CSS 735: HW 3

Due Date: *Oct 20 (F) 11:59 PM* ***Overdue submissions*****:** subject to penalties

Download the files **movies.csv** and **movie_ratings.csv** from Blackboard. Login to your Databricks Community Edition account. Follow the instructions from HW2 to upload the two files to your Databricks account (you may already have uploaded the file movies.csv from hw2).

1) Create a Notebook for this homework assignment with the name Asg3_YourFamiliyName_StudentID.**scala**.

2) Use the following commands to load the two files and register them as SQL tables:

```
val df1 = spark.read.option("header", "true")
  .option("inferSchema","true")
  .csv("/FileStore/tables/movies.csv")

df1.createOrReplaceTempView("movies_table")

val df2 = spark.read.option("header", "true")
  .option("inferSchema","true")
  .csv("/FileStore/tables/movie_ratings.csv")

df2.createOrReplaceTempView("movie_reviews_table")
```

3) **[10points]** Write DataFrame-based Spark code to find the number of distinct movies in the file movies.csv.

4) **[10 points]** Write DataFrame-based Spark code to find the titles of the movies that appear in the file movies.csv but do not have a rating in the file movie_ratings.csv. Remark: the answer could be empty.

5) **[10 points]** Write DataFrame-based Spark code to find the number of movies that appear in the ratings file (i.e., movie_ratings.csv) but not in the movies file (i.e., movies.csv).

6) **[10 points]** Write DataFrame-based Spark code to find the total number of distinct movies that appear in either movies.csv, or movie_ratings.csv, or both.

7) **[10 points]** Write DataFrame-based Spark code to find the title and year for movies that were remade. These movies appear more than once in the ratings file with the same title but different years. Sort the output by title.

8) **[10 points]** Write DataFrame-based Spark code to find the rating for every movie that the actor "Branson, Richard" appeared in. Schema of the output should be (title, year, rating)

9) **[20 points]** Write DataFrame-based Spark code to find the highest-rated movie per year and include all the actors in that movie. The output should have only one movie per year, and it should contain four columns: year, movie title, rating, and a list of actor names. Sort the output by year.

10) **[20 points]** Write DataFrame-based Spark code to determine which pair of actors worked together most. Working together is defined as appearing in the same movie. The output should have three columns: actor 1, actor 2, and count. The output should be sorted by the count in descending order.

Deliverables

- Upload to Blackboard your **notebook** and the corresponding **HTML** file named in the following format:
    - Asg3_YourFamiliyName_StudentID.**scala**
    - Asg3_YourFamiliyName_StudentID.**html**
- **Ensure that your notebook can run from start to finish without crashing** (e.g. click Run=>Run all and double check the output of each cell before submitting)
  To download, click File=>Export=>Source file, and File=>Export=>HTML
- To prove that your code works correctly, you will also submit screenshot of the output of your program in a file with an extension of **.png** or **.pdf**.

Policies and procedures

- Assignments must be completed **individually and independently,** and **NOT** in a group.
- Use comments to document and explain your code where needed.
- Assignments are to be handed in by the **due date**; otherwise, **penalty per day** will be incurred: 1 day late 10% penalty; 2 days late 20% penalty; after 2 days no credit will be given for a late assignment. Homework that is habitually late will not be accepted.
- Any incomplete work that is submitted will be considered for partial marks.
- If exceptional circumstances have occurred that have kept you from submitting your assignment on time, you should contact the course instructor as soon as possible. If you are ill, a valid **document**, e.g., a doctor note, must be provided.