Assignment 17.1

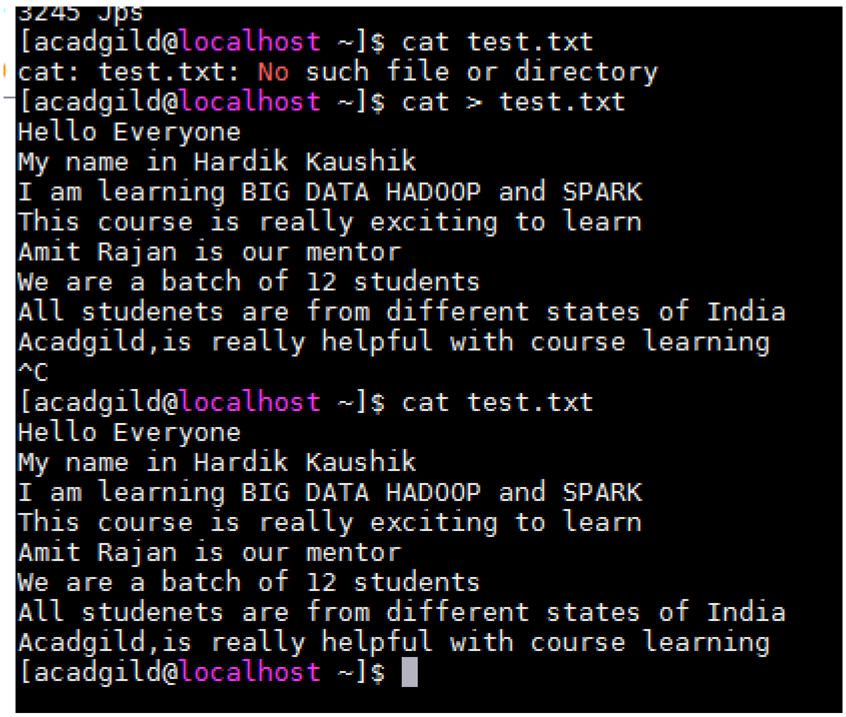
Problem Statement :-

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.

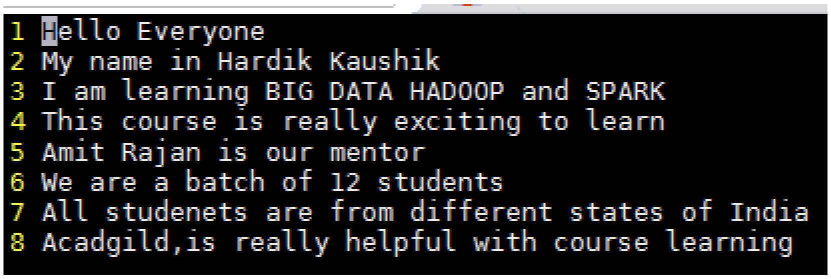
Solution:-

- **Program to read a text file and print the number of rows of data in the document**

Firstly,we create a test text file as follows:-



```
3245 Jps
[acadgild@localhost ~]$ cat test.txt
cat: test.txt: No such file or directory
[acadgild@localhost ~]$ cat > test.txt
Hello Everyone
My name in Hardik Kaushik
I am learning BIG DATA HADOOP and SPARK
This course is really exciting to learn
Amit Rajan is our mentor
We are a batch of 12 students
All studenets are from different states of India
Acadgild,is really helpful with course learning
^C
[acadgild@localhost ~]$ cat test.txt
Hello Everyone
My name in Hardik Kaushik
I am learning BIG DATA HADOOP and SPARK
This course is really exciting to learn
Amit Rajan is our mentor
We are a batch of 12 students
All studenets are from different states of India
Acadgild,is really helpful with course learning
[acadgild@localhost ~]$ 
```

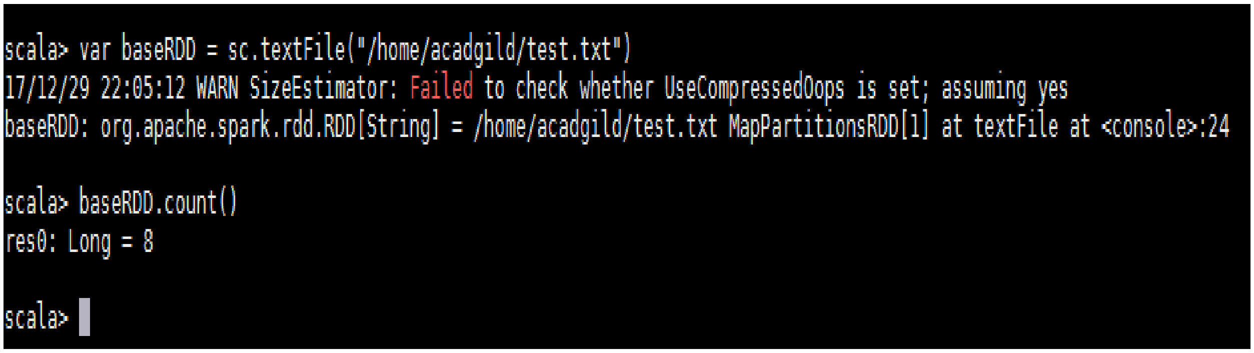


```
1 Hello Everyone
2 My name in Hardik Kaushik
3 I am learning BIG DATA HADOOP and SPARK
4 This course is really exciting to learn
5 Amit Rajan is our mentor
6 We are a batch of 12 students
7 All studenets are from different states of India
8 Acadgild,is really helpful with course learning
```

Above shows that that test.txt file contains 8 rows of data,Now we write a spark program for counting the number of rows of data in a file which as follows:-

var baseRDD = sc.textFile("/home/acadgild/test.txt")

baseRDD.count()

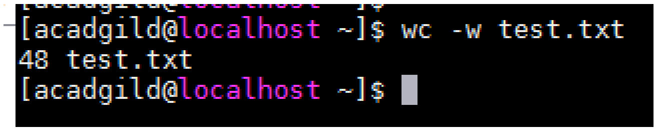


```
scala> var baseRDD = sc.textFile("/home/acadgild/test.txt")
17/12/29 22:05:12 WARN SizeEstimator: Failed to check whether UseCompressedOops is set; assuming yes
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/test.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> baseRDD.count()
res0: Long = 8

scala>
```

- **Program to read a text file and print the number of words in the document**



```
[acadgild@localhost ~]$ wc -w test.txt
48 test.txt
[acadgild@localhost ~]$
```

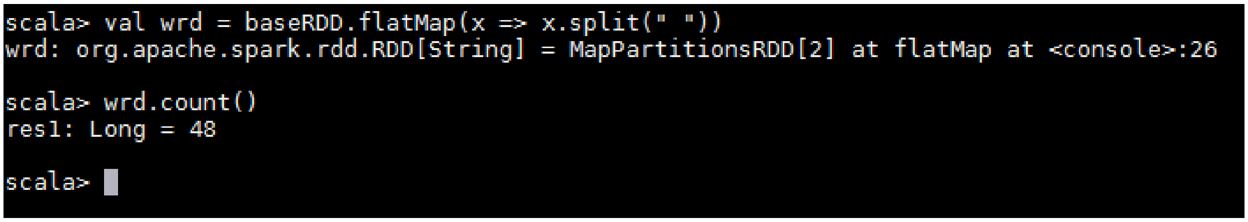
Above shows that that test.txt file contains 48 words,Now we write a spark program for counting the number of words in a file which as follows:-

var baseRDD = sc.textFile("/home/acadgild/test.txt")

Command used to count the number of words in the text file:

val wrd = baseRDD.flatMap(x => x.split(" "))

wrd.count()



```
scala> val wrd = baseRDD.flatMap(x => x.split(" "))
wrd: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:26

scala> wrd.count()
res1: Long = 48

scala> 
```

- **the count of the total number of words present in the document provided**

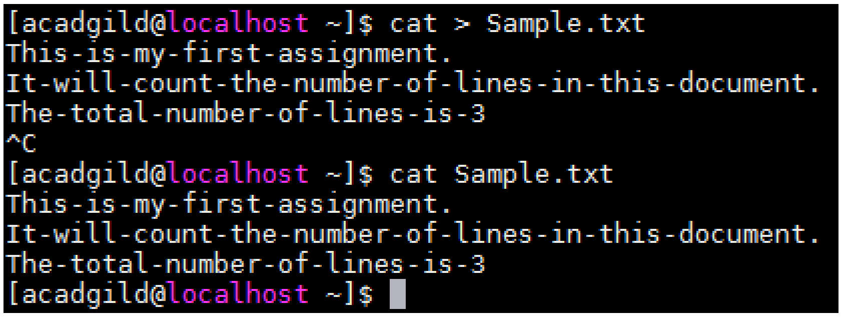
Sample Document:-

This-is-my-first-assignment.

It-will-count-the-number-of-lines-in-this-document.

The-total-number-of-lines-is-3

Created the Sample.txt file



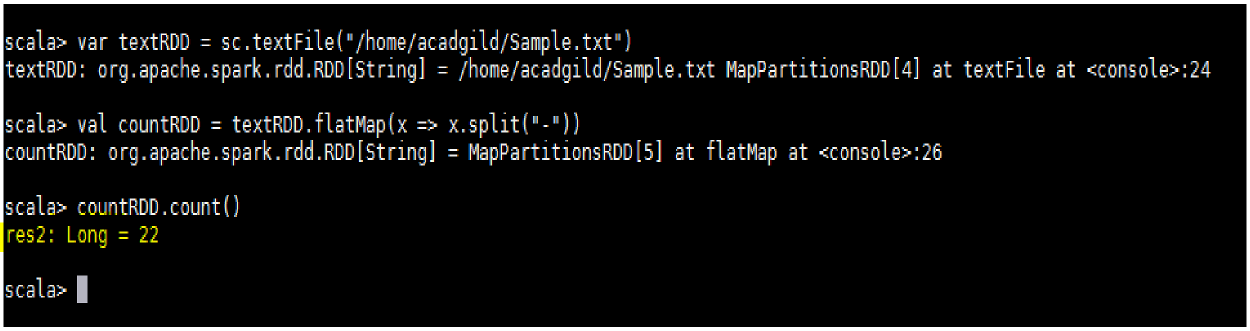
```
[acadgild@localhost ~]$ cat > Sample.txt
This-is-my-first-assignment.
It-will-count-the-number-of-lines-in-this-document.
The-total-number-of-lines-is-3
^C
[acadgild@localhost ~]$ cat Sample.txt
This-is-my-first-assignment.
It-will-count-the-number-of-lines-in-this-document.
The-total-number-of-lines-is-3
[acadgild@localhost ~]$ 
```

Spark Program for counting the number of words present in the above doucment is as follows:-

var textRDD = sc.textFile("/home/acadgild/Sample.txt")

val countRDD = textRDD.flatMap(x => x.split("-"))

countRDD.count()



```
scala> var textRDD = sc.textFile("/home/acadgild/Sample.txt")
textRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/Sample.txt MapPartitionsRDD[4] at textFile at <console>:24

scala> val countRDD = textRDD.flatMap(x => x.split("-"))
countRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at flatMap at <console>:26

scala> countRDD.count()
res2: Long = 22

scala> 
```

Submitted By:-

Hardik Kaushik