Assignment_45.1_Task3

Name : Silvi Dheer

Course: Big Data & Spark

Task 3

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

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

```
scala> val textFile = sc.textFile("file:///home/acadgild/Desktop/SampleDoc")
textFile: org.apache.spark.rdd.RDD[String] = file:///home/acadgild/Desktop/
SampleDoc MapPartitionsRDD[1] at textFile at <console>:24

scala> textFile.count
res0: Long =
3
```

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

```
scala> val lines = textFile.flatMap(line => line.split(" "))
lines: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at
<console>:25

scala> lines.collect
res3: Array[String] = Array(This-is-my-first-assignment., It-will-count-the-
number-of-lines-in-this-document., The-total-number-of-lines-is-3)

scala> val numberOfWords = lines.count
numberOfWords: Long = 3

scala> numberOfWords
res5: Long = 3
```

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.

```
scala> val words = textFile.flatMap(line => line.split("-"))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[3] at flatMap at
<console>:25

scala> words.collect
res6: Array[String] = Array(This, is, my, first, assignment., It, will,
count, the, number, of, lines, in, this, document., The, total, number, of,
lines, is, 3)

scala> val no_ofWordWithSeparator = words.count
no_ofWordWithSeparator: Long = 22

scala> no_ofWordWithSeparator
res7: Long = 22
```