Problem Statement

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

Solution:

Document Used:-Story

```
[acadgild@localhost ~]$ cat /home/acadgild/story
A Short Story : The Clever Crow

This Short Story The Clever Crow is quite interesting to all the people. Enjoy r
eading this story.

Once upon a time there lived a crow. She had built her nest on a tree. At the ro
ot of the same tree, a snake had built its home.

Whenever the crow laid eggs, the snake would eat them up. The crow felt helpless
. “That evil snake. I must do something. Let me go and talk to him," thought the
 crow.

The next morning, the crow went to the snake and said politely, “Please spare my
 eggs, dear friend. Let us live like good neighbors and not disturb each other."

“Huh! You cannot expect me to go hungry. Eggs are what I eat," replied the snake
, in a nasty tone.

The crow felt angry and she thought, “I must teach that snake a lesson."

The very next day, the crow was flying over the King’s palace. She saw the Princ
ess wearing an expensive necklace. Suddenly a thought flashed in her mind and sh
e swooped down, picked up the necklace in her beak and flew off to her nest.

When the Princess saw the crow flying off with her necklace, she screamed, “Some
body help, the crow has taken my necklace."

Soon the palace guards were running around in search of the necklace. Within a s
hort time the guards found the crow. She still sat with the necklace hanging fro
m her beak.

The clever crow thought, “Now is the time to act." And she dropped the necklace,
 which fell right into the snake’s pit of house.
```

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

Output:

```
SQL context available as sqlContext.

scala> val inputRDD = sc.textFile("/home/acadgild/story")
inputRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27

scala> val rows_count = inputRDD.count()
rows_count: Long = 29

scala>
```

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

Code:

```
scala> val inputRDD = sc.textFile("/home/acadgild/story")
inputRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[3] at textFile at <console>:27

scala> val words = inputRDD.flatMap(x => x.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[4] at flatMap at <console>:29

scala> val words_count = words.map(word => (word,1))
words_count: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[5] at map at <console>:31

scala> val words_number = words_count.reduceByKey((x,y) => x + y)
words_number: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[6] at reduceByKey at <console>:33

scala>
```

Output:

```
scala> words_number.foreach(println)
(next,2)
(reading,1)
(Let,2)
(it,1)
(The,9)
(Its,2)
(Soon,1)
(Short,2)
(politely,,1)
("I,1)
(tone.,1)
(help,,1)
(safe,",1)
(house.,2)
(This,1)
(necklace,,2)
(evil,1)
(over,1)
(With,1)
(crow.,3)
(Once,1)
(snake's,1)
(out,1)
(snake,,1)
(snake,6)
(felt,2)
(me,2)
(the,31)
(happy,,1)
(are,1)
(not,1)
(Crow,2)
(nasty,1)
(Clever,2)
(home.,1)
(morning,,1)
(eat,1)
(be,1)
(At,1)
(all,1)
(Then,1)
(sat,1)
(replied,1)
(them,1)
(:,1)
(palace,2)
(lived,1)
(eat",1)
(hungry.,1)
(lesson.",1)
(is,2)
(found,1)
(spare,1)
(disturb,1)
(short,1)
(pit,2)
(same,1)
(had,2)
(princess.,1)
(on,1)
(people.,1)
(built,2)
(would,1)
(him",1)
(Story,2)
(live,1)
(snake.,2)
(flashed,1)
(neighbors,1)
(talk,1)
(they,2)
(helpless.,1)
(root,1)
(my,3)
(swooped,1)
(noise,,1)
(,1)
(screamed,,1)
(has,1)
(us,1)
(running,1)
(killed,1)
(teach,1)
(I,2)
(flew,1)
(thought,,2)
(of,5)
(Whenever,1)
(must,2)
(saw,3)
(quite,1)
(flying,2)
(led,1)
(went,2)
(She,3)
(back,1)
(this,1)
(right,1)
(interesting,1)
(beak.,1)
(she,5)
(dear,1)
(good,1)
(wearing,1)
(taken,1)
(from,1)
(other.",1)
(hanging,1)
(Princess,2)
(shouted.,1)
(cannot,1)
(that,1)
(a,9)
(eggs,,2)
(will,1)
(to,7)
(friend.,1)
(,15)
(necklace.,2)
("Please,1)
(fell,1)
(very,1)
(crow,9)
(guards,4)
(were,1)
(and,10)
(angry,1)
(Kill,1)
(tree.,1)
(thought,3)
(it!",1)

scala>
```

- 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.

Document Used : sample

```
[acadgild@localhost ~]$ cat /home/acadgild/sample
This-is-my-first-assignment.
It-will-count-the-number-of-lines-in-this-document.
The-total-number-of-lines-is-3
[acadgild@localhost ~]$
```

Code:

```
scala> val inputRDD = sc.textFile("/home/acadgild/sample")
inputRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[10] at textFile at <console>:27

scala> val words = inputRDD.flatMap(x => x.split("-"))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[11] at flatMap at <console>:29

scala> val words_count = words.map(word => (word,1))
words_count: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[12] at map at <console>:31

scala> val words_number = words_count.reduceByKey((x,y) => x + y)
words_number: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[13] at reduceByKey at <console>:33
```

Output:

```
scala> words_number.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>
```

```
scala> words_number.count()
res3: Long = 18

scala>
```