

USN : 1BM20CS195

```
bmsce@bmsce-Precision-T1700:~$ val data=sc.textFile("py.txt")
bash: syntax error near unexpected token `{'
bmsce@bmsce-Precision-T1700:~$ spark-shell
23/06/10 09:18:56 WARN Utils: Your hostname, bmsce-Precision-T1700 resolves to a loopback address: 127.0.1.1; using 10.124.7.69 instead (on interface enp1s0)
23/06/10 09:18:56 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
23/06/10 09:18:57 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://10.124.7.69:4040
Spark context available as 'sc' (master = local[*], app id = local-1686368942545).
Spark session available as 'spark'.
Welcome to

  ____              __
 / ___/____  ____  /  /
/ /  / __ \/ __ \/  /
/ /__/ /_/ / /_/ /  /
\___/\____/____/\___/

 version 3.0.3

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 1.8.0_232)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

```
... 47 elided
scala> val data=sc.textFile("/home/bmsce/Desktop/py.txt")
data: org.apache.spark.rdd.RDD[String] = /home/bmsce/Desktop/py.txt MapPartitionsRDD[5] at textFile at <console>:24

scala> data.collect;
res3: Array[String] = Array(hi how are you, how is your job, how is your family, how is your brother, how is your sister)

scala> val splitdata=data.flatMap(line=>line.split(" "));
splitdata: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at flatMap at <console>:25

scala> splitdata.collect;
res4: Array[String] = Array(hi, how, are, you, how, is, your, job, how, is, your, family, how, is, your, brother, how, is, your, sister)

scala> val mapdata=splitdata.map(word=>(word,1));
mapdata: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[7] at map at <console>:25

scala> mapdata.collect;
res5: Array[(String, Int)] = Array((hi,1), (how,1), (are,1), (you,1), (how,1), (is,1), (your,1), (job,1), (how,1), (is,1), (your,1), (family,1), (how,1), (is,1), (your,1), (brother,1), (how,1), (is,1), (your,1), (sister,1))

scala> val reducedata=mapdata.reduceByKey(_+_);
reducedata: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[8] at reduceByKey at <console>:25

scala> reducedata.collect;
res6: Array[(String, Int)] = Array((are,1), (brother,1), (is,4), (sister,1), (family,1), (how,5), (job,1), (you,1), (hi,1), (your,4))

scala>
```

```
scala> val data=sc.textFile("/home/bmsce/Desktop/py.txt")
data: org.apache.spark.rdd.RDD[String] = /home/bmsce/Desktop/py.txt MapPartitionsRDD[5] at textFile at <console>:24

scala> data.collect;
res3: Array[String] = Array(hi how are you, how is your job, how is your family, how is your brother, how is your sister)

scala> val splitdata=data.flatMap(line=>line.split(" "));
splitdata: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at flatMap at <console>:25

scala> splitdata.collect;
res4: Array[String] = Array(hi, how, are, you, how, is, your, job, how, is, your, family, how, is, your, brother, how, is, your, sister)

scala> val mapdata=splitdata.map(word=>(word,1));
mapdata: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[7] at map at <console>:25

scala> mapdata.collect;
res5: Array[(String, Int)] = Array((hi,1), (how,1), (are,1), (you,1), (how,1), (is,1), (your,1), (job,1), (how,1), (is,1), (your,1), (family,1), (how,1), (is,1), (your,1), (brother,1), (how,1), (is,1), (your,1), (sister,1))

scala> val reducedata=mapdata.reduceByKey(_+_);
reducedata: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[8] at reduceByKey at <console>:25

scala> reducedata.collect;
res6: Array[(String, Int)] = Array((are,1), (brother,1), (is,4), (sister,1), (family,1), (how,5), (job,1), (you,1), (hi,1), (your,4))

scala> val textFile=sc.textFile("/home/bmsce/Desktop/py.txt");
textFile: org.apache.spark.rdd.RDD[String] = /home/bmsce/Desktop/py.txt MapPartitionsRDD[10] at textFile at <console>:24

scala> val counts=textFile.flatMap(line=>line.split(" ")).map(word=>(word,1)).reduceByKey(_+_);
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[13] at reduceByKey at <console>:25

scala> import scala.collection.immutable.ListMap
import scala.collection.immutable.ListMap

scala> val sorted=ListMap(counts.collect.sortWith(_. _2> _2):_*);
sorted: scala.collection.immutable.ListMap[String,Int] = ListMap(how -> 5, is -> 4, your -> 4, are -> 1, brother -> 1, sister -> 1, family -> 1, job -> 1, you -> 1, hi -> 1)

scala> println(sorted)
ListMap(how -> 5, is -> 4, your -> 4, are -> 1, brother -> 1, sister -> 1, family -> 1, job -> 1, you -> 1, hi -> 1)

scala> for((k,v)<-sorted)
| {
|   if(v>4)
|   {
|     print(k+",")
|     print(v)
|     println()
|   }
| }
```