

Roll No :31418
Batch: K4
Assignment 13

Step 1: Create a Folder to Work In

Open your terminal and run:

```
mkdir ~/spark-wordcount  
cd ~/spark-wordcount
```

This is your working directory.

Step 2: Create `passage.txt`

```
nano passage.txt
```

Paste this inside:

```
hello world  
hello spark  
hello scala  
I am Farkhanda  
spark is fast
```

Then:

- Press `Ctrl + O`, then `Enter` to save.
- Press `Ctrl + X` to exit.

You now have a file named `passage.txt`.

Step 3: Start the Spark Shell

From the **same folder** (`~/spark-wordcount`), run:

```
spark-shell
```

This opens a Scala REPL with Spark already set up (SC is available).

Step 4: Run the Word Count Program

In the spark-shell:

```
val input = sc.textFile("passage.txt")

val words = input.flatMap(line => line.split(" "))

val counts = words.map(word => (word, 1))

val reducedCounts = counts.reduceByKey(_ + _)

reducedCounts.foreach(println)
```

OUTPUT:

```
(hello,3)
(world,1)
(spark,2)
(is,1)
(fast,1)
(scala,1)
```

```
Welcome to
 _ _ _ _ _
|_/_/_/_/_| version 3.5.5
|_/_/_/_/_|
|_/_/_/_/_|

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 1.8.0_422)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val input = sc.textFile("passage.txt")
input: org.apache.spark.rdd.RDD[String] = passage.txt MapPartitionsRDD[1] at textFile at <console>:23

scala>

scala> val words = input.flatMap(line => line.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:23

scala>

scala> val counts = words.map(word => (word, 1))
counts: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:23

scala>

scala> val reducedCounts = counts.reduceByKey(_ + _)
reducedCounts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:23

scala>

scala> reducedCounts.foreach(println)
(scala,1)
(spark,2)
(is,1)
(I,1)
(am,1)
(fast,1)
(hello,3)
(Farkhanda,1)
(world,1)
```

