**INFS 770 Advanced Data Mining Applications**

**Assignment 2**

**Due 3/15/2022 (the grace period ends on 3/22/2022)**

In this assignment, you will need to implement a simple recommender system using a book rating data set DBbook\_train\_ratings.tsv (reference: <https://lists.w3.org/Archives/Public/public-rww/2013Dec/0002.html>). The first column of this data set contains user IDs. The second column contains itemIDs (i.e., book ids). The third column contains the rating scores (1 – 5). The purpose of studying this data set is to create a data mining model that recommend books to users. The data set (DBbook\_train\_ratings.tsv) can be downloaded from D2L.

**Please submit an iPython Notebook (you don’t really need to submit the dataset. If you want to change the original dataset, you need to write code to do that). Please use “run all” to run your code before you submit so that your iPython notebook will show the outputs of your code. You will lose 1 point if you do not “run all”.**  You probably want to copy and modify the code from the ipython notebook ml\_100k posted on D2L.

Please develop an ipython notebook titled **770\_hw2\_*yourlastname*** and finish the tasks below, where (C) indicates that you need to write code for the task, (O) indicates that you need to show output, and (A) that you need to type your answers using Markdown text. In your iPython notebook, at the beginning of each cell, you need to indicate which task the cell is about. For example, in the cell related to task 1, you should first type “# Task 1: Import data”. **If you do not clearly label the cells, you will lose 1-2 points (out of 18 points).** Whenever you see “print” in the questions, you need write print statement to print the intended outputs.

1. Please write code to print the number of unique users and the number of unique books in this data set. (C)(O)

2. Please write code to create the utility matrix. Each row of this matrix represents a user, and each column represents an item. Print the first 10 rows of matrix. Please write code to print **the number** and **the percentage** of cells in the utility matrix that are not populated. Please write code to fill these empty cells with 0s. (C)(O)

3. Please write code to print the top 5 similar users to userID 2 based on Euclidean distance. (C)(O)

4. Please write code to print the Euclidean distance between itemID 18 and itemID 1. Please write code to print the Enclidean distance between itemID36 and itemID 1. Write a print statement that tells me between itemID36 and itemID18, which is more similar to itemID 1 and why. For example, you can write a print statement like *print(“itemID36 is more similar to itemID 1 because some reason…” ). (C)(O)*

5. Please write code to print the top 5 similar items to itemID 8010. (C)(O)

6. Write code to remove books and users with less than 20 rating scores from the utility matrix by copying and maybe modifying the following codes. Write code to print the shape of the dataset. (C)(O)

df\_item\_fre = df\_data1.groupby("itemID").count()

df\_user\_fre = df\_data1.groupby("userID").count()

selected\_items = df\_item\_fre[df\_item\_fre["userID"]>20].index

dense\_matrix = dense\_matrix[selected\_items]

selected\_users = df\_user\_fre[df\_user\_fre["itemID"]>20].index

dense\_matrix = dense\_matrix.loc[selected\_users]

7. Please use the dataset you obtained from task 6 and write code to remove users that haven’t rated itemID8010, and then please write code to print the counts of the different rating scores of this item (hint: use the function value\_counts()). Print the shape of the dataset. (C)(O)

8. Write code to partition the data set you obtained from 7 for validating the performance on predicting rating on itemID 8010. Randomly select 25% of the users as the testing set and the others as the training set. Please print the dimensions of the training set and the testing set. Please write code to print the mean rating of itemID 8010 in the training set and its mean rating in the testing set. (Hint: use dense\_matrix[8010].mean() method to calculate the means) (C)(O)

9. Use the training and test dataset obtained in 8 and write code to 1) print the userID of the the user in the 5th row (not userID5) in the test dataset, and 2) predict this user’s rating of itemID 8010 based on the top 5 similar users in the training dataset, and print the user’s **predicted rating and the actual rating** of the book. (C)(O)

10. Create text boxes to write down the answers to the following questions. (A)

1. Use your own language (<= 6 sentences) to briefly describe content-based filtering.
2. There are different models that can used in model-based collaborative filter such as SVD and neural network. In this course, I just introduced SVD model based collaborative filtering, which is the most commonly used collaborative filtering algorithm. Use your own language (<= 6 sentences) to briefly describe SVD model-based collaborative filtering ()
3. Summarize the differences between content-based filtering and SVD model-based collaborative filtering? (at least two differences).