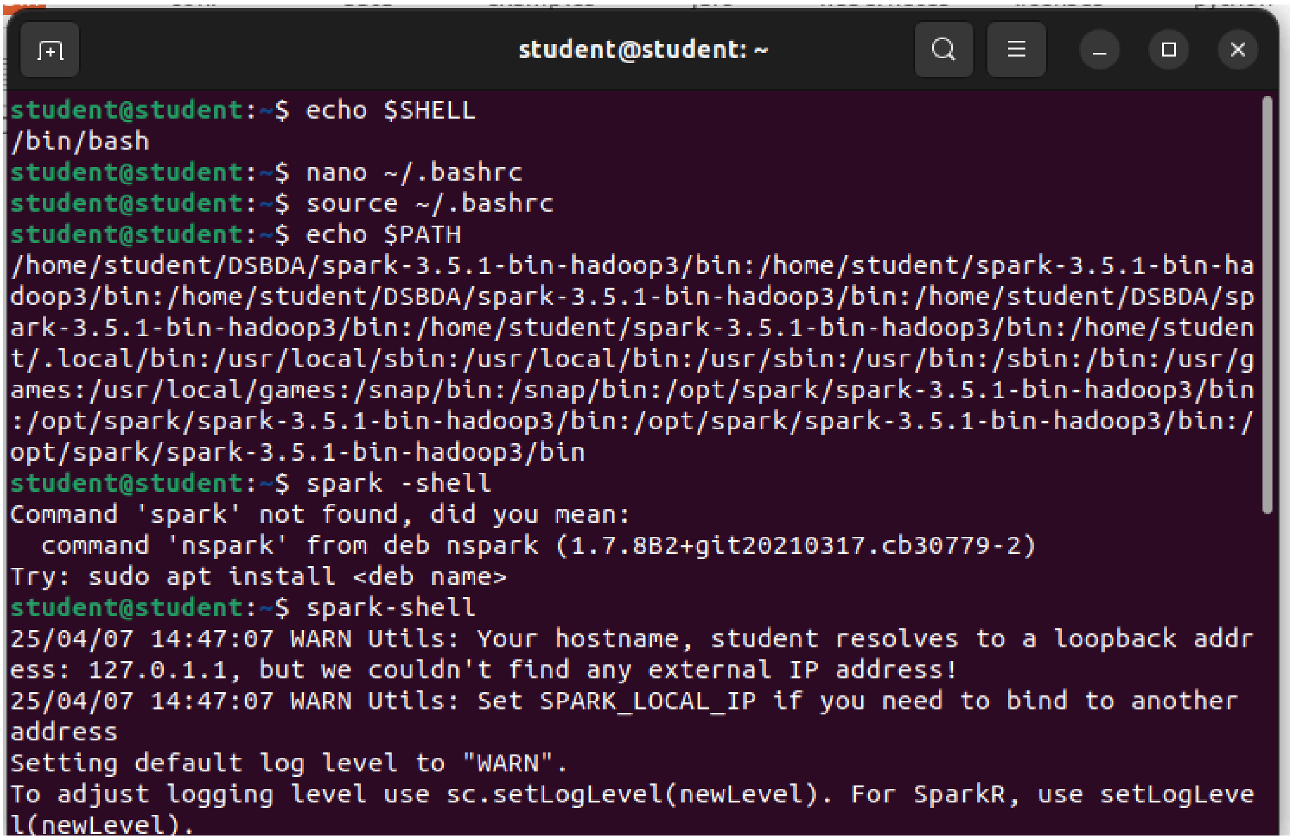
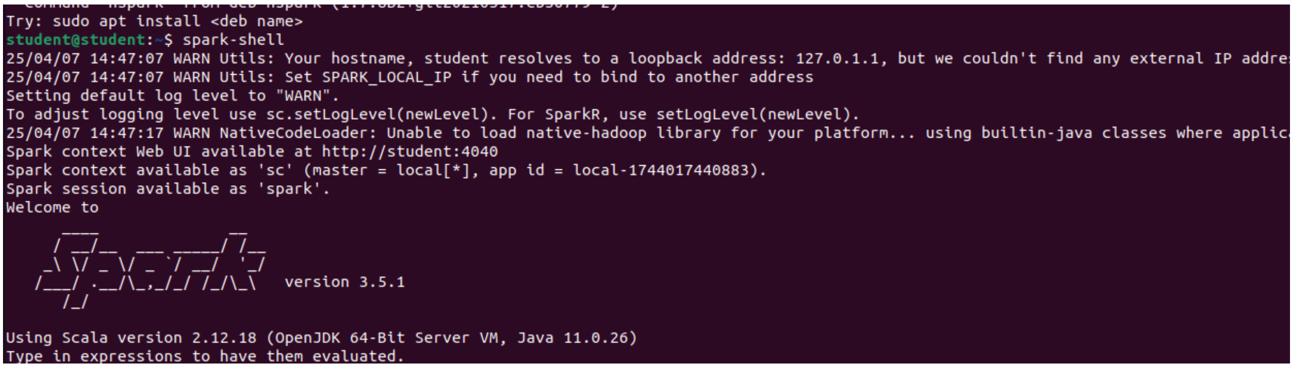
Input file:

One day I was returning from the school. When I was about 1 Km away from my house, it started raining. I had no umbrella. I stood under a shed in front of a shop. Half an hour passed. Rain became heavier. Street was full of water now. I could not stand like this forever. So I decided to go in rain. I drenched completely. Some children were playing with paper boats. I also joined them. I forgot everything. Suddenly I saw my mother was coming with an umbrella. She scolded me. But I was happy. I can’t forget that day.

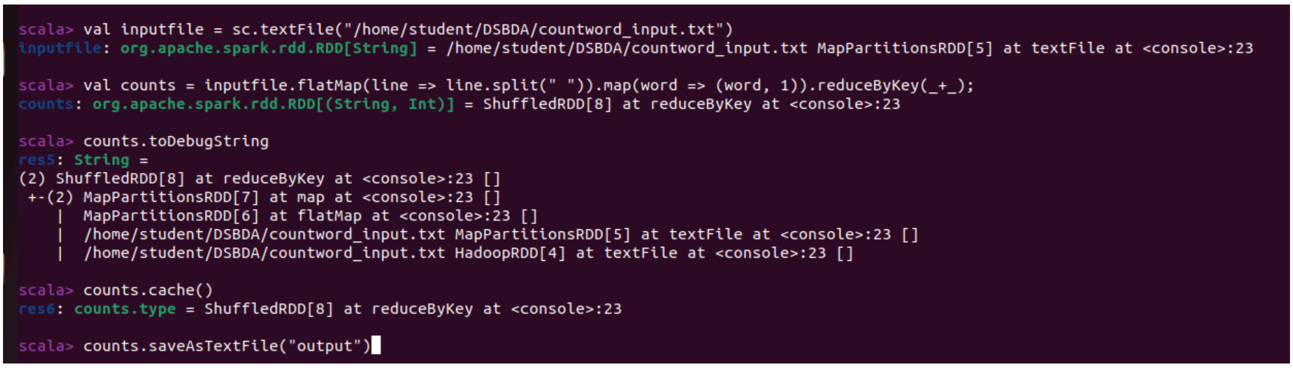
Scala Word count



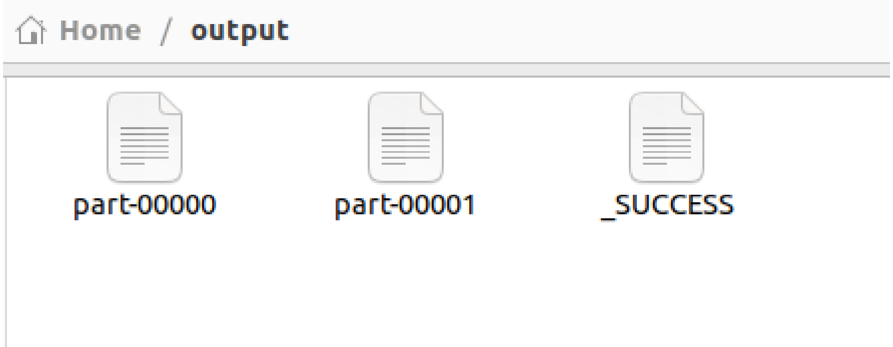
```
student@student:~$ echo $SHELL
/bin/bash
student@student:~$ nano ~/.bashrc
student@student:~$ source ~/.bashrc
student@student:~$ echo $PATH
/home/student/DSBDA/spark-3.5.1-bin-hadoop3/bin:/home/student/spark-3.5.1-bin-ha
doop3/bin:/home/student/DSBDA/spark-3.5.1-bin-hadoop3/bin:/home/student/DSBDA/sp
ark-3.5.1-bin-hadoop3/bin:/home/student/spark-3.5.1-bin-hadoop3/bin:/home/studen
t/.local/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/g
ames:/usr/local/games:/snap/bin:/snap/bin:/opt/spark/spark-3.5.1-bin-hadoop3/bin
:/opt/spark/spark-3.5.1-bin-hadoop3/bin:/opt/spark/spark-3.5.1-bin-hadoop3/bin:/
opt/spark/spark-3.5.1-bin-hadoop3/bin
student@student:~$ spark -shell
Command 'spark' not found, did you mean:
  command 'nspark' from deb nspark (1.7.8B2+git20210317.cb30779-2)
Try: sudo apt install <deb name>
student@student:~$ spark-shell
25/04/07 14:47:07 WARN Utils: Your hostname, student resolves to a loopback addr
ess: 127.0.1.1, but we couldn't find any external IP address!
25/04/07 14:47:07 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
```



```
  command 'nspark' from deb nspark (1.7.8B2+git20210317.cb30779-2)
Try: sudo apt install <deb name>
student@student:~$ spark-shell
25/04/07 14:47:07 WARN Utils: Your hostname, student resolves to a loopback address: 127.0.1.1, but we couldn't find any external IP addre
25/04/07 14:47:07 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/07 14:47:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applic
Spark context Web UI available at http://student:4040
Spark context available as 'sc' (master = local[*], app id = local-1744017440883).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
```



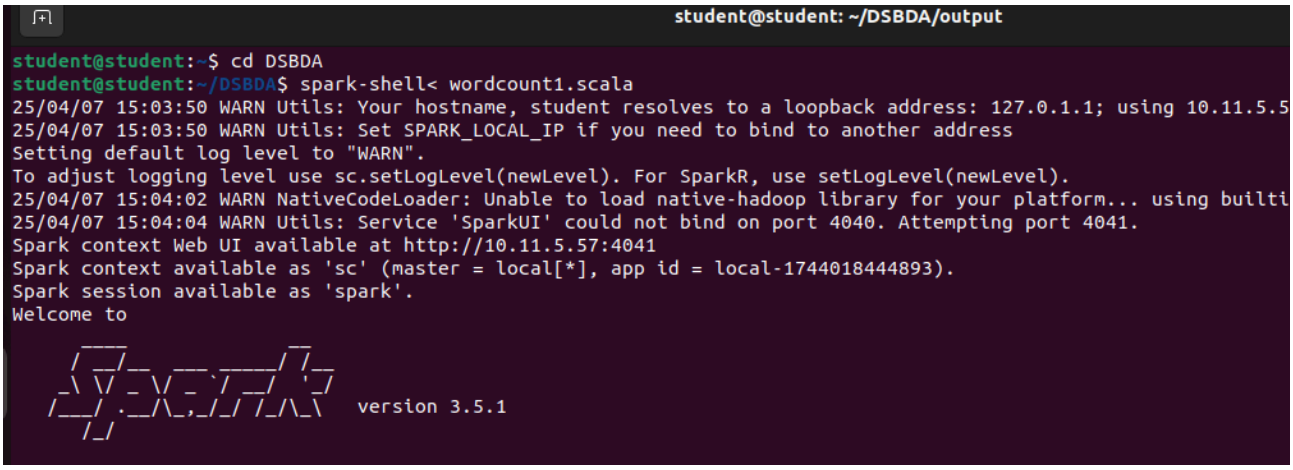
```
scala> val inputfile = sc.textFile("/home/student/DSBDA/countword_input.txt")
inputfile: org.apache.spark.rdd.RDD[String] = /home/student/DSBDA/countword_input.txt MapPartitionsRDD[5] at textFile at <console>:23

scala> val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_+_);
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[8] at reduceByKey at <console>:23

scala> counts.toDebugString
res5: String =
(2) ShuffledRDD[8] at reduceByKey at <console>:23 []
 +-(2) MapPartitionsRDD[7] at map at <console>:23 []
    |  MapPartitionsRDD[6] at flatMap at <console>:23 []
    |  /home/student/DSBDA/countword_input.txt MapPartitionsRDD[5] at textFile at <console>:23 []
    |  /home/student/DSBDA/countword_input.txt HadoopRDD[4] at textFile at <console>:23 []

scala> counts.cache()
res6: counts.type = ShuffledRDD[8] at reduceByKey at <console>:23

scala> counts.saveAsTextFile("output")
```



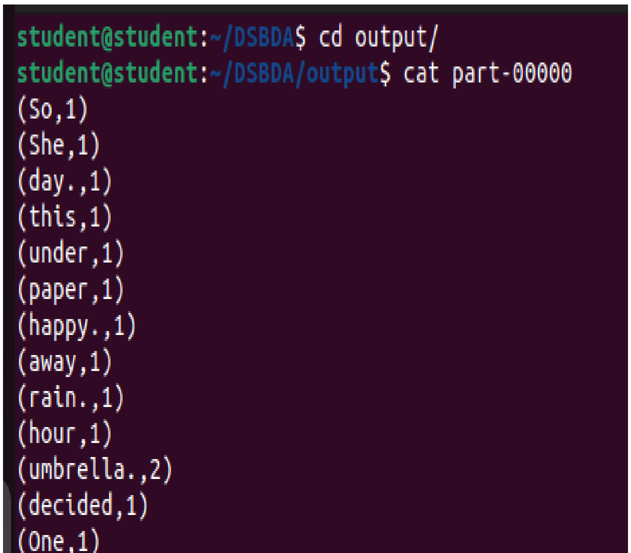
//Method 2

wordcount1.scala

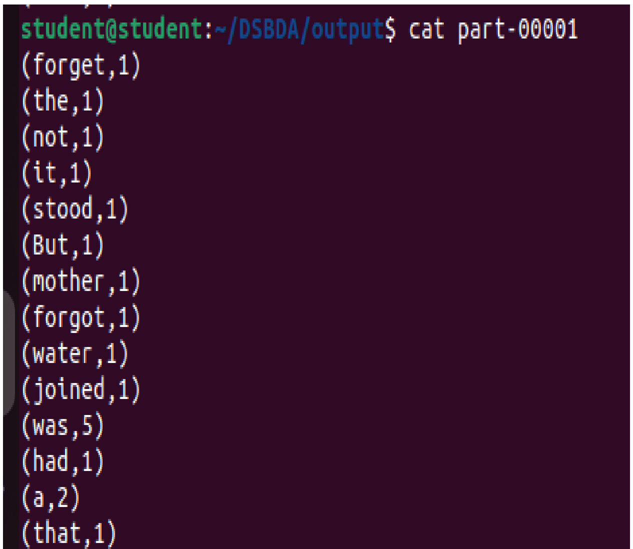
```
val inputfile = sc.textFile("/home/student/DSBDA/countword_input.txt")
val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word,
1)).reduceByKey(_+_);
counts.toDebugString
counts.cache()
counts.saveAsTextFile("output")
```



```
student@student:~$ cd DSBDA
student@student:~/DSBDA$ spark-shell< wordcount1.scala
25/04/07 15:03:50 WARN Utils: Your hostname, student resolves to a loopback address: 127.0.1.1; using 10.11.5.5
25/04/07 15:03:50 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/04/07 15:04:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builti
25/04/07 15:04:04 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Spark context Web UI available at http://10.11.5.57:4041
Spark context available as 'sc' (master = local[*], app id = local-1744018444893).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.1
      /_/
```



```
student@student:~/DSBDA$ cd output/
student@student:~/DSBDA/output$ cat part-00000
(So,1)
(She,1)
(day.,1)
(this,1)
(under,1)
(paper,1)
(happy.,1)
(away,1)
(rain.,1)
(hour,1)
(umbrella.,2)
(decided,1)
(One,1)
```



```
student@student:~/DSBDA/output$ cat part-00001
(forget,1)
(the,1)
(not,1)
(it,1)
(stood,1)
(But,1)
(mother,1)
(forgot,1)
(water,1)
(joined,1)
(was,5)
(had,1)
(a,2)
(that,1)
```