating'), plt.ylabel('Count')
    plt.show()
```



3. Secondly, we plotted an age distribution graph, in order to study the age distribution of the users in our database.

```
[ ] users.Age.hist(bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100], edgecolor = "black")
    plt.title('Age Distribution'), plt.xlabel('Age'), plt.ylabel('Count'), plt.grid(False)
    plt.show()
```



4. Now we started to process the available data. We filter out the users that have rated more than 200 books in our dataset.

```
[ ]  relevant_users = book_rating['User-ID'].value_counts() >= 200
     indexes_relevant_users = relevant_users[relevant_users].index
     indexes_relevant_users
```

```
Int64Index([ 11676, 198711, 153662,  98391,  35859, 212898, 278418,  76352,
            110973, 235105,
            ...
             28634,  59727, 268622, 188951, 225595,  83671, 252827,  99955,
             36554,  26883],
           dtype='int64', length=905)
```

5. Now based on the users that we sorted out above we reduced our dataset and printed the rating of books given by the relevant users (>=200).

```
[ ]  relevant_ratings = book_rating[book_rating['User-ID'].isin(indexes_relevant_users)]
     ratings_with_books = relevant_ratings.merge(books, on = 'ISBN')
```

```
ratings_with_books
```

	User-ID	ISBN	bookRating	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	277427	002542730X	10	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc
1	3363	002542730X	0	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc
2	11676	002542730X	6	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc
3	12538	002542730X	10	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc
4	13552	002542730X	0	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc
...	...	...	...	...	...	...	...
488751	275970	1892145022	0	Here Is New York	E. B. White	1999	Little Bookroom
488752	275970	1931868123	0	There's a Porcupine in My Outhouse: Misadventu...	Mike Tougias	2002	Capital Books (VA)
488753	275970	3411086211	10	Die Biene.	Sybil GrÃ?Â¤fin SchÃ?Â¶nfeldt	1993	Bibliographisches Institut, Mannheim
488754	275970	3829021860	0	The Penis Book	Joseph Cohen	1999	Konemann
488755	275970	4770019572	0	Musashi	Eiji Yoshikawa	1995	Kodansha International (JPN)

488756 rows × 7 columns

6. Now we displayed all the distinct books that we have in our database and displayed the count of the number of times the book has been rated by different users.

	Book-Title	Number of Ratings
0	A Light in the Storm: The Civil War Diary of ...	2
1	Always Have Popsicles	1
2	Apple Magic (The Collector's series)	1
3	Beyond IBM: Leadership Marketing and Finance ...	1
4	Clifford Visita El Hospital (Clifford El Gran...	1
...	...	...
160582	Ã?Â?ber die Pflicht zum Ungehorsam gegen den S...	3
160583	Ã?Â?lpiraten.	1
160584	Ã?Â?rger mit Produkt X. Roman.	1
160585	Ã?Â?stlich der Berge.	1
160586	Ã?Â?thique en toc	1

160587 rows × 2 columns

7. Final ratings.

```
final_ratings
```

	User-ID	ISBN	bookRating	Book-Title	Book-Author	Year-Of-Publication	Publisher	Number of Ratings
0	277427	002542730X	10	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc	82
1	3363	002542730X	0	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc	82
2	11676	002542730X	6	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc	82
3	12538	002542730X	10	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc	82
4	13552	002542730X	0	Politically Correct Bedtime Stories: Modern Ta...	James Finn Garner	1994	John Wiley & Sons Inc	82
...	...	...	...	...	...	...	...	...
488751	275970	1892145022	0	Here Is New York	E. B. White	1999	Little Bookroom	1
488752	275970	1931868123	0	There's a Porcupine in My Outhouse: Misadventu...	Mike Tougias	2002	Capital Books (VA)	1
488753	275970	3411086211	10	Die Biene.	Sybil GrÃ?Â¤fin SchÃ?Â¶nfeldt	1993	Bibliographisches Institut, Mannheim	1
488754	275970	3829021860	0	The Penis Book	Joseph Cohen	1999	Konemann	1
488755	275970	4770019572	0	Musashi	Eiji Yoshikawa	1995	Kodansha International (JPN)	1

488756 rows × 8 columns

8. Now we sorted out the books that have been rated more than 50 times.

```
[ ] final_ratings = final_ratings[final_ratings['Number of Ratings'] >= 50]
    final_ratings.drop_duplicates(['User-ID','Book-Title'], inplace = True)
```

```
/usr/local/lib/python3.8/dist-packages/pandas/util/_decorators.py:311: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  return func(*args, **kwargs)
```

9. Now in order to apply collaborative filtering(CF), we constructed a pivot table.

```
[ ] pivoted_books = final_ratings.pivot_table(columns = 'User-ID', index = 'Book-Title', values = 'bookRating')
    pivoted_books.fillna(0, inplace = True)
```

```
pivoted_books
```

User-ID	254	2276	2766	2977	3363	3757	4017	4385	6242	6251	...	274004	274061	274301	274308	274808	275970	277427	277478	277639	278418
Book-Title																					
1984	9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1st to Die: A Novel	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2nd Chance	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4 Blondes	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
84 Charing Cross Road	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Year of Wonders	0.0	0.0	0.0	7.0	0.0	0.0	0.0	0.0	7.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
You Belong To Me	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Zen and the Art of Motorcycle Maintenance: An Inquiry into Values	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Zoya	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
\O\" Is for Outlaw"	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

746 rows × 894 columns

10. Now we implemented the k-NN clustering algorithm on the given dataset.

```
[ ]  book_sparse = csr_matrix(pivoted_books)
     model = NearestNeighbors(algorithm = 'brute')
     model.fit(book_sparse)

     NearestNeighbors(algorithm='brute')
```

11. Now we implemented a function that returns the k nearest neighbors of a particular book.

```
# Function that returns the book names

def recommend_book(book_name):

    book_id = np.where(pivoted_books.index == book_name)[0][0]

    book_list = dist, sugg = model.kneighbors(pivoted_books.iloc[book_id, :].values.reshape(1, -1), n_neighbors = 6)

    # Looping over suggestions

    for i in range(len(book_list)):

      if not i:

        print(pivoted_books.index[sugg[i]])
```

Examples:-

```
[ ]  recommend_book('Exclusive')

     Index(['Exclusive', 'The Cradle Will Fall', 'The Long Road Home',
            'Jacob Have I Loved', 'No Safe Place', 'Eyes of a Child'],
           dtype='object', name='Book-Title')

[ ]  recommend_book('And Then You Die')

     Index(['And Then You Die', 'Long After Midnight', 'No Safe Place', 'Exclusive',
            'The Most Wanted', 'Executive Orders (Jack Ryan Novels)'],
           dtype='object', name='Book-Title')

[ ]  recommend_book('1984')

     Index(['1984', 'No Safe Place', 'A Civil Action', 'Foucault's Pendulum',
            'Long After Midnight', 'Abduction'],
           dtype='object', name='Book-Title')

     recommend_book('You Belong To Me')

     Index(['You Belong To Me', 'Exclusive', 'The Cradle Will Fall',
            'Loves Music, Loves to Dance', 'While My Pretty One Sleeps',
            'Before I Say Good-Bye'],
           dtype='object', name='Book-Title')

[ ]  recommend_book('While My Pretty One Sleeps')

     Index(['While My Pretty One Sleeps', 'The Cradle Will Fall', 'Exclusive',
            'Long After Midnight', 'No Safe Place', 'Dragon Tears'],
           dtype='object', name='Book-Title')
```