National University of Sciences and Technology (NUST)
School of Electrical Engineering and Computer Science

CS-471 Machine Learning Assignment # 02

Assignment Title: Book Recommendation System (Content-Based + Collaborative Filtering)

Name	Irfa Farooq
CMS ID	412564
Date	December 19, 2025
Submitted to	Ma'am Neelma Naz

Table of Contents

Introduction

In this assignment, we implemented a **Book Recommendation System** using two different approaches:

1. **Content-Based Filtering (CBF)**
2. **Collaborative Filtering (CF)**

The goal was to recommend books based on book information as well as user rating behaviour and compare both approaches.

Part 1: Dataset Selection

For this assignment, the Book-Crossing Dataset was used. This dataset contains information about books, users, and their ratings. The main fields used include Title, Author, Publisher, ISBN, User ID, and Rating. The dataset has many users and books, making it suitable for building and testing recommendation systems. Initial Dataset details are shown in figure 1.

```
        ISBN                                          Title  \
0  0195153448                                Classical Mythology
1  0002005018                                       Clara Callan
2  0060973129                               Decision in Normandy
3  0374157065  Flu: The Story of the Great Influenza Pandemic...
4  0393045218                             The Mummies of Urumchi

               Author  Year                Publisher
0  Mark P. O. Morford  2002  Oxford University Press
1  Richard Bruce Wright  2001    HarperFlamingo Canada
2        Carlo D'Este  1991          HarperPerennial
3    Gina Bari Kolata  1999     Farrar Straus Giroux
4     E. J. W. Barber  1999  W. W. Norton & Company
   User-ID        ISBN  Rating
0   276725  034545104X       0
1   276726  0155061224       5
2   276727  0446520802       0
3   276729  052165615X       3
4   276729  0521795028       6
Book Column Shape:  (271379, 5)
Ratings Column Shape:  (1149780, 3)
```

Figure 1: Dataset Details

Part 2: Data Preprocessing

The Books and Ratings datasets were loaded and cleaned before modelling. Missing values were removed, and book ratings were converted into numeric format. Zero ratings were removed since they do not represent user preference. To reduce processing time and improve recommendation quality, only books with at least 50 ratings and users with at least 30 ratings were selected. These preprocessing decisions helped in creating a cleaner and more reliable dataset. The reduced dataset size is shown in Figure 2.

```
Reduced dataset shape: (896, 3)
```

Figure 2: Reduced Dataset Shape

Part 3: Content-Based Filtering

Content-Based Filtering was implemented using book metadata. The Title, Author, and Publisher were combined into a single text feature for each book. This combined text was converted into numerical form using TF-IDF vectorization. After that, cosine similarity was calculated between all books to measure similarity. The TF-IDF matrix shape was displayed to show the number of books and features used. For a selected book, the system generated the top 10 most similar books based on similarity scores as shown in figure 3.

	Title
278	Wicked: The Life and Times of the Wicked Witch...
338	Everything's Eventual : 14 Dark Tales
145	Life of Pi
126	Life of Pi
2160	A Beautiful Mind: The Life of Mathematical Gen...
63	The Secret Life of Bees
1005	The Beach House
324	The Secret Life of Bees
999	Cradle and All
2318	The Beach House

dtype: object

Figure 3: Cosine Similarity used to refer 10 books

Part 4: Collaborative Filtering

Collaborative Filtering was implemented using the Item-Item approach. A pivot table was created where rows represent users, columns represent books, and values represent ratings. Missing values were filled with zeros. Cosine similarity was used to compute similarity between books based on user ratings. A small portion of the similarity matrix was displayed as shown in figure 4.

ISBN	002542730X	0060096195	006016848X	0060173289	0060175400
ISBN					
002542730X	1.000000	0.000000	0.029427	0.000000	0.087169
0060096195	0.000000	1.000000	0.000000	0.000000	0.061962
006016848X	0.029427	0.000000	1.000000	0.000000	0.073099
0060173289	0.000000	0.000000	0.000000	1.000000	0.091464
0060175400	0.087169	0.061962	0.073099	0.091464	1.000000

Figure 4: Similarity matrix based on user rating

For a selected user, the system recommended top 5 books that the user had not rated yet.

	Title	Author
37	To Kill a Mockingbird	Harper Lee
90	The Catcher in the Rye	J.D. Salinger
2143	Harry Potter and the Sorcerer's Stone (Harry P...	J. K. Rowling
5506	Harry Potter and the Order of the Phoenix (Boo...	J. K. Rowling
6933	Harry Potter and the Goblet of Fire (Book 4)	J. K. Rowling

Figure 5: Top 5 Recommended Books

Comparison of CBF and CF

Content-Based Filtering recommends books based on book features and works well even for new users but may lack diversity. Collaborative Filtering recommends books based on user behaviour and provides more diverse results but suffers from the cold-start problem. Both methods have advantages and using them together can improve recommendation quality.

Conclusion

In this assignment, both Content-Based Filtering and Collaborative Filtering techniques were successfully implemented to recommend books. The results show that Content-Based Filtering relies on book information, while Collaborative Filtering depends on user rating patterns. Each approach performs well in different situations and combining both can lead to better recommendations.