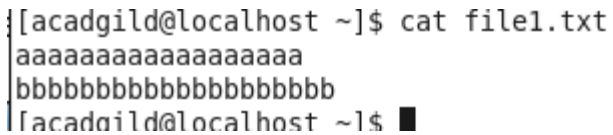
ASSIGNMENT17.1

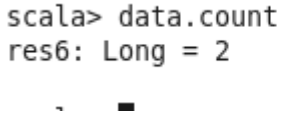
1 Write a program to read a text file and print the number of rows of data in the document.

```
scala> val inputfile="file:///home/acadgild/file1.txt"
inputfile: String = file:///home/acadgild/file1.txt

scala> val data=sc.textFile(inputfile)
data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at textFile at <console>:29

scala> data.count
res6: Long = 2
```

Screen shot



Output



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

```
scala> val inputfile="file:///home/acadgild/file1.txt"
inputfile: String = file:///home/acadgild/file1.txt

scala> val data=sc.textFile(inputfile)
data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[5] at textFile at <console>:29


scala> val words=data.flatMap(_.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at flatMap at <console>:31
```

scala> words.collect
res15: Array[String] = Array(save, open, undo, view, tools, system, help, plain, text)

scala> words.count
res16: Long = 9

Output Screen shots

```
[acadgild@localhost ~]$ cat file1.txt
save open undo view tools system help plain text
[acadgild@localhost ~]$
```

```
scala> words.collect
res15: Array[String] = Array(save, open, undo, view, tools, system, help, plain, text)

scala> words.count
res16: Long = 9

scala>
```

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

scala> val inputfile="file:///home/acadgild/text"
inputfile: String = file:///home/acadgild/text

scala> val data=sc.textFile(inputfile)
data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[34] at textFile at <console>:29

scala> val words = data.flatMap(_.split("-"))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[35] at flatMap at <console>:31

scala> words.count
res39: Long = 22

scala> val wordCounts = words.map(x => (x, 1)).reduceByKey(_ + _)
wordCounts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[37] at reduceByKey at

<console>:33

scala> wordCounts.collect
res36: Array[(String, Int)] = Array((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))

Output for above sample document

```
scala> words.count
res39: Long = 22

scala> wordCounts.collect
res40: Array[(String, Int)] = Array((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>
```