Assignment 6 - Spark Tutorial

Submission Deadline:_

10th May (For students with presentation on 3rd May 2018)

3rd May (For students with presentation on 10th May 2018)

This assignment is a spark tutorial. It will help you to grasp the basic concept of functionalities and operations offered by the APACHE SPARK.

Question 1 (35 marks)

You are provided dataset “Movies.csv” that contains information about 1600 movies with properties such as year, length, main actor and actress, director and popularity. Your task is to read the file in SPARK RDD and efficiently perform the following queries on the dataset using Spark (Scala).

You must use the spark functionality of persisting data in memory and partitioning. Try to write an efficient query.

1. Find the total number of movies of each genre. Consider only the movies released after 1960 and have length greater than 90 minutes.
2. For each actress, find the movie she acted it. You must print the names of the movies and the year of the movie.
3. Find the 25 most popular movies released during 1980-1990
4. Find the average length of the movies of each genre.
5. Find the lead actor and lead actress pair who have acted in more than one movie together.
6. Find the names of directors who directed movies of both ‘Mystery’ and ‘Drama’ Genre.
7. Find each director, find the average, max, min ranking of his movies.

Question 2 (15 marks)

In this example, you will use the given “Movies.csv” data file. Your task is to write an **efficient** spark code for finding movies (released after 1970) that might be similar or related to a movie made in the 1960’s (1960-1969).

To keep things simple, for comparing two movies we will use the following fields: Genre, director, lead actor, lead actress.

So, we can say two movies might be similar if they have same genre, director and same lead actor **or** actress.

Submission Details: *You must submit Source code, Jar file and Output file. No form of plagiarism will be tolerated.*