Homework 1 (Due June 10)

CS4412: Data Mining
Kennesaw State University
Summer 2022

In this project, we will explore the MovieLens dataset, available at:

`https://grouplens.org/datasets/movielens/`

The MovieLens dataset is composed of over 25 million movie ratings, covering 62,000 movies by 162,000 users. For this particular assignment, we will consider a smaller version of this dataset composed of 100,000 ratings over 9,000 movies and 600 users.

Included in this homework are two files: the file `ratings.csv` which includes all of the movie ratings, and the file `movies.csv` which contains information about the movies being rated. Each line of the ratings dataset consists of a single movie rating. For example, consider the first three rows of the ratings dataset:

userId	movieId	rating	timestamp
1	1	4.0	964982703
1	3	4.0	964981247
1	6	4.0	964982224

Each movie rating is associated with:

- a user (represented as an ID)
- a movie (represented as an ID)
- a rating
- a timestamp (in UNIX time)

In this homework, we ignore the timestamp.

We can associate each movie ID of the ratings dataset, with an entry in the movies dataset. Each line of the movie dataset consists of a single movie. For example, consider the first three rows in the movie dataset:

movieId	title	genres
1	Toy Story (1995)	Adventure, Animation, Children, Comedy, Fantasy
2	Jumanji (1995)	Adventure, Children, Fantasy
3	Grumpier Old Men (1995)	Comedy, Romance

Each movie is associated with:

- a movie ID (the same IDs used in the ratings dataset)
- a title
- a list of genre's

In this homework assignment, our goal is to get some practice navigating a dataset, and answering some basic questions about a dataset.

The python programming language is popular for working with datasets, whether it is in the context of data mining, or in related fields such as machine learning and artificial intelligence. A short sample python script, called `hw1.py`, is included for loading this dataset and doing some elementary analysis on this dataset. This assignment can be completed by modifying this python script. Learning python is not required to complete this assignment, and you are welcome to use any other programming language to complete this assignment. However, it is strongly suggested that one starts to become familiar with python, as many data mining libraries and tools are available in python, which we will also be using in this class.

Questions: Answer the following questions regarding this dataset.

1. Which movie has the highest average rating, among all movies that are included in the Children genre, that also has at least 100 ratings?
2. Which movie has the lowest average rating, among all movies that are included in the Sci-Fi genre, that also has at least 100 ratings?
3. Consider all users who rated the movie "Toy Story (1995)" a rating of 5.0. Consider all of the movie ratings that these users provided. Among these movies that were rated by at least 10 of these users, what is the movie (besides "Toy Story (1995)") with the highest average rating? Such a movie could be thought of as a recommendation for users who liked the movie Toy Story.
4. A movie recommendation is referred to as *serendipitous* if it (1) recommends a movie that a user would like, but (2) it is also a movie that the user would not have found for themselves. Provide an example of a serendipitous movie recommendation. (This is an open-ended question).

Turn in: the answers to the above questions as a text file or as a .pdf file (no Microsoft Word .doc's please), onto the course website under "Assignments" and "Homework 1." Assignments are due Friday, June 10 by 11:59pm. Please start early in case you encounter any unexpected difficulties. Late projects are accepted without penalty until solutions are posted.

Included files:

- `hw1.pdf`: this document
- `hw1.py`: a python script that reads in the datasets, and computes some basic statistics, to help you get started
- `data/movies.csv`: the dataset of movies
- `data/ratings.csv`: the dataset of movie ratings