

HW#2 MovieLens Data Analysis

MovieLens 100k data set is most common in the recommender systems research area. This dataset consists of:

- ✓ 100,000 ratings (1 ~ 5) from 943 users on 1,682 movies.
- ✓ Simple demographic info for the users (age, gender, occupation, zip).
- ✓ Each movie has one or more genres.

Please refer to the README.txt file for detailed data description.

Step-by-step instructions for completing assignments

1. Download “ml-100k.zip” file from <https://grouplens.org/datasets/movielens/100k/>
2. Make a list of the top 100 most popular movies
 - A. Extract a list of movies with the highest number of ratings 1) for all users, 2) by gender, and 3) by occupation.
 - B. Extract a list of movies with the highest average ratings 1) for all users, 2) by gender, and 3) by occupation.
3. Calculate the ratings by the genre
 - A. Count the number of ratings by each genre for 1) all users, 2) genders, and 3) all occupations and sort them in descending order respectively.
 - B. Measure the average ratings by genre for 1) all users, 2) genders, and 3) all occupations and sort them in descending order respectively.