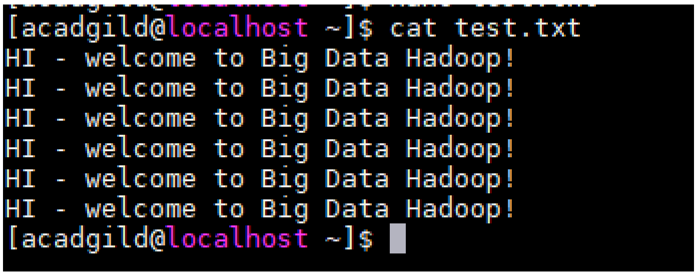
Assignment - 19.1

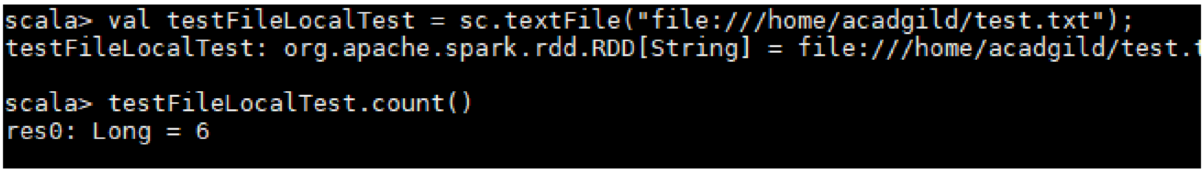
Task1.1: Write a program to read a text file and print the number of rows of data in the document.

Answer: Now first we create a file in local and note the no. of records over here to be 6 rows of records as shown in the below screenshot.



```
[acadgild@localhost ~]$ cat test.txt
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
[acadgild@localhost ~]$
```

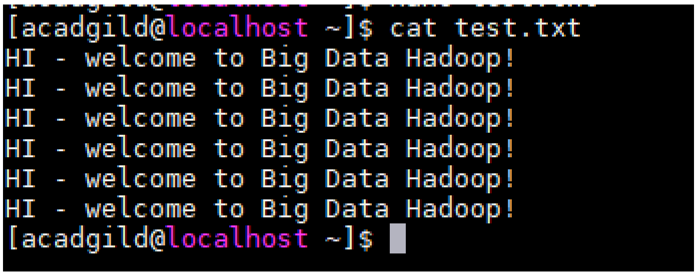
Now we will load this file to spark and get the no. of rows of record using the below commands as shown in below screenshot.



```
scala> val testFileLocalTest = sc.textFile("file:///home/acadgild/test.txt");
testFileLocalTest: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/test.t

scala> testFileLocalTest.count()
res0: Long = 6
```

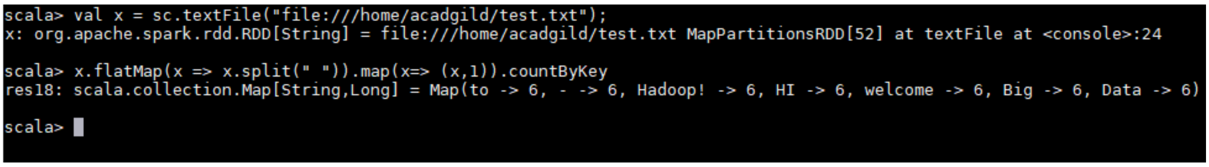
Task1.2: Write a program to read a text file and print the number of words in the document.

Answer: Input file content is as below.



```
[acadgild@localhost ~]$ cat test.txt
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
HI - welcome to Big Data Hadoop!
[acadgild@localhost ~]$
```

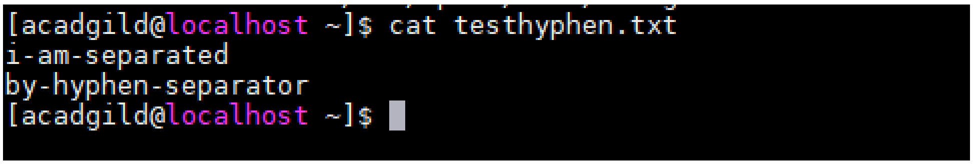
Wordcount is as mentioned below.



```
scala> val x = sc.textFile("file:///home/acadgild/test.txt");
x: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/test.txt MapPartitionsRDD[52] at textFile at <console>:24

scala> x.flatMap(x => x.split(" ")).map(x=> (x,1)).countByKey
res18: scala.collection.Map[String,Long] = Map(to -> 6, - -> 6, Hadoop! -> 6, HI -> 6, welcome -> 6, Big -> 6, Data -> 6)

scala>
```

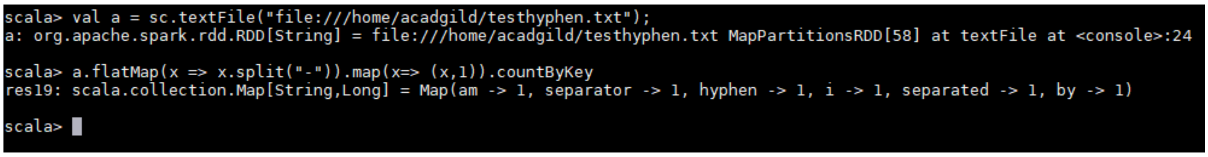
Task1.3: Write a program to read a text file and print the number of words in the document. Write a spark code, to obtain the count of the total number of words present in the document.

Answer: input file we can see that the words are separated by hyphen



```
[acadgild@localhost ~]$ cat testhyphen.txt
i-am-separated
by-hyphen-separator
[acadgild@localhost ~]$ 
```

Word count is as follows.

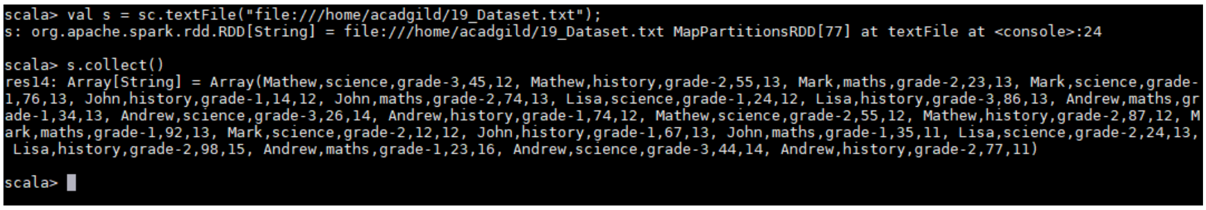


```
scala> val a = sc.textFile("file:///home/acadgild/testhyphen.txt");
a: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/testhyphen.txt MapPartitionsRDD[58] at textFile at <console>:24

scala> a.flatMap(x => x.split("-")).map(x=> (x,1)).countByKey
res19: scala.collection.Map[String,Long] = Map(am -> 1, separator -> 1, hyphen -> 1, i -> 1, separated -> 1, by -> 1)

scala> 
```

Task 2

Problem1.1: Read the text file, and create a tupled rdd.

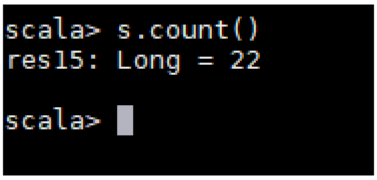
Answer:



```
scala> val s = sc.textFile("file:///home/acadgild/19_Dataset.txt");
s: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/19_Dataset.txt MapPartitionsRDD[77] at textFile at <console>:24

scala> s.collect()
res14: Array[String] = Array(Mathew,science,grade-3,45,12, Mathew,history,grade-2,55,13, Mark,maths,grade-2,23,13, Mark,science,grade-
1,76,13, John,history,grade-1,14,12, John,maths,grade-2,74,13, Lisa,science,grade-1,24,12, Lisa,history,grade-3,86,13, Andrew,maths,gr
ade-1,34,13, Andrew,science,grade-3,26,14, Andrew,history,grade-1,74,12, Mathew,science,grade-2,55,12, Mathew,history,grade-2,87,12, M
ark,maths,grade-1,92,13, Mark,science,grade-2,12,12, John,history,grade-1,67,13, John,maths,grade-1,35,11, Lisa,science,grade-2,24,13,
 Lisa,history,grade-2,98,15, Andrew,maths,grade-1,23,16, Andrew,science,grade-3,44,14, Andrew,history,grade-2,77,11)

scala> 
```

Problem1.2: Find the count of total number of rows present.

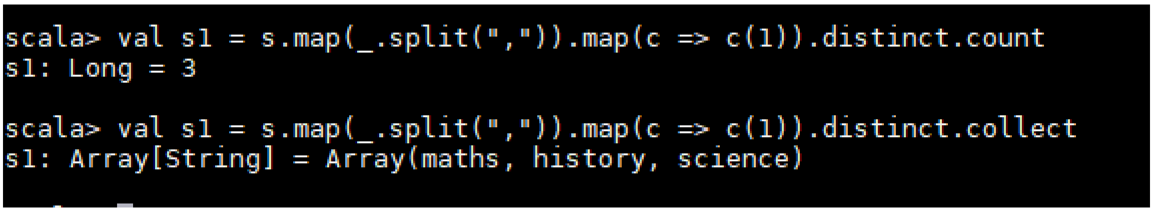
Answer:



```
scala> s.count()
res15: Long = 22

scala> 
```

Problem1.3: What is the distinct number of subjects present in the entire school?

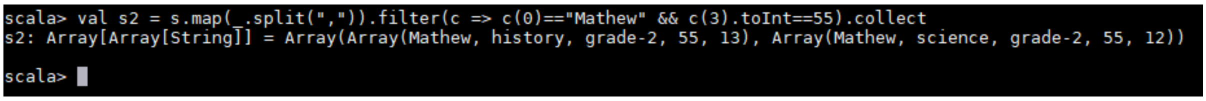
Answer:



```
scala> val s1 = s.map(_.split(",")).map(c => c(1)).distinct.count
s1: Long = 3

scala> val s1 = s.map(_.split(",")).map(c => c(1)).distinct.collect
s1: Array[String] = Array(maths, history, science)
```

Problem1.4: What is the count of the number of students in the school, whose name is Mathew and marks is 55?

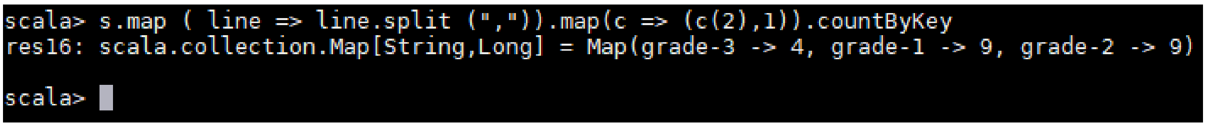
Answer:



```
scala> val s2 = s.map(_.split(",")).filter(c => c(0)=="Mathew" && c(3).toInt==55).collect
s2: Array[Array[String]] = Array(Array(Mathew, history, grade-2, 55, 13), Array(Mathew, science, grade-2, 55, 12))

scala> 
```

Problem2.1: What is the count of students per grade in the school?

Answer:



```
scala> s.map ( line => line.split (",")).map(c => (c(2),1)).countByKey
res16: scala.collection.Map[String,Long] = Map(grade-3 -> 4, grade-1 -> 9, grade-2 -> 9)

scala> 
```