PySpark Assignments

Spark Core – RDD

1. From the given dataset of log-messages, find out how many times each type of error occurred. Consider only "error" messages and group the errors by the type. (ex; how may php errors, how may mysql errors etc.)
    - Dataset: server_log

2. Extend the Wordcount program to do the following: From a given data file count how many times each word occurs (i.e normal wordcount). Then filter out all words that occurs less than a given threshold. From among those filtered words, count how many times each letter occurs.
    - Dataset: You can use the same dataset you used for wordcount.

3. From the given Car details dataset, compute the "Average Weight" of "American Cars" by "Make". The output should be like: (ford, 3540), (buick, 2800) etc..
    - Dataset: cars.tsv

Spark SQL

1. From the given **movies.csv** and **ratings.csv** datasets find out the **top 20 movies with the highest average user ratings** from those movies **that received at least 10 user reviews**. Show the following data in the output: Movie Id, Movie Name, Average Rating & Total Number of Ratings.
    a. Datasets: movies.csv, ratings.csv

2. Do the same assignment as above but with the datasets that does not have header row. Apply your own schema to the datasets and do the same as above. (Save the above datasets as separate files after removing the header and use the saved files for this assignment)

3. Perform the same functionality mentioned in the assignment 1, using DataFrame API alone. (Do not use SQL). Store the output into a PARQUET file format. Now load the data from the stored PARQUET file into a dataframe and store the contents of this dataframe into a JSON file.