NORTHEASTERN UNIVERSITY, COLLEGE OF COMPUTER AND INFORMATION SCIENCE

CS 6220 Data Mining — Assignment 1
Due: May 24, 2018 (100 points)

Exploring the MovieLens Dataset

In 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 anything 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 the movie ratings for each particular genre.

- [10 pts] The top 5 highest ranked genre by women.

- [10 pts] The top 5 highest ranked genre by men.

- [20 pts] A breakdown of a movie's ratings by age, using any movie of your choice.

- [30 pts] A function that given a `user_id` and `movie_id`, returns a list of `user_ids` for other users that rated the movie identified by the `movie_id` with the same score.

- [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:
Follow the same steps as in Assignment 0 and submit your GitHub link through Slack.