Assignment 1

CS-UH-2218: Algorithmic Foundations of Data Science

Assignments are to be submitted in groups of two or three. *Upload the solutions on NYU classes as one PDF file for theoretical assignments and separate source code files for each programming assignment. Submit only one copy per group. Clearly mention the participant names on each file you submit.*

Mandatory Reading: Chapters 1 and 2 of the textbook [MMDS].

Optional Reading: Section 12.4 of the textbook [FoDS] (`https://www.cs.cornell.edu/jeh/book.pdf`) provides an excellent short review of basic probability theory.

Problem 1 (10 points)**.**
Learn to use `pyplot` from `https://matplotlib.org/users/pyplot_tutorial.html` (or any other resource you prefer). Plot the function $f(x) = 1-(1-x^r)^b$ for $(r, b) \in \{(1, 1), (2, 2), (5, 5), (1, 5), (5, 1), (5, 20)\}$ and $x \in [0, 1]$. Plot the six functions obtained for the different values of r and b as subplots of the same figure.

Experiment with different values of r and b to determine a pair of (not too large) positive integers r and b so that the function $f(x)$ has the maximum slope around $x = 0.4$.

Problem 2 (10 points)**.**
Solve Exercise 1.2.1 of [MMDS].

Problem 3 (10 points)**.**
Solve Exercise 2.3.1 of [MMDS].

Problem 4 (20 points)**.**
Download `ml-latest-small.zip` from `https://grouplens.org/datasets/movielens/latest/` and unzip it into a folder. Read the corresponding `README.html` file in the above website to understand the format of the files contained in the folder. The first line in each csv file is the header and the rest of lines contain data. Remove the first line from each of the csv files manually to make programming easier.

Write *a* Spark program to compute the following:

1. The average number of users a movie is rated by.
2. For each genre, the average rating of all movies in that genre.
3. The names of the top three movies (by average user rating) in each genre.
4. Top ten movie watchers ranked by the number of movies they have rated.
5. Top ten pairs of users ranked by the number of movies both of them has watched.

When considering the ratings for a movie, we only consider the users that have rated the movie. For instance, to compute the average rating for a movie we add all the ratings of the movie and divide by the number of users that have rated the movie.

Your program should assume that the csv files are located in the same directory as your program. Along with your code give brief explanations for the algorithm used for each of the above tasks.