First Admission Exercise – Hadoop Movie Analysis

Solve for the given .csv movie file the following tasks in a programmatically way. Before you start, try to understand in which way the MapReduce paradigm can be applied. All requested tasks have to be integrated in compiled .jar file.

Individually (THE SELECTION CAN BE MADE ON THE ELEARNING PORTAL):

Solve individually one of the following tasks

1. Which "production company" produced the most movies in every year, since 1980? Show a chronological order, since 1980, for each year the production company, amount of movies and exact movie titles need to be depicted.
2. In which country have the most movies been produced (according to the file)? Show a distribution of all countries together with the amount of produced movies and the for each the movie with the highest **profit** (use the revenue and budget**).**
3. How often was the "original_title" used as the actual title? Determine the overall requested amount and provide a detailed list containing the id and movie title.
4. What is the average runtime of each movie genre? Give a list of all genres together with the average runtime and the movie which comes closest to this value.
5. What is the average revenue of each movie genre? Give a list of all genres and the average revenue per movie of each of those. Additionally present to each genre the movie with the biggest revenue.
6. Based on the given budget and the revenue, what is the most successful movie for each genre? Show a distribution of all genres together with the best performing movie in terms of their **profit** (use the revenue and budget)**.**
7. How many musician movies have been produced since 1980? Show a chronological order, containing for each year the amount of movies and exact titles.
8. What is the most prevailing original language, spoken in movies? Show a distribution containing all languages together with the amount of related movies and the title of the highest budgeted one.
9. How much **profit** (use the revenue and budget) did each production studio make? Show a distribution of all companies and the most profitable movie for each studio.
10. What are the most rated movies since 1970? Show a distribution containing the most rated movie for each year since 1970 calculate the profit for each of those. In which year was the most profitable one released?

For all(!) students:

i. Write down a second program that takes the IMDB_id and release_date as parameters. Identify each movie that has similar runtime to the provided one +- 10 minutes. The procedure shall start with movies from the given release_date. Each entry of the list shall include: id, runtime, original title and release date.

ii. Write down a third program that performs a count of the occurrence for each of three input terms within the field "overview". Such as for "me, agent, mafia". Those terms need to be provided as parameters. Each occurrence of the defined term(s) should be counted and the respected movie id tracked. The program should work with one, two and three terms. If no parameters are provided, instead a warning should be displayed. The final list should provide the movie id as well as the number of for each matched word (count).

In any case, if the genre needs to be considered in the respective question, multi-labeled movies such as “Drama|Crime|Music” can be counted for each genre separately. If a specific movie is required and nothing special requested take the id and title.