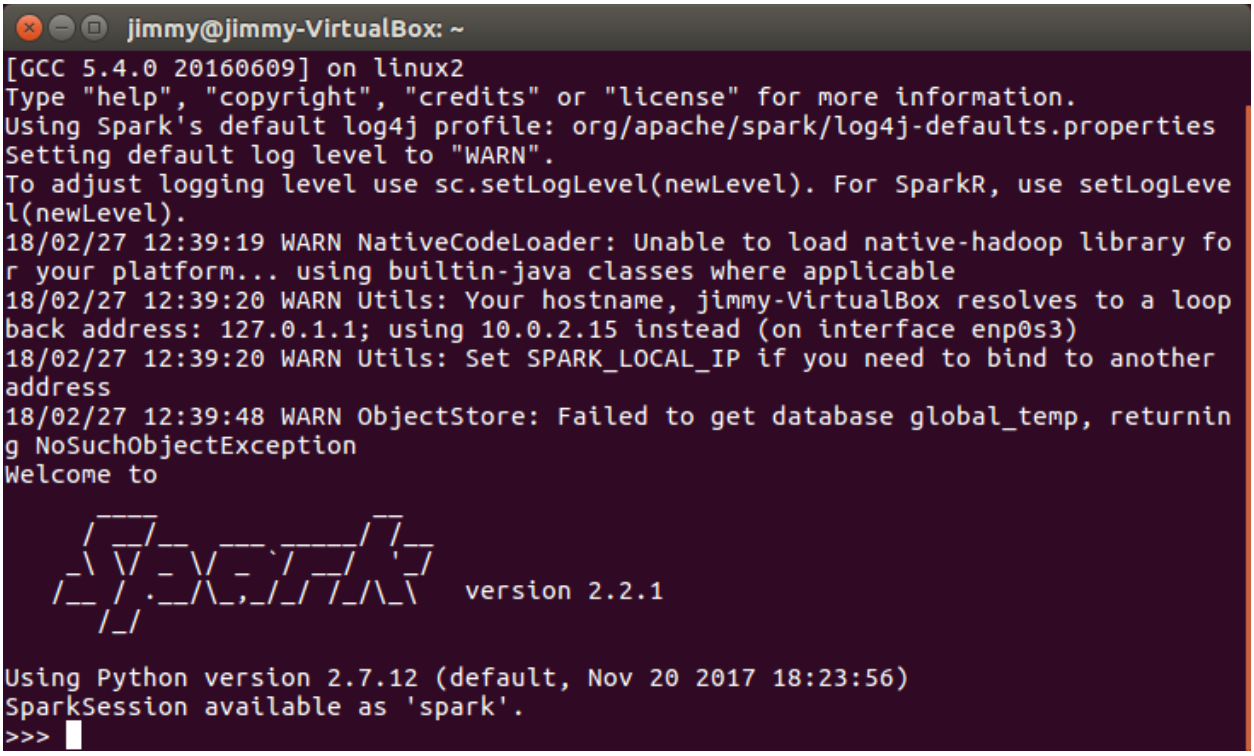
Jimmy Tran

CSC 4760

28 February 2018

Assignment 3 Report

1) Spark installed successfully



2) test.txt output

```
18/02/27 12:43:18 INFO DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/jimmy/Documents/WordCount.py:22) finished in 3.720 s
18/02/27 12:43:18 INFO DAGScheduler: looking for newly runnable stages
18/02/27 12:43:18 INFO DAGScheduler: running: Set()
18/02/27 12:43:18 INFO DAGScheduler: waiting: Set(ResultStage 1)
18/02/27 12:43:18 INFO DAGScheduler: failed: Set()
18/02/27 12:43:18 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:48), which has no missing parents
18/02/27 12:43:18 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 7.8 KB, free 366.0 MB)
18/02/27 12:43:18 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 4.9 KB, free 366.0 MB)
18/02/27 12:43:18 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 10.0.2.15:38337 (size: 4.9 KB, free: 366.3 MB)
18/02/27 12:43:18 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006
18/02/27 12:43:18 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:48) (first 15 tasks
 are for partitions Vector(0))
18/02/27 12:43:18 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/02/27 12:43:18 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 4621 bytes)
18/02/27 12:43:18 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/02/27 12:43:19 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/02/27 12:43:19 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 41 ms
18/02/27 12:43:19 INFO PythonRunner: Times: total = 65, boot = -2141, init = 2204, finish = 2
18/02/27 12:43:19 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1675 bytes result sent to driver
18/02/27 12:43:19 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 364 ms on localhost (executor driver) (1/1)
18/02/27 12:43:19 INFO DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:455) finished in 0.369 s
18/02/27 12:43:19 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/02/27 12:43:19 INFO DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:455, took 5.875679 s
[(u'hadoop', 4), (u'spark', 3), (u'pig', 2), (u'hbase', 1), (u'hive', 1)]
18/02/27 12:43:19 INFO SparkContext: Invoking stop() from shutdown hook
18/02/27 12:43:19 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/02/27 12:43:19 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/02/27 12:43:19 INFO MemoryStore: MemoryStore cleared
18/02/27 12:43:19 INFO BlockManager: BlockManager stopped
18/02/27 12:43:19 INFO BlockManagerMaster: BlockManagerMaster stopped
18/02/27 12:43:19 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/02/27 12:43:19 INFO SparkContext: Successfully stopped SparkContext
18/02/27 12:43:19 INFO ShutdownHookManager: Shutdown hook called
18/02/27 12:43:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a31f7c5-847e-4c6e-820e-74b924b0077f/pyspark-f2d9b0e2-7596-4208-9dd3
-45a29fd6e640
18/02/27 12:43:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-9a31f7c5-847e-4c6e-820e-74b924b0077f
jimmy@jimmy-VirtualBox:~/Documents$
```

peterpan.txt output

```
18/02/27 12:44:56 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:48), which has no missing parents
18/02/27 12:44:56 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 7.8 KB, free 366.0 MB)
18/02/27 12:44:56 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 4.9 KB, free 366.0 MB)
18/02/27 12:44:56 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 10.0.2.15:37875 (size: 4.9 KB, free: 366.3 MB)
18/02/27 12:44:56 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006
18/02/27 12:44:56 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at RDD at PythonRDD.scala:48) (first 15 tasks
 are for partitions Vector(0))
18/02/27 12:44:56 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
18/02/27 12:44:56 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 4621 bytes)
18/02/27 12:44:56 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
18/02/27 12:44:57 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/02/27 12:44:57 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 46 ms
18/02/27 12:44:57 INFO PythonRunner: Times: total = 175, boot = -1686, init = 1738, finish = 123
18/02/27 12:44:57 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 2046 bytes result sent to driver
18/02/27 12:44:57 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 522 ms on localhost (executor driver) (1/1)
18/02/27 12:44:57 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/02/27 12:44:57 INFO DAGScheduler: ResultStage 1 (runJob at PythonRDD.scala:455) finished in 0.539 s
18/02/27 12:44:57 INFO DAGScheduler: Job 0 finished: runJob at PythonRDD.scala:455, took 6.328777 s
[(u'the', 2331), (u'', 2259), (u'and', 1396), (u'to', 1214), (u'a', 962), (u'of', 929), (u'was', 898), (u'he', 866), (u'in', 683), (u'that', 5
64), (u'had', 498), (u'it', 473), (u'they', 465), (u'she', 465), (u'his', 455), (u'you', 403), (u'but', 378), (u'for', 377), (u'not', 375), (u
'her', 361), (u'with', 361), (u'is', 350), (u'on', 329), (u'at', 322), (u'as', 315), (u'I', 253), (u'be', 249), (u'have', 247), (u'were', 243)
, (u'Peter', 238)]
18/02/27 12:44:57 INFO BlockManagerInfo: Removed broadcast_2_piece0 on 10.0.2.15:37875 in memory (size: 4.9 KB, free: 366.3 MB)
18/02/27 12:44:57 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 10.0.2.15:37875 in memory (size: 5.8 KB, free: 366.3 MB)
18/02/27 12:44:57 INFO SparkContext: Invoking stop() from shutdown hook
18/02/27 12:44:57 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/02/27 12:44:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/02/27 12:44:58 INFO MemoryStore: MemoryStore cleared
18/02/27 12:44:58 INFO BlockManager: BlockManager stopped
18/02/27 12:44:58 INFO BlockManagerMaster: BlockManagerMaster stopped
18/02/27 12:44:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/02/27 12:44:58 INFO SparkContext: Successfully stopped SparkContext
18/02/27 12:44:58 INFO ShutdownHookManager: Shutdown hook called
18/02/27 12:44:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-6cf7ac2e-5dd9-4309-86c1-50beb04852e6/pyspark-f4d931c7-53a4-413d-92f7
-dfb465c05535
18/02/27 12:44:58 INFO ShutdownHookManager: Deleting directory /tmp/spark-6cf7ac2e-5dd9-4309-86c1-50beb04852e6
jimmy@jimmy-VirtualBox:~/Documents$
```

3) Report

The commands passed in the terminal were in the format “pyspark (.py file) (.txt file location) (top # of words, where # is the number the user inputs).

pyspark is the process/program called to run.

The .py file is the file being run by pyspark.

The .txt file is the location of the .txt file that the .py file is run on.

The last parameter is the number of top words the user wants returned or output after the program completes. For example, entering 15 would mean the user wants the 15 most common/repeated words found by the program in the given .txt file. In the case of the two files above, the top 5 and top 30 were output by the program in the middle of each screenshot in the format (word, count).