ASSIGNMENT 2

The file available for this assignment reviews.txt contains data of movie reviews. Perform following tasks using Spark with Java.

1. Read the data from file reviews.txt and store into a Java RDD in Spark. Print the count of records in the RDD. Also print the number of partitions present in the RDD.
2. Remove all of the following special characters from the records [$&+,:;=?@#|'<>.-^*()%!]. Use these cleansed records for further processing.
3. Print the number of records that do not contain any numeric character.
4. Print total count of the occurrence of word ‘movie’ in all records.
5. Print the minimum and the maximum length of the review.
6. Write the data of cleansed records from task 2 in a file named reviews_cleansed.txt.
7. Note down the Turn Around Time (TAT) for each of the above operations.
8. Take the screenshot of the DAG generated by Spark by browsing Spark UI.
9. Thoroughly understand the jobs, stages and tasks created by Spark after exploring the Spark UI. Appropriately answer the questions regarding these during the assignment presentation.