Assignment 3 - Spark Data Frames and Clustering

Submission Deadline: ***11th May 2020***

Submit the code along with the ouput of each part in a word file on google classroom.

Question 1: SPARK DATAFRAMES

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.

Load the given dataset into Spark Data-Frames and answer the following queries using Data Frame functions only. You are not allowed to write the SparkSQL queries.

1. Find the title, year and director of action films that won an award.
2. For each award-winning actor, find the movies he acted it. Print the names of the movies and the director of the movie.
3. Find the top 10 most popular movies that did not win an award.
4. Find the 10 least popular movies that were released before 1980.
5. Find the average length of the movies of each genre.
6. Find the actor and actress pair who have acted in more than three Comedies together.
7. Find the names of actors who acted in movies of both 'Comedy' **and** 'Drama' Genre.
8. Find the names of actors who acted in movies of both 'Comedy' or 'Drama' Genre.
9. Find the names of actors who did not acted in any 'Comedy'.
10. Find each actor, find the mean, max, min ranking of his movies.
11. List the number of movies released in each decade starting from 1960's.
12. Find the number of movies released in each year.
13. Find the number of movies released in each year of each genre. Consider only the movies with length greater than 100 minutes.
14. Sort the movie's release before 1990 by the title.
15. Find the movies with long titles. A movie title is considered long if it is greater than 50 alphabets.