CS 6220 Data Mining — Assignment 1

Exploring the MovieLens Dataset

On this assignment, you'll describe how a simple recommendation system can be crafted to leverage user ratings and generate individual recommendations for movies, products, and any thing else that can be rated.

Below are two links that will serve as a solid starting point for what you will submit:

1. A notebook with some initial data exploration of the MovieLens dataset. LINK 1
2. A blog post containing some further analysis of that dataset. LINK 2

Objective:
Download the data (link 1 above contains instructions on how you can do that) and extend the presented data exploration to include:

- [10 pts] An aggregate of movie ratings by men of age above 25 for each particular genre, e.g., Action, Adventure, Drama, Science Fiction, ... Note, Action|Drama|Thriller' is not considered a unique genre. The movie that has a genre like this belongs to all three genres.
- [5 pts] The top 5 ranked movies by the most number of ratings (not the highest rating).
- [20 pts] Average movie ratings between users of different age groups (<18, 18-29, 30-49, 50-69, 70 and above>)
- [25 pts] Pick a movie of your choice and for all movies of the same year, provide a breakdown of the number of unique movies rated by 3 ranges of age of reviewers (a) under 18 (b) 19 to 45, including 45 (c) Above 45.
- [20 pts] A function that takes in a `user_id` and a `movie_id`, and returns a list of all the other movies that the user rated similarly to the given movie, i.e. with the same rating. Demonstrate that your function works.
- [20 pts] Some other statistic, figure, aggregate, or plot that you created using this dataset, along with a short description of what interesting observations you derived from it.

You may use the code available on both of the provided notebook files as a starting point.

Submission:

Submit your ipynb file with a pdf of your assignment solution through the assignment submission portal on Canvas. No need to zip the files.