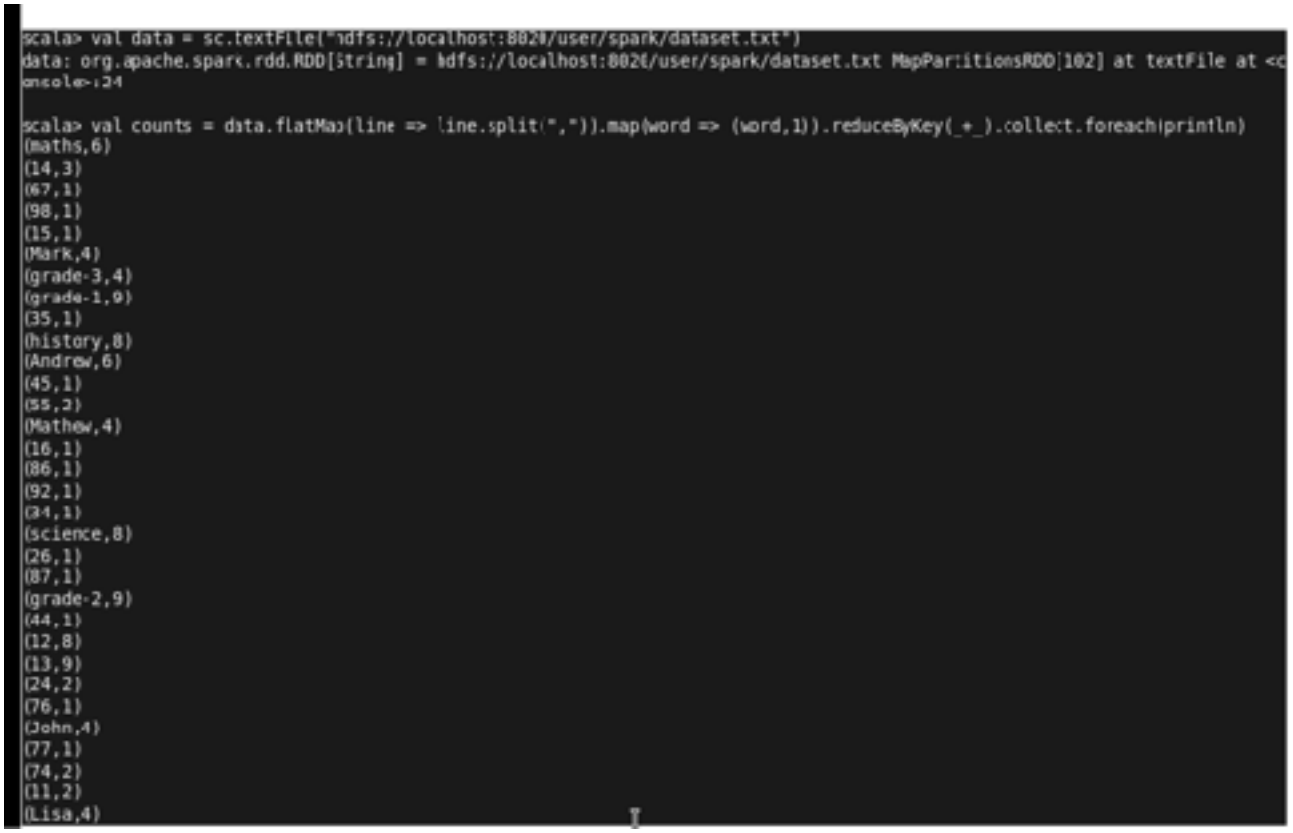
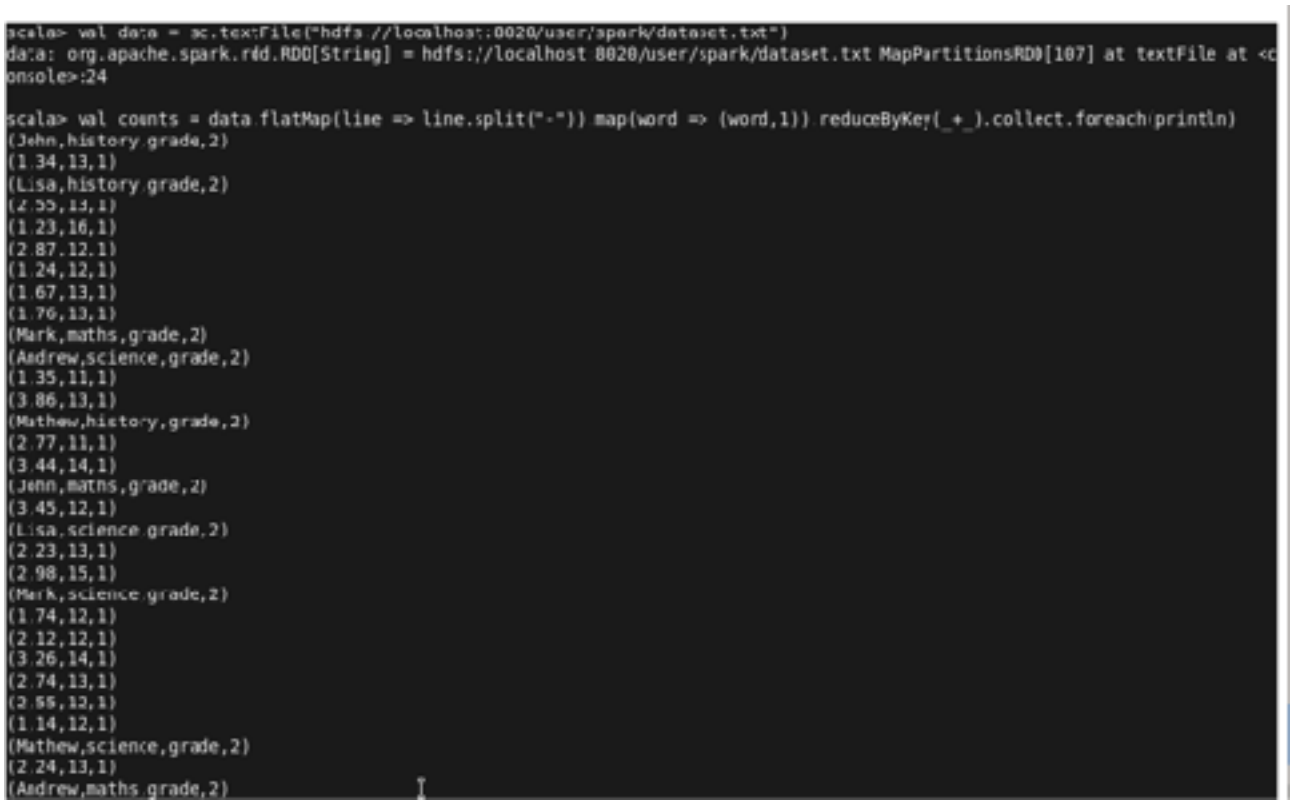
Task 1

1.Write a program to read a text file and print the number of rows of data in the document.

2. Write a program to read a text file and print the number of words in the document.



3. We have a document where the word separator is -, instead of space. Write a spark code, to obtain the count of the total number of words present in the document.



Task 2

Problem Statement 1:

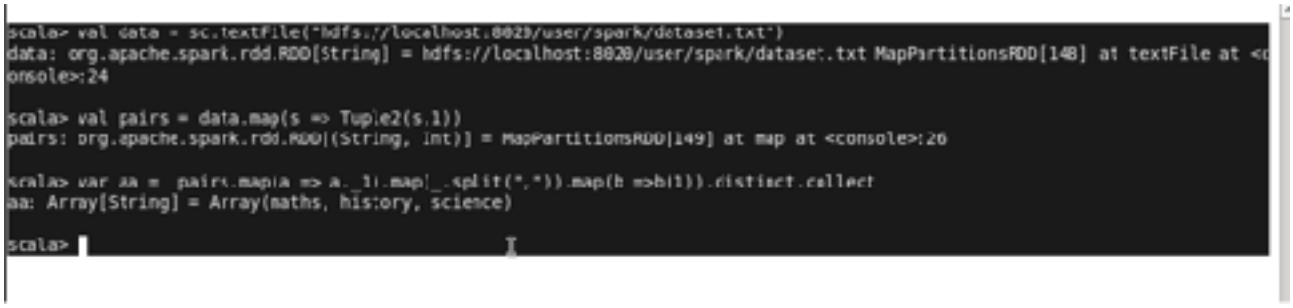
1. Read the text file, and create a tupled rdd.

```
scala> val data = sc.textFile("hdfs://localhost:8020/user/spark/dataset.txt")
data: org.apache.spark.rdd.RDD[String] = hdfs://localhost:8020/user/spark/dataset.txt MapPartitionsRDD[121] at textFile at <c
onsole>:24

scala> val pairs = data.map(s => Tuple2(s,1))
pairs: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[122] at map at <console>:26

scala> pairs.collect().foreach(println)
(Mathew,science,grade-3,45,12,1)
(Mathew,history,grade-2,55,13,1)
(Mark,maths,grade-2,23,13,1)
(Mark,science,grade-1,75,13,1)
(John,history,grade-1,14,12,1)
(John,maths,grade-2,74,13,1)
(Lisa,science,grade-1,24,12,1)
(Lisa,history,grade-3,85,13,1)
(Andrew,maths,grade-1,34,13,1)
(Andrew,science,grade-3,26,14,1)
(Andrew,history,grade-1,74,12,1)
(Mathew,science,grade-2,55,12,1)
(Mathew,history,grade-2,87,12,1)
(Mark,maths,grade-1,92,13,1)
(Mark,science,grade-2,12,12,1)
(John,history,grade-1,67,13,1)
(John,maths,grade-1,35,11,1)
(Lisa,science,grade-2,24,13,1)
(Lisa,history,grade-2,93,15,1)
(Andrew,maths,grade-1,23,16,1)
(Andrew,science,grade-3,44,14,1)
(Andrew,history,grade-2,77,11,1)
```

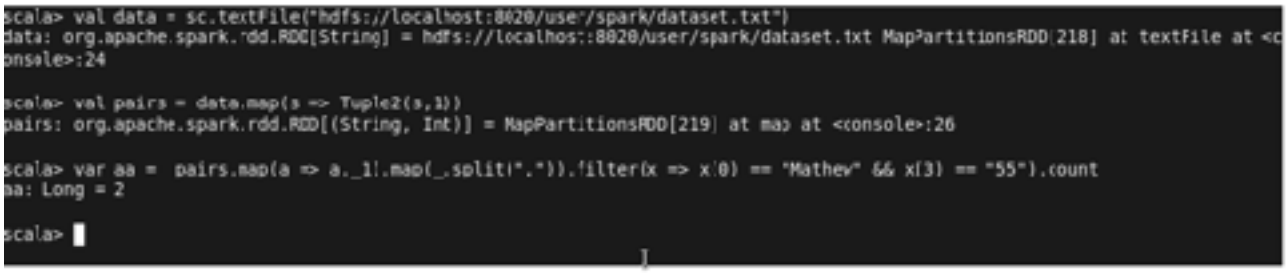
2. Find the count of total number of rows present.

```
scala> val data = sc.textFile("hdfs://localhost:8020/user/spark/dataset.txt")
data: org.apache.spark.rdd.RDD[String] = hdfs://localhost:8020/user/spark/dataset.txt MapPartitionsRDD[133] at textFile at <c
onsole>:24

scala> val pairs = data.map(s => Tuple2(s,1))
pairs: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[133] at map at <console>:26

scala> val counts = pairs.reduceByKey((a,b) => a + b)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[134] at reduceByKey at <console>:28

scala> counts.count
res36: Long = 22

scala>
```

3. What is the distinct number of subjects present in the entire school

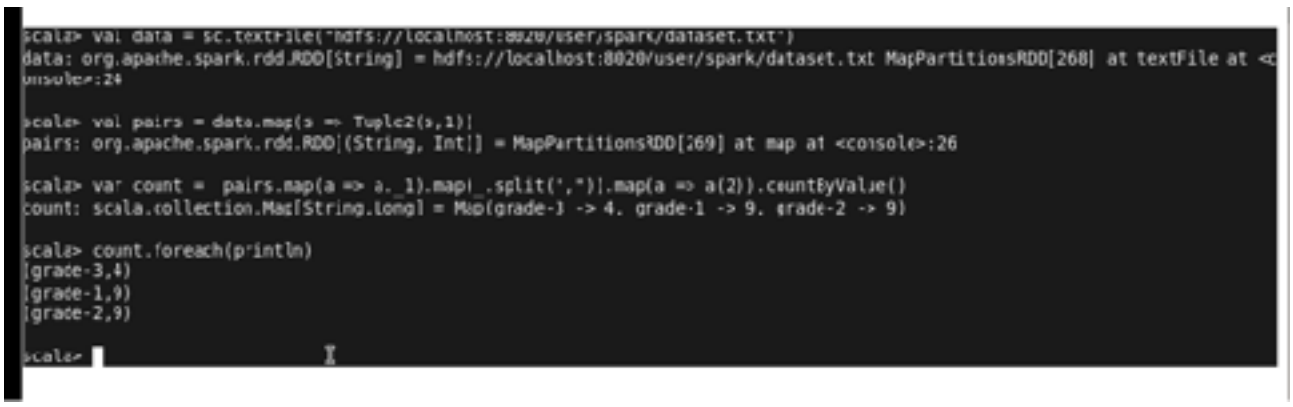


4. What is the count of the number of students in the school, whose name is Mathew and marks is 55



Problem Statement 2:

1. What is the count of students per grade in the school?



2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)

3. What is the average score of students in each subject across all grades?