**‘’AIE425 Intelligent Recommender Systems, Fall Semester 24/25’’**

**‘’Assignment 1Neighborhood CF models (user, item-based CF).’’**

**‘’ Name: Mohamed Mosad Fawzy /ID: 221100611‘’**

**1-**I searched for Amazon and Netflix

**2-**I choose Amazon as a data source for my assignment

**3-**By product reviews and asking customer questions about the product, they use 5 stars rating

**4-   
to clean it and express the feedback in the form of integer values.**

I web scrape from amazon dataset about health and personal care, step one was data collection, step two data cleaning preprocessing like (remove duplicates, filter outliers), step three dropping the columns that is not important like ('text', ‘parent\_asin','title','images','parent\_asin','helpful\_vote','timestamp','verified\_purchase’)

**5-**My data consists of 10 columns (rating, title, text, images, asin, parent\_asin, user\_id , timestamp, helpful\_vote, verified\_purchase) First step was dropping the columns that is not important like ('text', 'parent\_asin','title','images','parent\_asin' , 'helpful\_vote','timestamp','verified\_purchase’) second step make sure there are no empty values

**6-**I generate 6\*6 matrix but I deleted user3 and product 6 because the have many null values

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 |
| User1 | 3 | 3 | 1 | 3 | NULL |
| User2 | NULL | 2 | 2 | 5 | NULL |
| User3 | 5 | 4 | NULL | 4 | 2 |
| User4 | NULL | 4 | 5 | 4 | 4 |
| User5 | NULL | 3 | 5 | 3 | 3 |

*Community Verified icon*

**7- Give a complete description of the created dataset.**

My dataset has 10 columns:-

1. Rating: A numerical score given by user to the product (from 1to 5 stars)
2. Title: a short review
3. Text: the review that all the users give the product and their opinion
4. Images
5. Asin: Amazon identification number, special for each product
6. parent\_asin: This refer to a parent product if the item is part of a larger product family or variation.
7. user\_id: A unique identifier for the user who wrote the review.
8. timestamp: A numerical representation of the date and time when the review was submitted.
9. helpful\_vote: Indicates how many other users found the review helpful, which can influence its visibility.
10. verified\_purchase: A boolean (TRUE/FALSE)

**8- Compute the average rating and copy the results into your report under the "Assignment Results" section**

User-average ratings:-

User1: (3+3+1+3) / 4= 2.5 User2: (2+2+5) / 3= 3.0

User3: (5+4+4+2) / 4= 3.75 User4: (4+5+4+4) / 4= 4.25

User5: (3+5+3+3) / 4= 3.5

Product- average ratings:-

P1: (3+5) / 2= 4.0 P2: (3+2+4+4) / 4= 3.25

P3: (1+2+5+5) / 4= 3.25 P4: (3+5+4+4+3) / 5= 3.8

P5: (2+4+3) / 3= 3.0

**9- Give a complete background/overview about user-based and item-based CF algorithms, and their detailed analytical solutions.**

|  |  |  |
| --- | --- | --- |
|  | **User-based collaborative filtering:-** | **Item-based collaborative filtering:-** |
| **Basic idea** | The basic idea is to determine users, who are similar to the target user A, and recommend ratings for the unobserved ratings of A by computing weighted averages of the ratings of this peer group. Generally, the k most similar users to B can be used to make rating predictions for B. | The idea is, in order to make the rating predictions for target item B by user A, the first step is to determine a sets of items that are most similar to target item B.  The ratings in item set S, which are specified by A, are used to predict whether the user A will like item B. |
| **similarity** | Similarity functions are computed between the rows of the ratings matrix to discover similar users. | Similarity functions are computed between the columns of the ratings matrix to discover similar items. |
| **Type of similarity** | In the user-based we use cosine\_similarity and pearson correlation coefficient | In the item-based we use Adjusted cosine\_similarity |
| **Equation** |  |  |

**‘’Assignment results section’’**

**10-**

**11-**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | P1 | P2 | P3 | P4 | P5 | cosine | pearson | Mean  Rating |
| User1 | 3 | 3 | 1 | 3 | NULL | **1** | **1** | **2.5** |
| User2 | NULL | 2 | 2 | 5 | NULL |  |  | **3** |
| User3 | 5 | 4 | NULL | 4 | 2 |  | **0.777** | **3.75** |
| User4 | NULL | 4 | 5 | 4 | 4 |  | **-1** | **4.25** |
| User5 | NULL | 3 | 5 | 3 | 3 |  |  | **3.5** |
|  | **1** |  |  |  |  |  |  |  |

as we can see from the previous table that cosine is better than pearson and pearson in not good with data that have a lot null values

**13- the prediction rating:-**

The **top\_N** in the cosine\_similarity = **(0.9941,0.9185)**

The **top\_N** in the pearson =(**0.777,0.4923)**

The **top\_N** in the Adjusted cosine\_similarity = **(0.7474,0.636)**

**Cosine prediction** 🡪

**pearson prediction** 🡪

**Adjusted cosine prediction** 🡪

**14-**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **cosine\_similarity** | **Pearson correlation** | **Adjusted cosine\_similarity** |
| **Predication** |  |  | **3** |
| **Top\_N** | **0.9941,0.9185** | **0.777,0.4923** | **0.7474,0.636** |