Accadgild_Session_17_Assignment_17.1_Solution

## 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.

Sample document :

This-is-my-first-assignment.

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

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

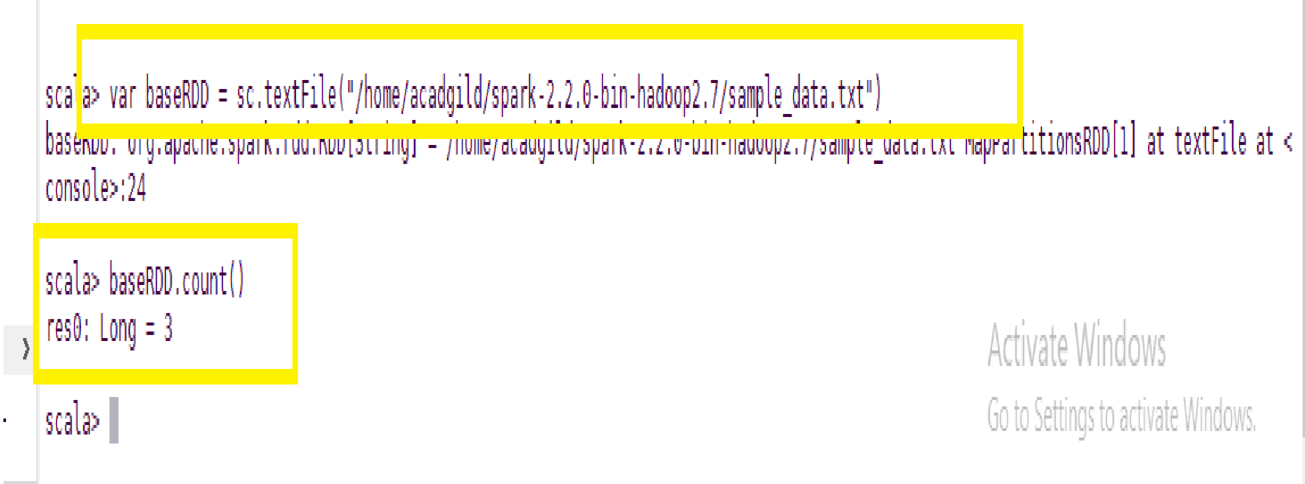
### Question 1 Solution:

Code:

```
var baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt")

baseRDD.count()
```

Screen-Shot



```
scala> var baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt")
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt MapPartitionsRDD[1] at textFile at <
console>:24

scala> baseRDD.count()
res0: Long = 3

scala>
```

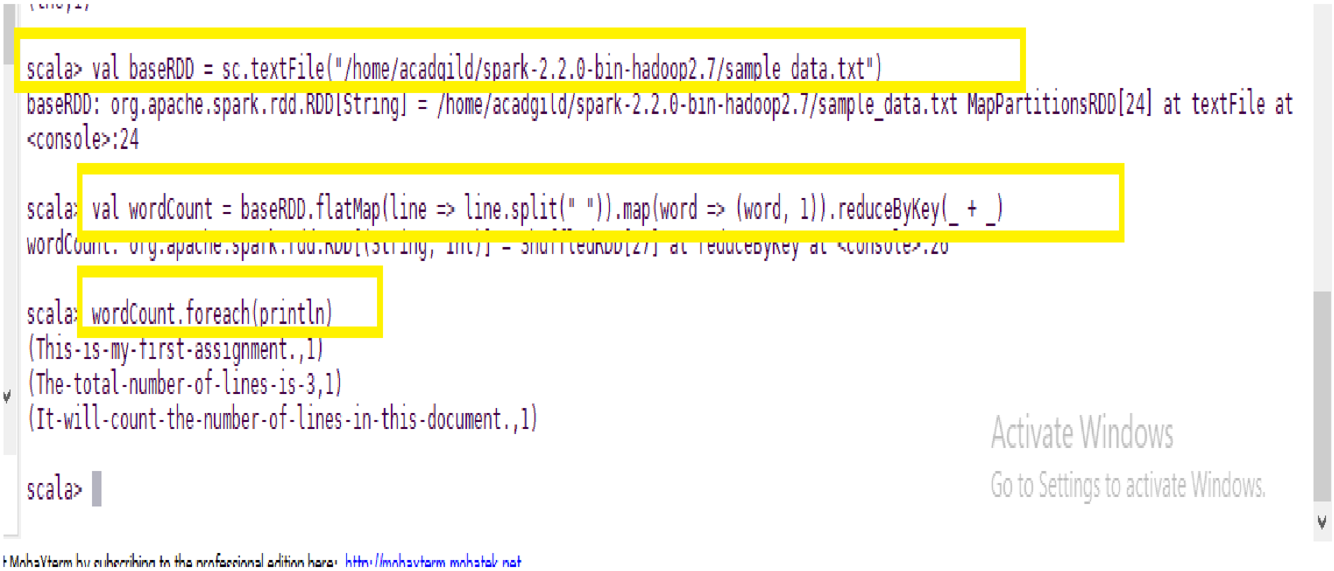
Question 2 Solution:

Code:

val baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt")

val wordCount = baseRDD.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)

wordCount.foreach(println)



```
scala> val baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample data.txt")
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt MapPartitionsRDD[24] at textFile at
<console>:24

scala> val wordCount = baseRDD.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
wordCount: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[27] at reduceByKey at <console>:26

scala> wordCount.foreach(println)
(This-is-my-first-assignment.,1)
(The-total-number-of-lines-is-3,1)
(It-will-count-the-number-of-lines-in-this-document.,1)

scala>
```

t MobaXterm by subscribing to the professional edition here: http://mobaxterm.mobatek.net

Question 3 Solution:

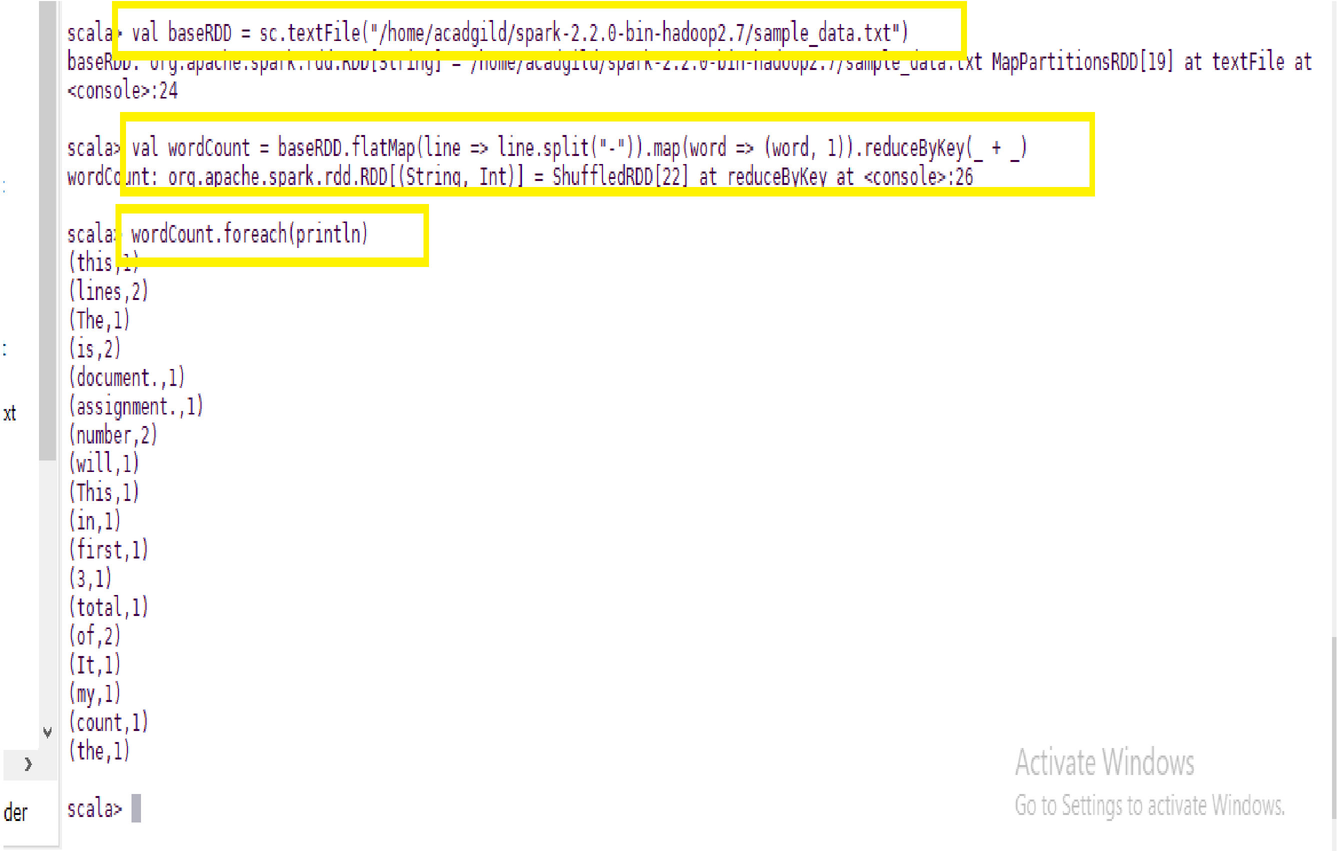
Code:

val baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt")

val wordCount = baseRDD.flatMap(line => line.split("-")).map(word => (word, 1)).reduceByKey(_ + _)

wordCount.foreach(println)

Accadgild_Session_17_Assignment_17.1_Solution



```
scala> val baseRDD = sc.textFile("/home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt")
baseRDD: org.apache.spark.rdd.RDD[String] = /home/acadgild/spark-2.2.0-bin-hadoop2.7/sample_data.txt MapPartitionsRDD[19] at textFile at
<console>:24

scala> val wordCount = baseRDD.flatMap(line => line.split("-")).map(word => (word, 1)).reduceByKey(_ + _)
wordCount: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[22] at reduceByKey at <console>:26

scala> wordCount.foreach(println)
(this,1)
(lines,2)
(The,1)
(is,2)
(document.,1)
(assignment.,1)
(number,2)
(will,1)
(This,1)
(in,1)
(first,1)
(3,1)
(total,1)
(of,2)
(It,1)
(my,1)
(count,1)
(the,1)

scala>
```

Submitted By

Shishir Jha