**Project Part I: Dataset Study**

The first part of the project is to choose a proper dataset for making a recommendation system and studying then preprocessing the dataset. First, you can choose a dataset in the following section *Datasets We Provide*. Or, alternatively, you can search for other datasets that are suitable for building a recommendation system. Second, you will be asked to utilize **Python** libraries to analyze and preprocess the dataset, including but not limited to data cleaning, integration and transformation. Finally, you should prepare a **4-page** report regarding how you execute the preprocessing, and then show the statistical results in figures or tables. Besides all these, any other things you've tried to make the quality of the dataset nicer can be included in the report and might get you some bonus points.

**Datasets We Provide**

We select 2 default options for you, or you can find a dataset that satisfies those requests listed in part **3.** ***You only need to do 1 dataset!***

The training datasets of part **1.** and **2.** can be found [here](https://drive.google.com/drive/folders/1Gn1giWRGGf_vq47n7eAt4BxqntRo-FPg?usp=sharing).

**1. Book-Crossing Dataset**

This is a dataset collected from [a book crossing (圖書漂流) community](https://www.bookcrossing.com/?), containing 278,858 users with 1,149,780 ratings about 271,379 books. There are 3 csv files in the folder above: **Ratings.csv** contains 3 columns *User-ID*, *ISBN* and *Book-Rating*. This file is the training data, containing 1,034,802 (90%) records, by random and stratified sampling. **Users.csv** contains the detailed information for all 278,858 users, including their locations and ages. **Books.csv** contains the titles, authors, and much more detailed information about all 271,379 books. Further descriptions can be found in the original dataset website [here](http://www2.informatik.uni-freiburg.de/~cziegler/BX/). Finally, for those students using this dataset, your recommendation system would be tested by the 114,978 records preserved, and ranked by the accuracy to compare with other students. *±1 from the ground truth rating* would still be considered accurate.

**2. The Echo Nest Taste Profile Subset**

This is the official user dataset of the renowned **Million Song Dataset (MSD).** The dataset contains 1,019,318 users and 384,546 songs included in MSD, with each record presented as a *"user - song - play count"* triplet. There are totally 48,373,586 such triplets in the dataset. The training data, **training\_triplets.csv** contains 43,536,227 triplets (90%), randomly sampled. There are 2 other files: **users.csv** and **songs.csv**. These 2 files contain the IDs of all 1,019,318 users and 384,546 songs, respectively. That is, no matter in training data or in testing data, the users and songs appeared can all be found here. For those students using this dataset, the remaining 4,837,359 triplets (10%) would be preserved as test cases to examine your recommendation system. The performance will be ranked by the accuracy to compare with other students. *±10 from the ground truth counts* would still be considered accurate. Further detailed descriptions can be found in the original website [here](http://millionsongdataset.com/tasteprofile/). Finally, some tips for those using this dataset:

a. This file is big (2.55GB), so please deal with this carefully.

b. The training data provided only contains the song ID and user ID. It is recommended that you also take other related information into account, such as similar songs, tags, etc. Regarding these extra features, there are several datasets as well as their descriptions on the MSD sites [here](http://millionsongdataset.com/pages/getting-dataset/) and [here](http://millionsongdataset.com/pages/additional-datasets/). You might find them helpful in boosting the performance of your system.

**3. Other datasets**

If you are not interested in the above datasets we provided at all, you are allowed to find a dataset on your own. Nonetheless, the dataset you chose should satisfy the following features, so that it can be used to construct a recommendation system based on collaborative filtering.

a. The dataset *must* have **a set of users** and at least **a set of items**, in terms of products, movies, point-of-interests, etc.

b. The dataset *better* has other detailed information about the users and the items. For example, locations, ages and genders of the users, or titles, producers of the items, etc.

**Dataset Preprocessing and Analysis**

After you choose a dataset, you are required to do some preprocessing and analysis on your data. Following are some questions that you need to answer.

*A. Basic Questions (60%)*

**Q1.** (5%) Which labels/features/records can be used to represent user preference (i.e., how much a user likes the item)? Also, what is the total range of the labels/ features/records? Does the higher the value is, the better a user prefers an item, or conversely the lower the better?

**Q2.** (25%) What did you do to clean the data? Have you done any transformation, integration or deletion? Briefly explain why you did all these cleaning actions.

**Q3.** (15%) Sort the users by the number of interactions, and observe what is the minimum number of interactions (*Y*) generated by the top-*X*% users? Draw some appropriate figures and briefly explain what your insight is.

(*Hint: You can set X-axis as percentage of users, with 10% for a class interval, and set Y-axis as number of interactions. The interactions are for instance, books a user read in Book-Crossing Dataset or Songs a user listened in Echo Nest Dataset.)*

**Q4.** (15%) Sort the items by the number of interactions, and observe what is the minimum number of times (*Y*) that the top-*X*% items have been interacted with? Draw some appropriate figures and briefly explain what your insight is.

(*Hint: You can set X-axis as percentage of items, with 10% for a class interval, and set Y-axis as times of items that have been interacted with, which are for instance, number of users that a book has been read in Book-Crossing Dataset, or number of users that a song has been listened in Echo Nest Dataset.)*

*B. Free Exploration (40%)*

**Q5.** (20%) Did you use other datasets or resources to get more information regarding the users and items? If you do, what did you use and why did you choose them? Where did you apply these extra features to help the analysis? And if you don't, what features do you think would be beneficial and where to find them?

**Q6.** (20%) Freely study your dataset and come up with a question/idea to analyze.

a. The motivation of the question needs to be meaningful and valuable. Why do you want to study this question? What do you expect to bring out from the analysis?

b. Define and formulate your question clearly. It is suggested that you provide an example for easy understanding.

c. At least 1 figure to show your analysis results and several sentences for explanations. Make sure that the figure is representative enough for the question.

d. Did the results turn out to be the same as your expectation? Why or why not?

**Hand-in Rules**

*The project will be divided into 2 ~ 3 parts, each with full credits (100%). The final score of your project will be the arithmetic average of the grades of these parts.*

You shall hand in a report with **a maximum of 4 pages in pdf format**, as well as your **Python code.** You are allowed to write your report in either English or Chinese. While turning in, name both your files as your **STUDENT ID.** For instance,

106304027.pdf + 106304027.py or 106304027.pdf + 106304027.ipynb

You don't need to zip these files.

Finally, the deadline of Project Part I will be May 15, 23:59:59.

**Penalty**

You will get a **20% discount** for your grades for **late submission**. For those who don't even bother doing the make-up submission, you will get **0** points after the make-up deadline (announced after the original deadline).

You will get **0 points** without any review if the files you handed in are in the **wrong format**.