Session 17 Assignment 1

Sample File:

```
[acadgild@localhost Session17Assignment1]$ cat sampledata.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 Session17Assignment1]$
```

Task 1 – To count the numbers of lines in the input file

```
scala> val data = sc.textFile("/home/acadgild/deepak/assignments/Session17Assignment1/sampledata.txt")
data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27

scala> data.count
res0: Long = 3

scala>
```

Task 2 – Count of each word in the file

```
scala> val wordcount = data.flatMap(line => line.split("-")).map(word => (word,1)).reduceByKey(_+_)
wordcount: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[16] at reduceByKey at <console>:29

scala> wordcount.saveAsTextFile("/home/acadgild/deepak/assignments/Session17Assignment1/Task2output")

scala>
```

```
[acadgild@localhost Task2output]$ pwd
/home/acadgild/deepak/assignments/Session17Assignment1/Task2output
[acadgild@localhost Task2output]$ cat part-00000
(this,1)
(lines,2)
(The,1)
(is,2)
(document.,1)
(assignment.,1)
(number,2)
(will,1)
(This,1)
(in,1)
(first,1)
(total,1)
(of,2)
(3.,1)
(It,1)
(my,1)
(count,1)
(the,1)
[acadgild@localhost Task2output]$
```

Task 3 – Total word count in the file

```
scala> val wordcount = data.flatMap(line => line.split("-")).filter(x => x.equals(x))
wordcount: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[19] at filter at <console>:29

scala> wordcount.count()
res9: Long = 22

scala>
```