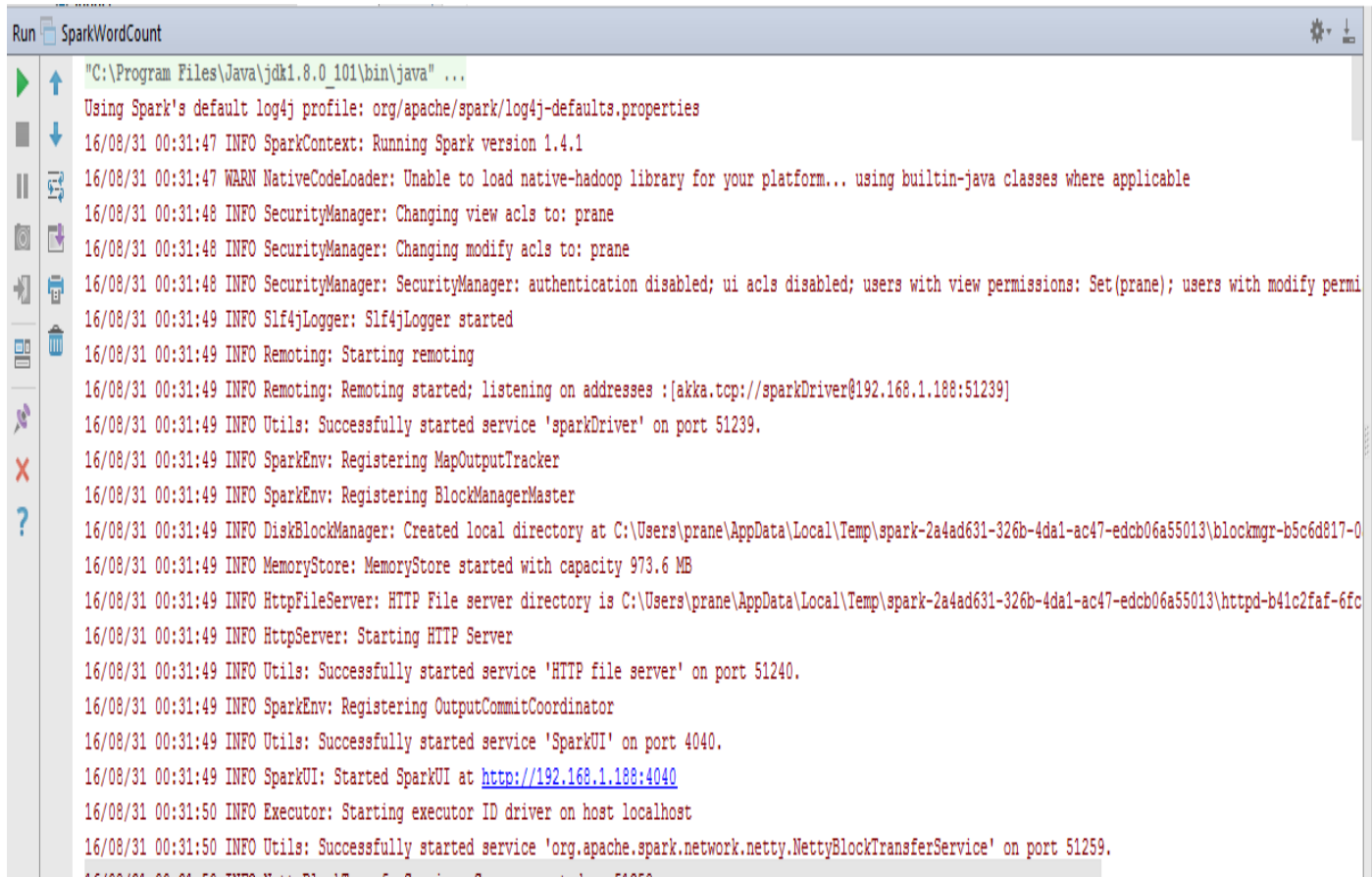


Real Time Big Data Analytics Lab Assignment-1

- In this lab we are going to count sentences from a given input file and display them in a sorted order.
- The program for accomplishing the above has been written in scala using IntelliJ IDEA
- The below screenshot shows the running of the program.

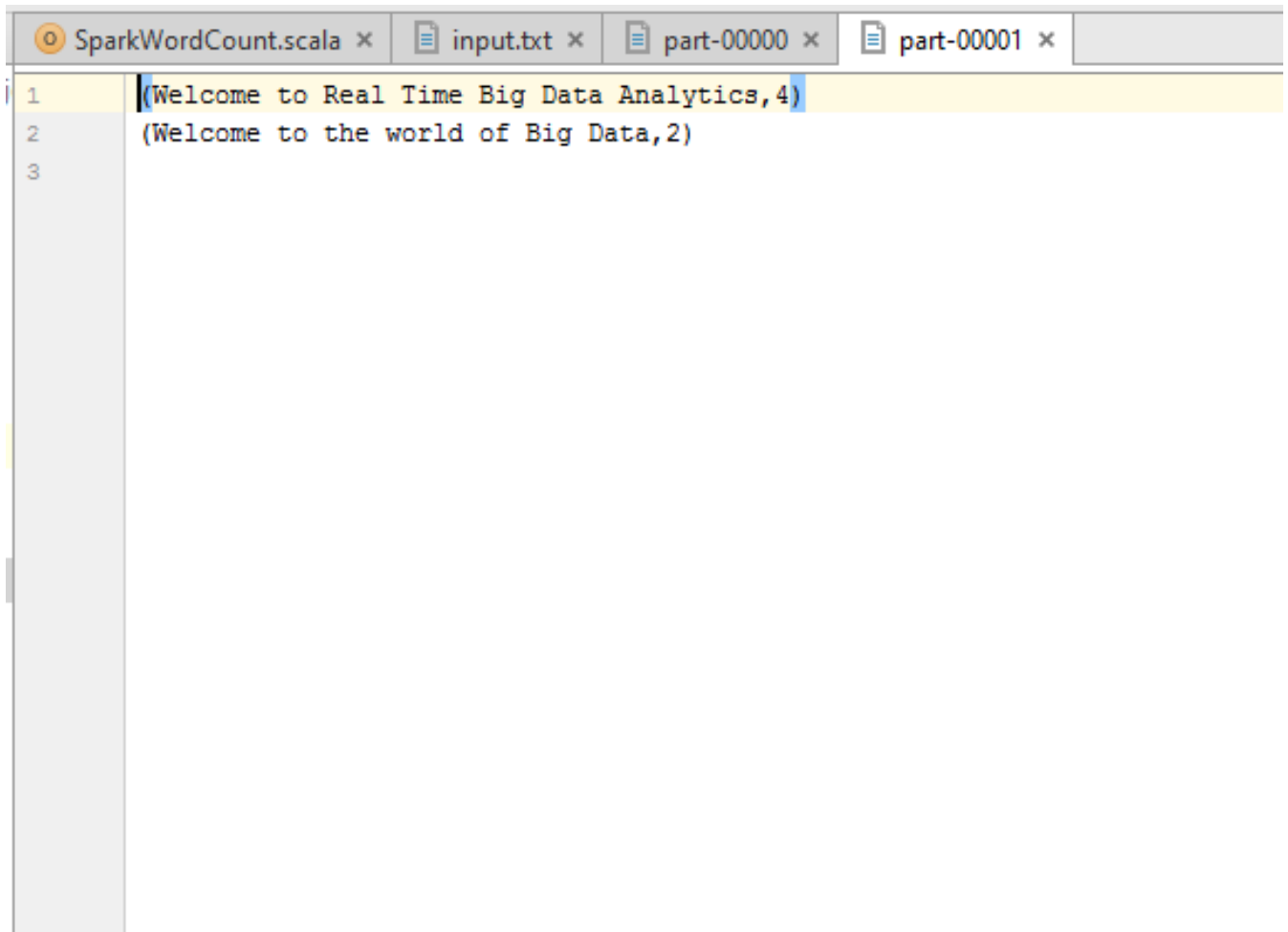


```
Run SparkWordCount
"C:\Program Files\Java\jdk1.8.0_101\bin\java" ...
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
16/08/31 00:31:47 INFO SparkContext: Running Spark version 1.4.1
16/08/31 00:31:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/08/31 00:31:48 INFO SecurityManager: Changing view acls to: prane
16/08/31 00:31:48 INFO SecurityManager: Changing modify acls to: prane
16/08/31 00:31:48 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(prane); users with modify permissions: Set(prane)
16/08/31 00:31:49 INFO Slf4jLogger: Slf4jLogger started
16/08/31 00:31:49 INFO Remoting: Starting remoting
16/08/31 00:31:49 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@192.168.1.188:51239]
16/08/31 00:31:49 INFO Utils: Successfully started service 'sparkDriver' on port 51239.
16/08/31 00:31:49 INFO SparkEnv: Registering MapOutputTracker
16/08/31 00:31:49 INFO SparkEnv: Registering BlockManagerMaster
16/08/31 00:31:49 INFO DiskBlockManager: Created local directory at C:\Users\prane\AppData\Local\Temp\spark-2a4ad631-326b-4da1-ac47-edcb06a55013\blockmgr-b5c6d817-0
16/08/31 00:31:49 INFO MemoryStore: MemoryStore started with capacity 973.6 MB
16/08/31 00:31:49 INFO HttpFileServer: HTTP File server directory is C:\Users\prane\AppData\Local\Temp\spark-2a4ad631-326b-4da1-ac47-edcb06a55013\httpd-b41c2faf-6fc
16/08/31 00:31:49 INFO HttpServer: Starting HTTP Server
16/08/31 00:31:49 INFO Utils: Successfully started service 'HTTP file server' on port 51240.
16/08/31 00:31:49 INFO SparkEnv: Registering OutputCommitCoordinator
16/08/31 00:31:49 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/08/31 00:31:49 INFO SparkUI: Started SparkUI at http://192.168.1.188:4040
16/08/31 00:31:50 INFO Executor: Starting executor ID driver on host localhost
16/08/31 00:31:50 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 51259.
16/08/31 00:31:50 INFO NettyBlockTransferService: Server started on 51259
```

- The input file for the sentence count is as below
- The input is a paragraph with multiple similar sentences.

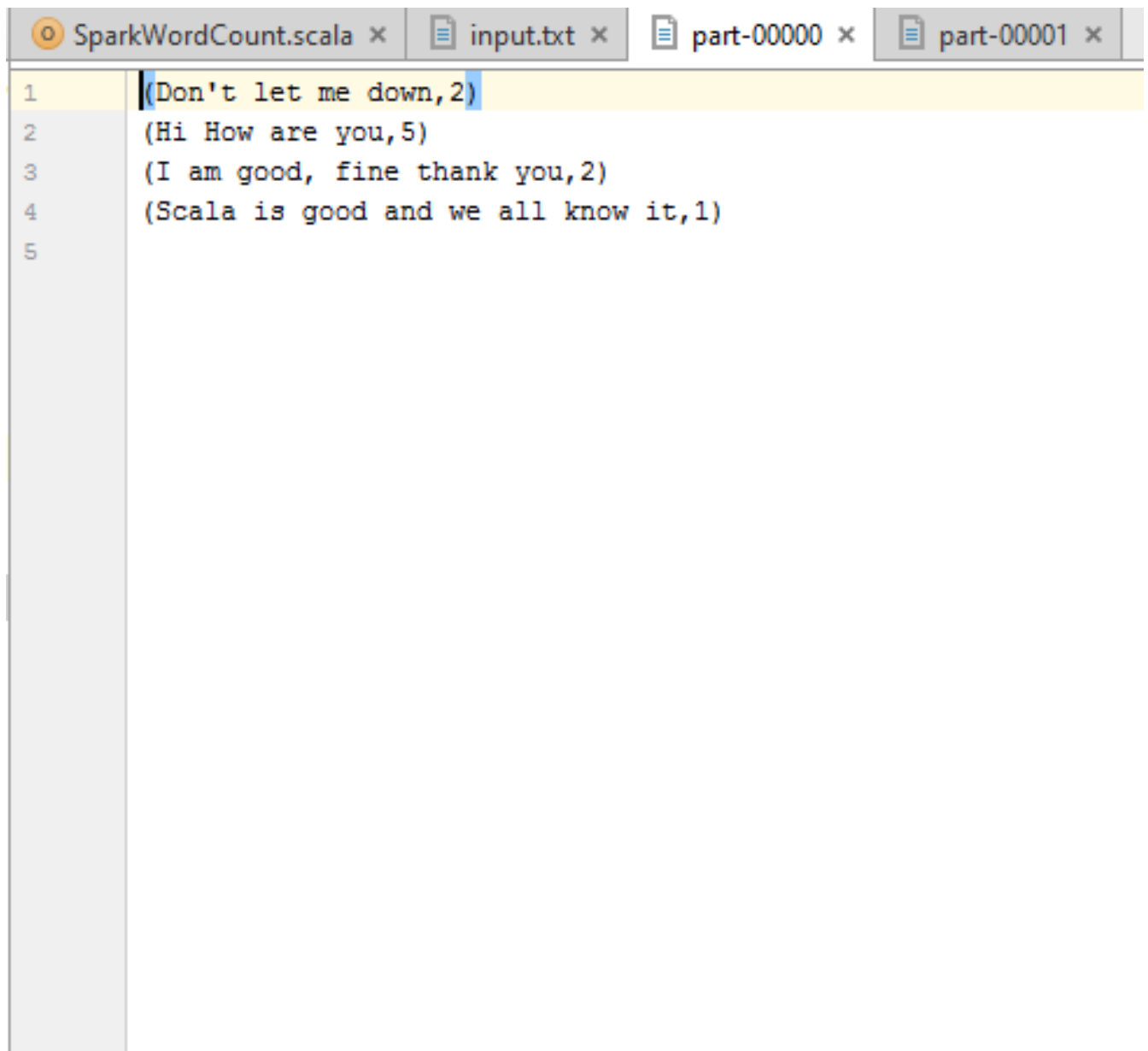
```
Welcome to the world of Big Data. Hi How are you. Welcome to Real Time Big Data Analytics. I am good, fine thank you. Welcome to Real Time Big Data Analytics.
Hi How are you. Hi How are you. Don't let me down. Hi How are you. Hi How are you. Welcome to Real Time Big Data Analytics. I am good, fine thank you.
Welcome to Real Time Big Data Analytics. Scala is good and we all know it. Don't let me down. Welcome to the world of Big Data
```

- Once the program is run Spark generates two output part files.
- You can see the key value pair as below with the key being the sentence and value being the number of occurrences of the sentence.
- Also the sentences are sorted by the key.
- Output part file-1



```
1 (Welcome to Real Time Big Data Analytics,4)
2 (Welcome to the world of Big Data,2)
3
```

- Output part file-2



```
1 (Don't let me down,2)
2 (Hi How are you,5)
3 (I am good, fine thank you,2)
4 (Scala is good and we all know it,1)
5
```