## -------------------------------------------------------------------------------------------------------

## CS5542 - Lab Assignment 1 – Lab Report

## Sub team # 1-2 | Avni Mehta (Class Id: 11) | Raji Muppala (Class Id: 14)

## -------------------------------------------------------------------------------------------------------

**Objective:**

For Movielens dataset, write Spark Transformations and Actions to find the users who have rated more than 25 items.

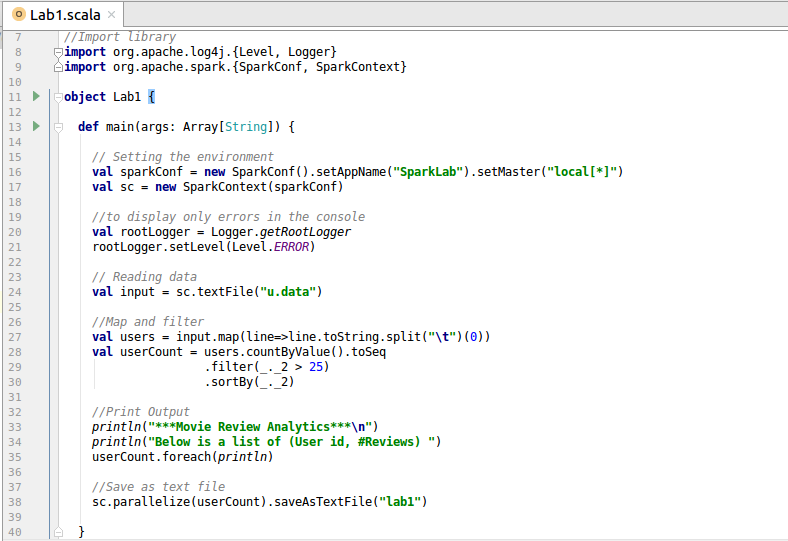
**Input:**

The input for this task is movielens dataset having 100k records of tab-separated list of user id | item id | rating | timestamp.

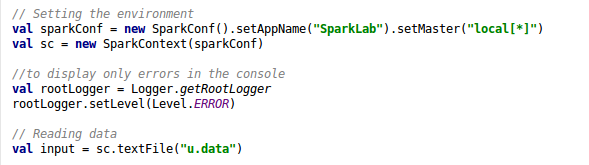
**Code:**

Both team members came up with different ways to achieve the goal.

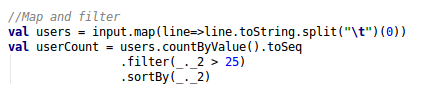
One way:



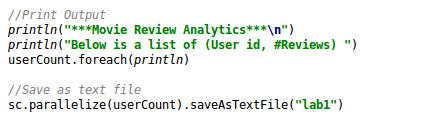
After importing the libraries, we set the spark configuration and spark context. Data from text file is read and stored into a variable.



A RDD of user ids is obtained by mapping the input after splitting by ‘\t’ and getting the first element in the row. The user ids RDD is transformed into a Map of (userids, count of reviews) by countByValue action. The output is filtered where count > 25 and the result is sorted by asc order.

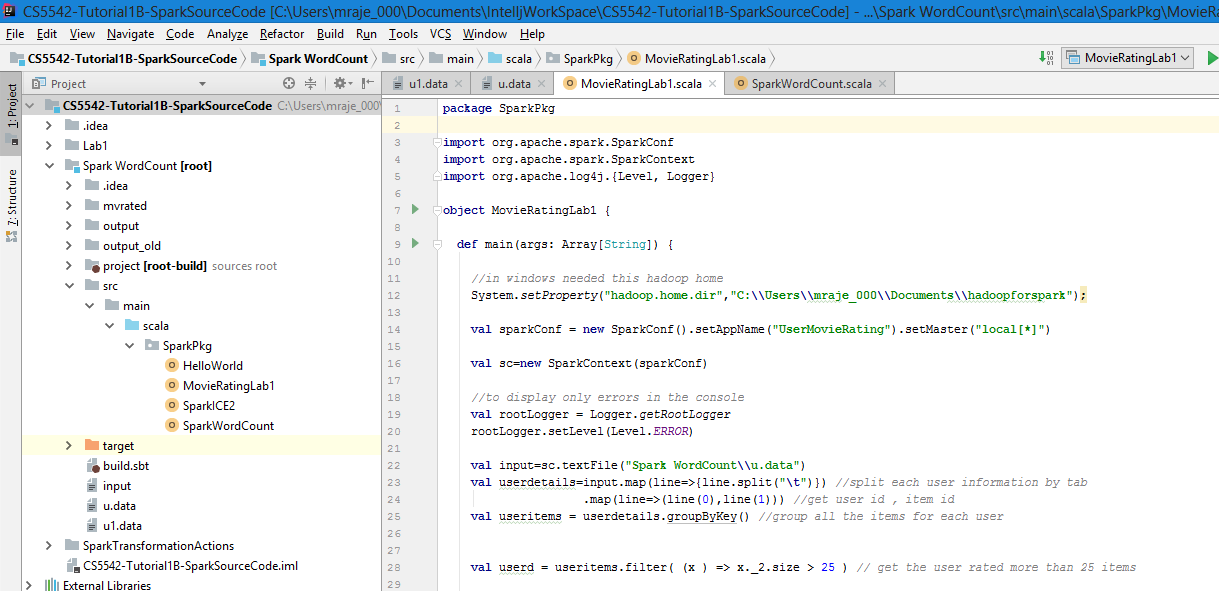


The output is printed on the console and is also stored into a text file.

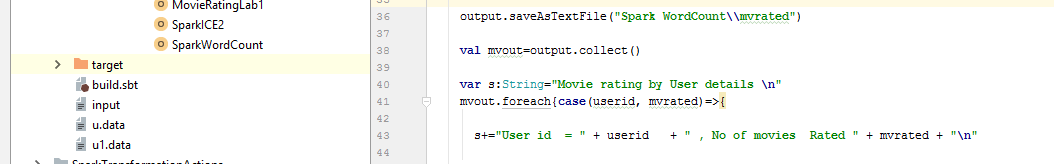


Another way:

In the below code: Reading file, splitting each line data for ‘\t’. Applying map transformation to generate RDD with userid and item details. Applying groupByKey transformation to group data by userid. Then filtering users where items size is greater than 25. Creating new RDD with just user id and no. of movies rated.



Lastly saving data to text file. Collecting all the data and printing on the console.



**Output:**

Output is printed on the console as well as stored as a text file.

