You are required to analyze the Movie Dataset using PySpark. The dataset can be downloaded on Moodle. These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset consists of the following files:

movies_metadata.csv: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

keywords.csv: Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.

credits.csv: Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

links.csv: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset. You may obtain more information by using a crawler program with the TMDB and IMDB IDs:

- imdbId is an identifier for movies used by http://www.imdb.com. E.g., the movie Toy Story has the link http://www.imdb.com/title/tt0114709/.
- tmdbId is an identifier for movies used by https://www.themoviedb.org. E.g., the movie Toy Story has the link https://www.themoviedb.org/movie/862.

ratings.csv: This file contains 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Questions and tasks:

1. Build regression models to predict movie revenue and vote averages based on a certain metric. (40 marks)
2. Analyze that what movies tend to get higher vote averages on TMDB. Try to use more figures with data visualization methods to illustrate your analysis. (20 marks)
3. Use collaborative filtering to build a movie recommendation system with two functions:
    a. Suggest top N movies similar to a given movie title (20 marks).

b. Predict user rating for the movies they have not rated for. You may use a test set to test your prediction accuracy, in which the test ratings can be regarded as not rated during training (20 marks).

Show all steps including data preprocessing, modeling, testing, evaluations with concise explanation in Markdown cell. You may also try different models and compare them in different ways with discussion. If your personal computer is not powerful enough to handle this project, you may try to use some public computation resources like Google Colab.