CSC 735 – HW 3  
Due Date: Tuesday, November 20, 2018 at 9:30 am

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 **hw3.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")

1. **[10points]** Write DataFrame-based Spark code to find the number of distinct movies in the file movies.csv.
2. **[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.
3. **[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).
4. **[10 points]** Write a DataFrame-based Spark code to find the total number of distinct movies that appear in either movies.csv, or movie\_ratings.csv, or both.
5. **[10 points]** Write a 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.
6. **[10 points]** Write a 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)
7. **[20 points]** Write a 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.
8. **[20 points]** Write a 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.
9. **[Extra Credit 20 points]** Think of a nontrivial data analytics task that you can perform on the movies.csv and movie\_ratings.csv datasets. State the task and provide a DataFrame-based solution for it. Do not repeat what you proposed for hw2. The number of points that you will receive will be based on how interesting and nontrivial your task is. You may use rollups, pivots, window functions or any material mentioned in the textbook even if that material has not been covered in class. If you use material not covered in class, please reference the page numbers in the book where that material is mentioned.

**What to turn in**: upload to Blackboard a file named **hw3.scala** with your answers to the homework questions. Make sure to **also turn in a stapled hard copy of your solution in class on the due date**. Write your name at the top of your file. Notice: you may use File => Export => Source File to download your Databricks Notebook for the file to upload to Blackboard and print out for a hard copy.