Assignment 17.1

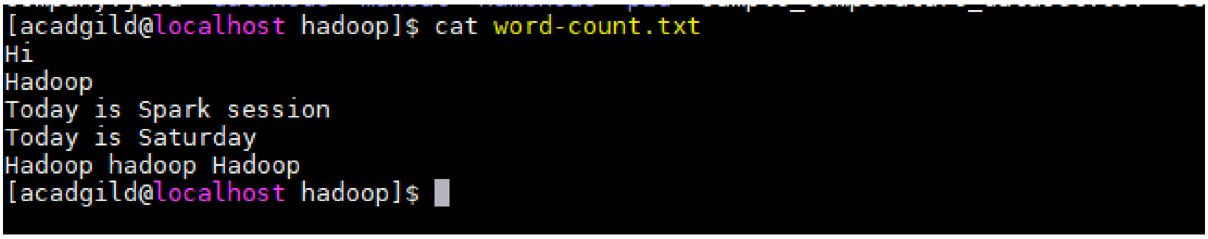
Problem Statement:

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

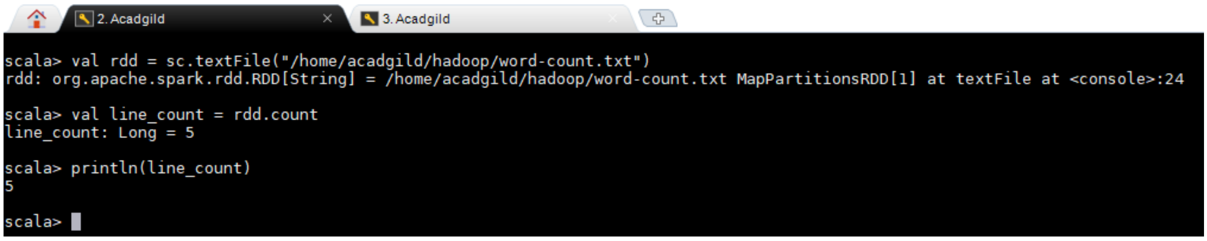
   val rdd = sc.textFile("/home/acadgild/hadoop/word-count.txt")

   val line_count = rdd.count

   println(line_count)



```
[acadgild@localhost hadoop]$ cat word-count.txt
Hi
Hadoop
Today is Spark session
Today is Saturday
Hadoop hadoop Hadoop
[acadgild@localhost hadoop]$ 
```



```
scala> val rdd = sc.textFile("/home/acadgild/hadoop/word-count.txt")
rdd: org.apache.spark.rdd.RDD[String] = /home/acadgild/hadoop/word-count.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> val line_count = rdd.count
line_count: Long = 5

scala> println(line_count)
5

scala> 
```

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

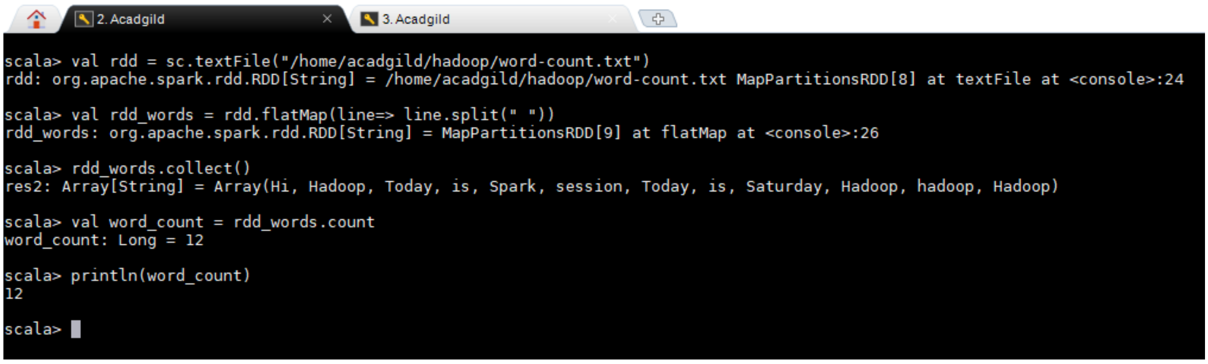
   val rdd = sc.textFile("/home/acadgild/hadoop/word-count.txt")

   val rdd_words = rdd.flatMap(line=> line.split(" "))

   rdd_words.collect()

   val word_count = rdd_words.count

   println(word_count)



```
scala> val rdd = sc.textFile("/home/acadgild/hadoop/word-count.txt")
rdd: org.apache.spark.rdd.RDD[String] = /home/acadgild/hadoop/word-count.txt MapPartitionsRDD[8] at textFile at <console>:24

scala> val rdd_words = rdd.flatMap(line=> line.split(" "))
rdd_words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[9] at flatMap at <console>:26

scala> rdd_words.collect()
res2: Array[String] = Array(Hi, Hadoop, Today, is, Spark, session, Today, is, Saturday, Hadoop, hadoop, Hadoop)

scala> val word_count = rdd_words.count
word_count: Long = 12

scala> println(word_count)
12

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.**

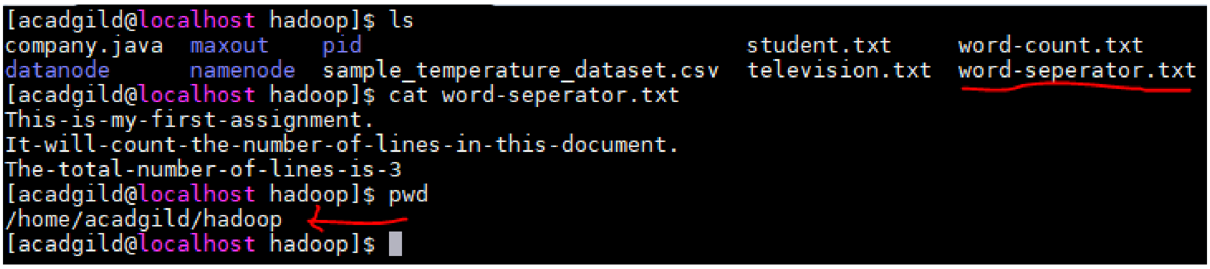
   val rdd_hyphen = sc.textFile("/home/acadgild/hadoop/word-seperator.txt")

   val rdd_hyphen_words = rdd_hyphen.flatMap(line=> line.split("-"))

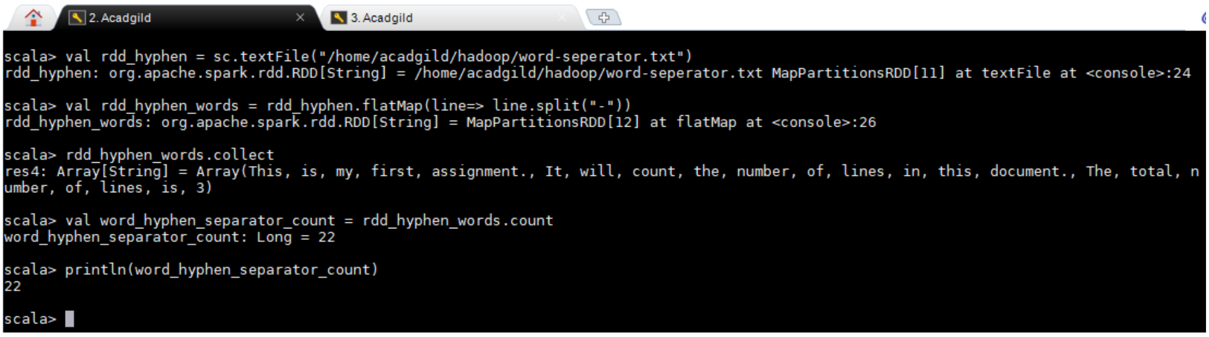
   rdd_hyphen_words.collect

   val word_hyphen_separator_count = rdd_hyphen_words.count

   println(word_hyphen_separator_count)



```
[acadgild@localhost hadoop]$ ls
company.java  maxout    pid                            student.txt     word-count.txt
datanode      namenode  sample_temperature_dataset.csv  television.txt  word-seperator.txt
[acadgild@localhost hadoop]$ cat word-seperator.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 hadoop]$ pwd
/home/acadgild/hadoop
[acadgild@localhost hadoop]$ 
```



```
scala> val rdd_hyphen = sc.textFile("/home/acadgild/hadoop/word-seperator.txt")
rdd_hyphen: org.apache.spark.rdd.RDD[String] = /home/acadgild/hadoop/word-seperator.txt MapPartitionsRDD[11] at textFile at <console>:24

scala> val rdd_hyphen_words = rdd_hyphen.flatMap(line=> line.split("-"))
rdd_hyphen_words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[12] at flatMap at <console>:26

scala> rdd_hyphen_words.collect
res4: Array[String] = Array(This, is, my, first, assignment., It, will, count, the, number, of, lines, in, this, document., The, total, n
umber, of, lines, is, 3)

scala> val word_hyphen_separator_count = rdd_hyphen_words.count
word_hyphen_separator_count: Long = 22

scala> println(word_hyphen_separator_count)
22

scala> 
```