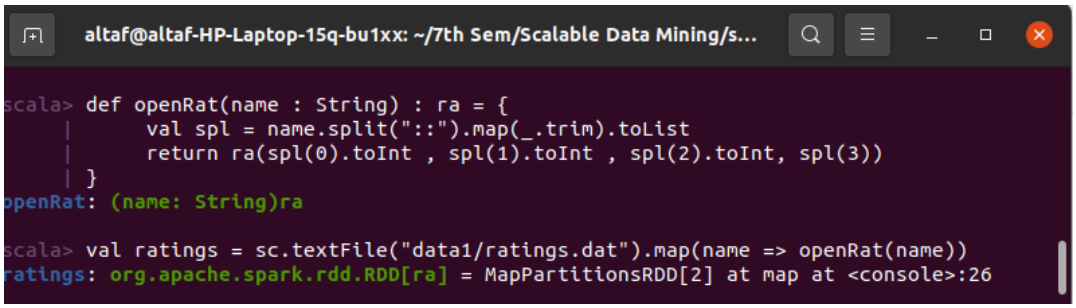
Scalable Data Mining (Autumn 2021)
Assignment 1

Name : Altaf Ahmad
Roll no: 18MA20005

Question 1

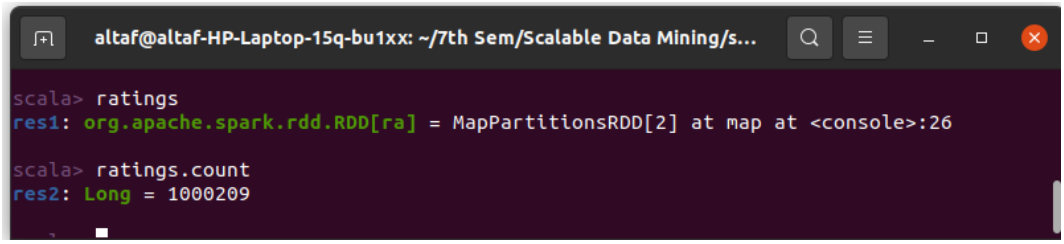
(a) Download the ratings file, parse it and load it in an RDD named ratings.

Ans: First we download the given file and then parse it by separating the elements from the format UserID::MovieID::Rating::Timestamp , using split with "::"



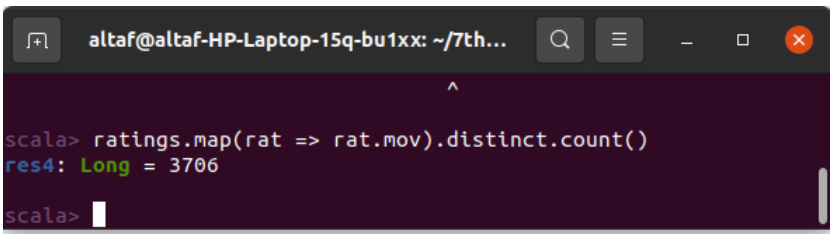
(b)How many lines does the ratings RDD contain?

Ans : We found that the number of lines of RDD is 1000209



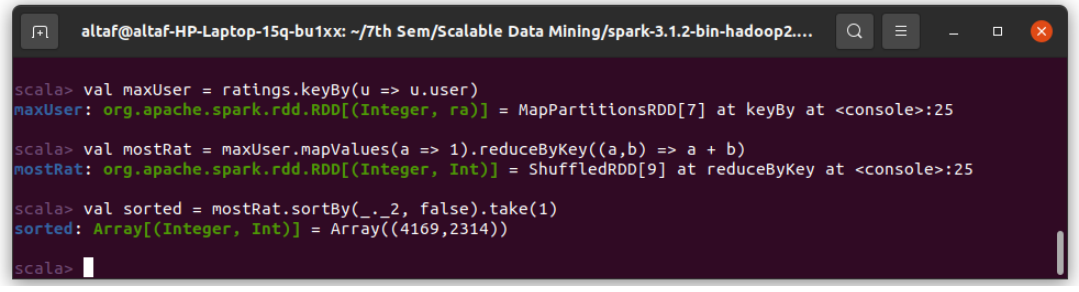
(c) Count how many unique movies have been rated.

Ans : We will have to count the number of distinct movies. We can create the map and count of that using distinct. The number of unique movies is 3706



(d) Which user gave the most ratings? Return the userID and number of ratings.

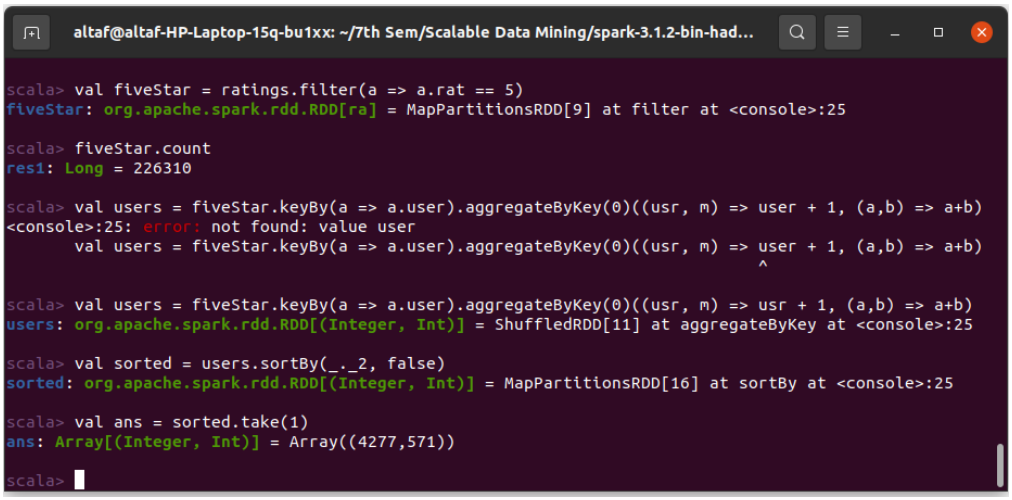
Ans : In order to find the user who gave the most ratings, we first calculcate the number of ratings of each user and then sort it in descending order. After this we can find out the maximum rating user along with the number of ratings.



The userID is 4169 and the number of ratings is 2314

(e) Which user gave the most '5' ratings? Return the userID and number of ratings.

Ans : Similarly in this question, we count the number of '5' star ratings given and then sort it in desceding order. After that we can find the userID and the number of ratings.

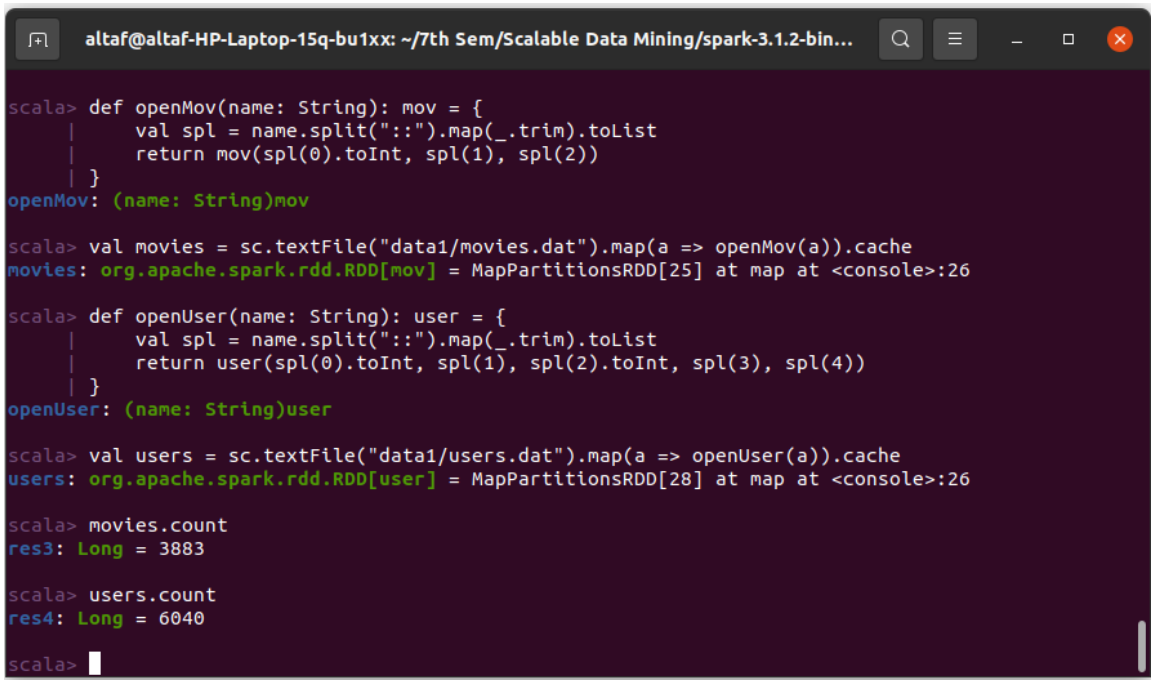


userID = 4277 and number of ratings = 571

Question 2

(a) Read the movies and users files into RDDs. How many records are there in each RDD?

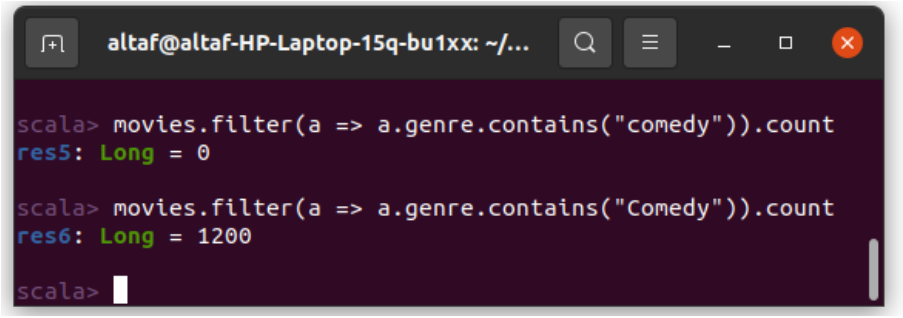
Ans :



movies.dat has 3883 records and users.dat has 6040 records

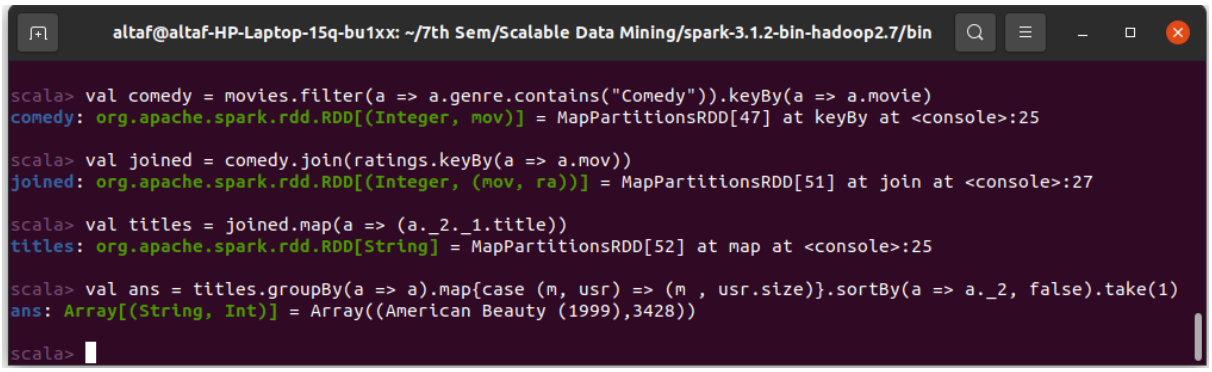
(b) How many of the movies are a comedy?

Ans : There are a total of 1200 movies with genre "Comedy"



(c) Which comedy has the most ratings? Return the title and the number of rankings. Answer this question by joining two datasets.

Ans : For solving this, we will have to join the movies RDD with the ratings RDD and then count the number of movies with genre "Comedy". Sort this in decreasing order and then take the first value,



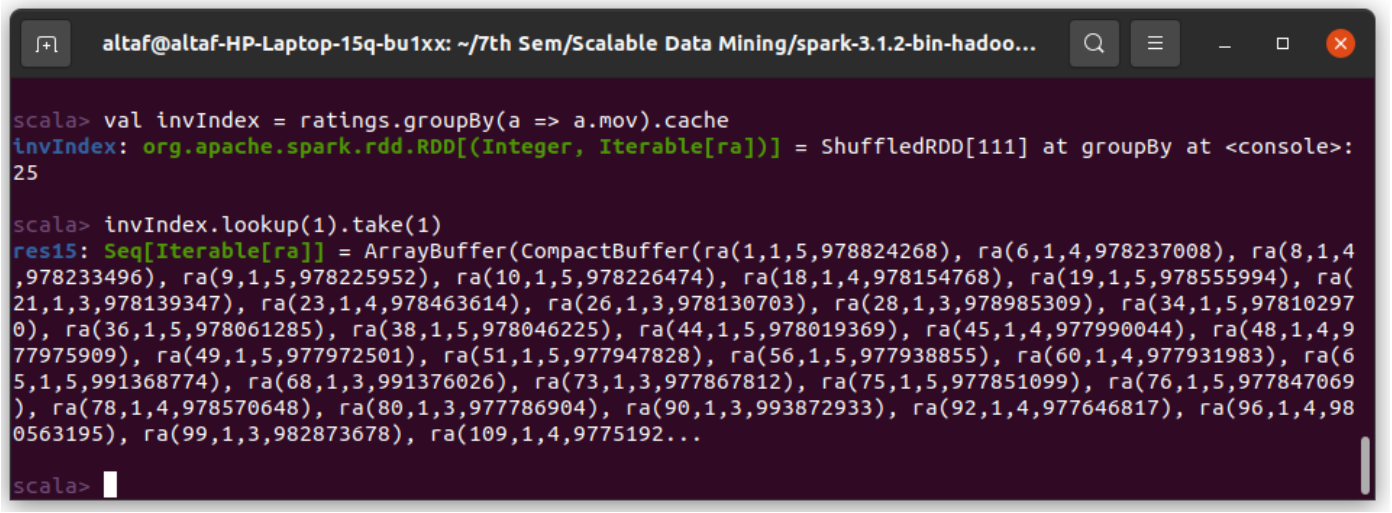
The most rated comedy is American Beauty (1999) with number of rankings -> 3428

(e) Compute the number of unique users that rated the movies with movie_IDs 2858, 356 and 2329 without using an inverted index. Measure the time (in seconds) it takes to make this computation.

Ans : The number of unique users is 4213

It takes around 0.3 - 0.38 seconds to make this computation

(f) Create an inverted index on ratings, field movie_ID. Print the first item.



(g) Compute the number of unique users that rated the movies with movie_IDs 2858, 356 and 2329 using the above calculated index. Measure the time (in seconds) it takes to compute the same result using the index.

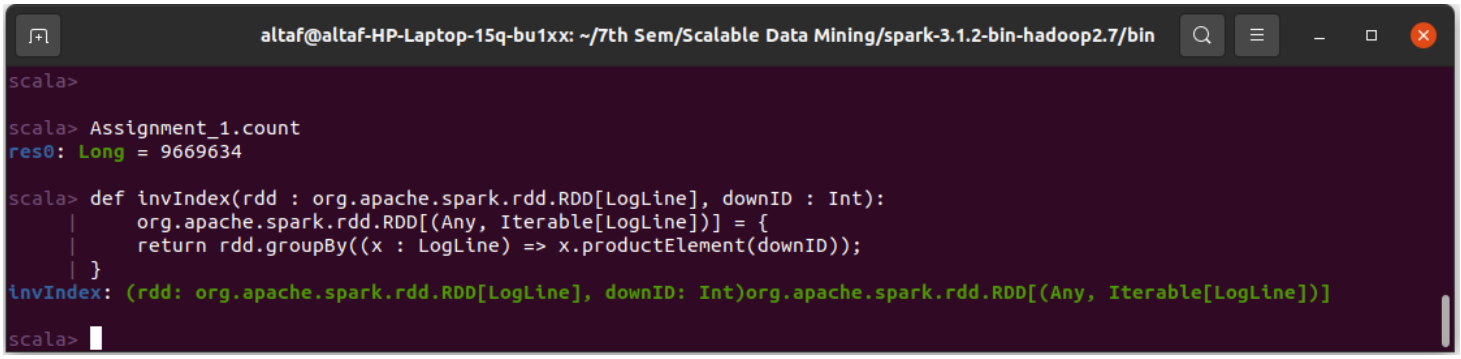
Ans:

The number of unique users is still the same but the time taken is approximately 7.9 secs

---

Question 3

(a). Create a function that given an RDD and a field (e.g. download_id), it computes an inverted index on the RDD for efficiently searching the records of the RDD using values of the field as keys.



(b) Compute the number of different repositories accessed by the client ‘ghtorrent-22’ (without using the inverted index).