Northeastern University
CS6220 – Data Mining Techniques
Fall 2017, Derbinsky

Homework 3, Part 2

Include the single notebook (`starter-3.4.ipynb`) for this part of the assignment, along with all work from Part 1, in a **single zip** for submission to Blackboard.

4 Recommending Movies

You will use Pandas, NumPy, and SciKit-Learn to analyze and recommend movies from the MovieLens 100K dataset. You will first read in portions of the dataset as DataFrames, clean the data, and then perform exploratory queries. Then you will implement evaluation, distance/similarity, and neighborhood functions. And finally you will implement and evaluate two approaches to recommender systems.

At the end of the assignment you are required to extend one of the approaches, per your choice (plenty of room has been left for improvement). You can also implement an actual end-to-end movie recommender for extra credit.

You will do all the work inside the `starter-3.4.ipynb` Jupyter notebook file. Make sure to complete all items flagged as `TODO`.

The starter file has not been cleared of output – meaning that when you open it, you will still see results from the last successful run. This is intentional, such that you know what is expected as output. However, you will **NOT** receive any credit for items in which you do not supply code to reproduce these outputs. For example, during EDA you are asked some questions (e.g. "Which movie has the most ratings?") – simply providing the answer (which is given to you in the output) will gain you no credit, instead you must provide code that finds the answer from the data (and thus reproduces the output).