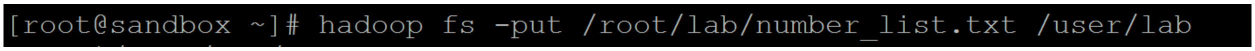
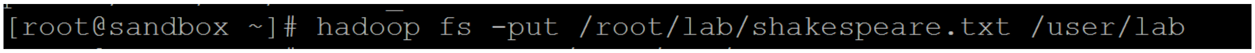
Spark

Summary: Using the datasets from two files, a script was written to yield certain results in Spark. The first and second results indicated the number of even and odd numbers within the dataset. The other half of this project included finding the top 10 and bottom 10 words of the dataset.

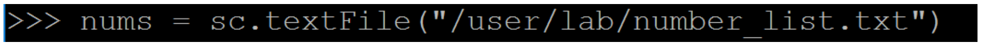
# Below is the script to find the even and odd counts within the dataset.



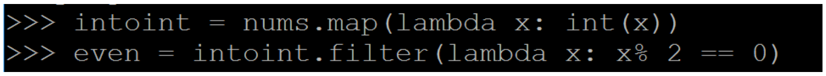
```
[root@sandbox ~]# hadoop fs -put /root/lab/number_list.txt /user/lab
```



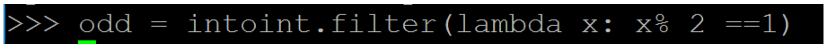
```
[root@sandbox ~]# hadoop fs -put /root/lab/shakespeare.txt /user/lab
```



```
>>> nums = sc.textFile("/user/lab/number_list.txt")
```

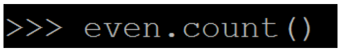


```
>>> intoint = nums.map(lambda x: int(x))
>>> even = intoint.filter(lambda x: x% 2 == 0)
```

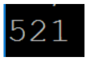


```
>>> odd = intoint.filter(lambda x: x% 2 ==1)
```

The even count result is:

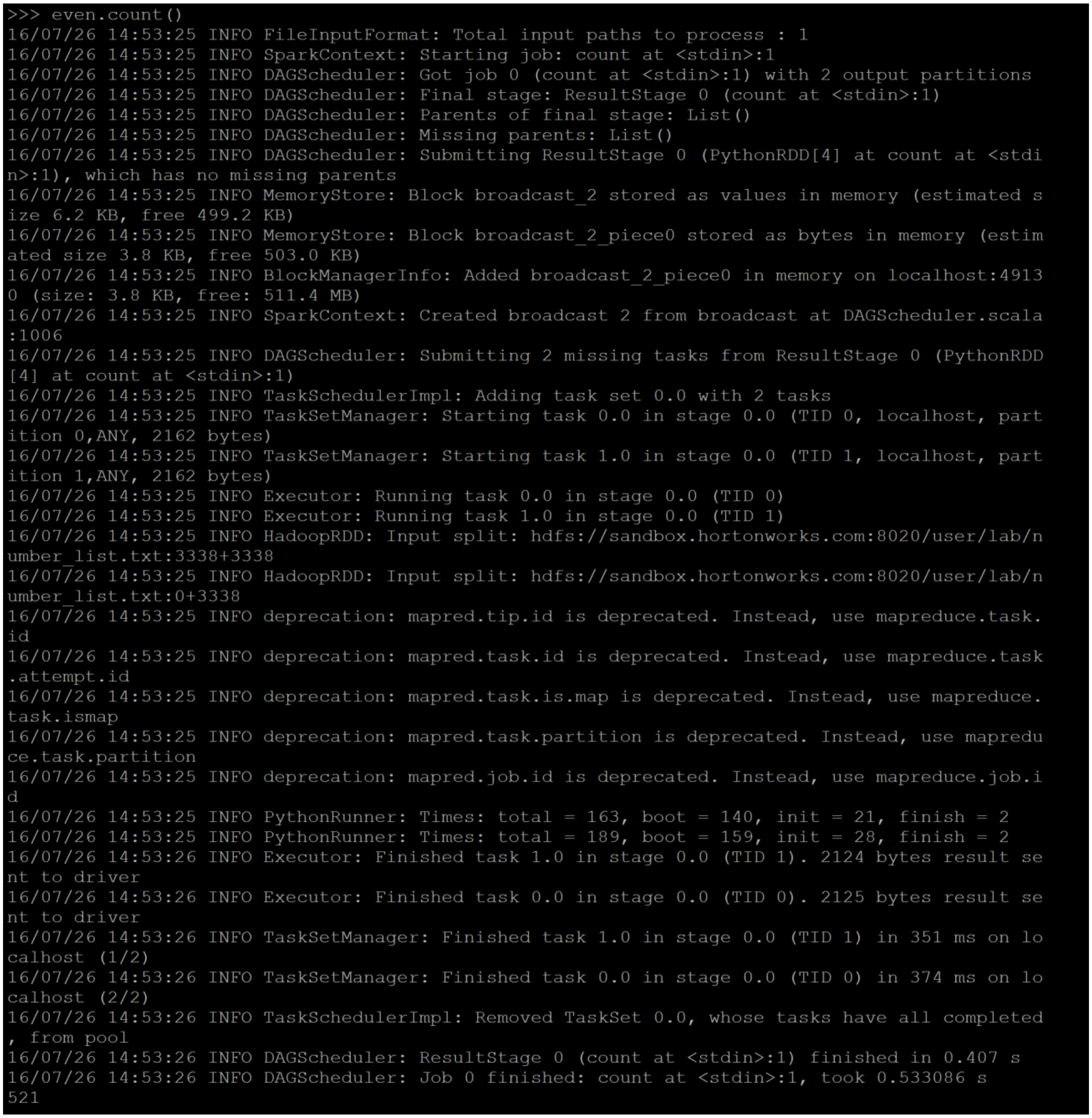


```
>>> even.count()
```



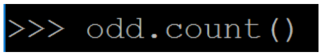
```
521
```

The following (below) shows the process that pyspark carried out and shows the answer as mentioned (above):

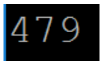


```
>>> even.count()
16/07/26 14:53:25 INFO FileInputFormat: Total input paths to process : 1
16/07/26 14:53:25 INFO SparkContext: Starting job: count at <stdin>:1
16/07/26 14:53:25 INFO DAGScheduler: Got job 0 (count at <stdin>:1) with 2 output partitions
16/07/26 14:53:25 INFO DAGScheduler: Final stage: ResultStage 0 (count at <stdin>:1)
16/07/26 14:53:25 INFO DAGScheduler: Parents of final stage: List()
16/07/26 14:53:25 INFO DAGScheduler: Missing parents: List()
16/07/26 14:53:25 INFO DAGScheduler: Submitting ResultStage 0 (PythonRDD[4] at count at <stdi
n>:1), which has no missing parents
16/07/26 14:53:25 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated s
ize 6.2 KB, free 499.2 KB)
16/07/26 14:53:25 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estim
ated size 3.8 KB, free 503.0 KB)
16/07/26 14:53:25 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:4913
0 (size: 3.8 KB, free: 511.4 MB)
16/07/26 14:53:25 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala
:1006
16/07/26 14:53:25 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 0 (PythonRDD
[4] at count at <stdin>:1)
16/07/26 14:53:25 INFO TaskSchedulerImpl: Adding task set 0.0 with 2 tasks
16/07/26 14:53:25 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, part
ition 0,ANY, 2162 bytes)
16/07/26 14:53:25 INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, part
ition 1,ANY, 2162 bytes)
16/07/26 14:53:25 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
16/07/26 14:53:25 INFO Executor: Running task 1.0 in stage 0.0 (TID 1)
16/07/26 14:53:25 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/lab/n
umber_list.txt:3338+3338
16/07/26 14:53:25 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/lab/n
umber_list.txt:0+3338
16/07/26 14:53:25 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.
id
16/07/26 14:53:25 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task
.attempt.id
16/07/26 14:53:25 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.
task.ismap
16/07/26 14:53:25 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapredu
ce.task.partition
16/07/26 14:53:25 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.i
d
16/07/26 14:53:25 INFO PythonRunner: Times: total = 163, boot = 140, init = 21, finish = 2
16/07/26 14:53:25 INFO PythonRunner: Times: total = 189, boot = 159, init = 28, finish = 2
16/07/26 14:53:26 INFO Executor: Finished task 1.0 in stage 0.0 (TID 1). 2124 bytes result se
nt to driver
16/07/26 14:53:26 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 2125 bytes result se
nt to driver
16/07/26 14:53:26 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 351 ms on lo
calhost (1/2)
16/07/26 14:53:26 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 374 ms on lo
calhost (2/2)
16/07/26 14:53:26 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed
, from pool
16/07/26 14:53:26 INFO DAGScheduler: ResultStage 0 (count at <stdin>:1) finished in 0.407 s
16/07/26 14:53:26 INFO DAGScheduler: Job 0 finished: count at <stdin>:1, took 0.533086 s
521
```

The odd count result is:

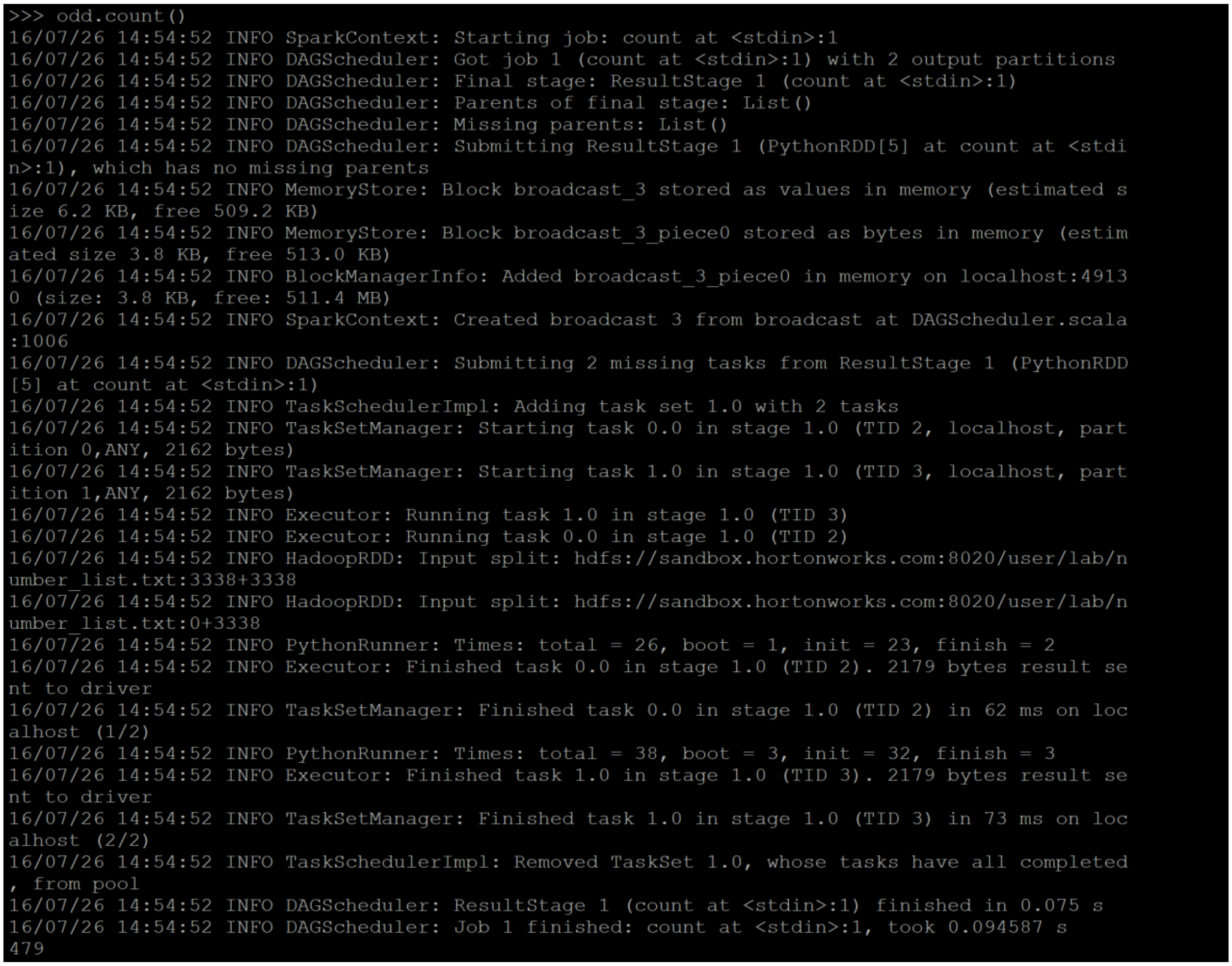


```
>>> odd.count()
```



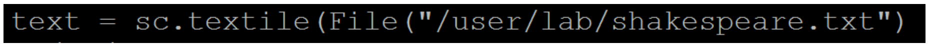
```
479
```

The following (below) shows the process that pyspark carried out and shows the answer as mentioned (above):

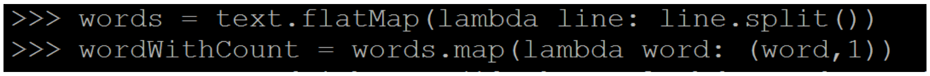


```
>>> odd.count()
16/07/26 14:54:52 INFO SparkContext: Starting job: count at <stdin>:1
16/07/26 14:54:52 INFO DAGScheduler: Got job 1 (count at <stdin>:1) with 2 output partitions
16/07/26 14:54:52 INFO DAGScheduler: Final stage: ResultStage 1 (count at <stdin>:1)
16/07/26 14:54:52 INFO DAGScheduler: Parents of final stage: List()
16/07/26 14:54:52 INFO DAGScheduler: Missing parents: List()
16/07/26 14:54:52 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[5] at count at <stdi
n>:1), which has no missing parents
16/07/26 14:54:52 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated s
ize 6.2 KB, free 509.2 KB)
16/07/26 14:54:52 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estim
ated size 3.8 KB, free 513.0 KB)
16/07/26 14:54:52 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:4913
0 (size: 3.8 KB, free: 511.4 MB)
16/07/26 14:54:52 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala
:1006
16/07/26 14:54:52 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (PythonRDD
[5] at count at <stdin>:1)
16/07/26 14:54:52 INFO TaskSchedulerImpl: Adding task set 1.0 with 2 tasks
16/07/26 14:54:52 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2, localhost, part
ition 0,ANY, 2162 bytes)
16/07/26 14:54:52 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3, localhost, part
ition 1,ANY, 2162 bytes)
16/07/26 14:54:52 INFO Executor: Running task 1.0 in stage 1.0 (TID 3)
16/07/26 14:54:52 INFO Executor: Running task 0.0 in stage 1.0 (TID 2)
16/07/26 14:54:52 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/lab/n
umber_list.txt:3338+3338
16/07/26 14:54:52 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/lab/n
umber_list.txt:0+3338
16/07/26 14:54:52 INFO PythonRunner: Times: total = 26, boot = 1, init = 23, finish = 2
16/07/26 14:54:52 INFO Executor: Finished task 0.0 in stage 1.0 (TID 2). 2179 bytes result se
nt to driver
16/07/26 14:54:52 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 62 ms on loc
alhost (1/2)
16/07/26 14:54:52 INFO PythonRunner: Times: total = 38, boot = 3, init = 32, finish = 3
16/07/26 14:54:52 INFO Executor: Finished task 1.0 in stage 1.0 (TID 3). 2179 bytes result se
nt to driver
16/07/26 14:54:52 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 73 ms on loc
alhost (2/2)
16/07/26 14:54:52 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed
, from pool
16/07/26 14:54:52 INFO DAGScheduler: ResultStage 1 (count at <stdin>:1) finished in 0.075 s
16/07/26 14:54:52 INFO DAGScheduler: Job 1 finished: count at <stdin>:1, took 0.094587 s
479
```

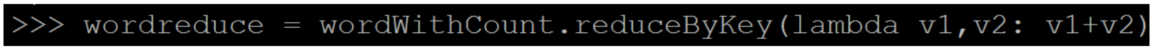
# Below is the script to find the top 10 and bottom 10 words of the dataset.



```
text = sc.textile(File("/user/lab/shakespeare.txt")
```

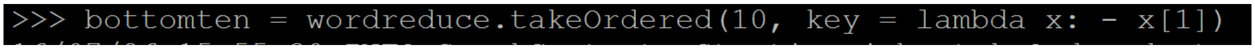


```
>>> words = text.flatMap(lambda line: line.split())
>>> wordWithCount = words.map(lambda word: (word,1))
```

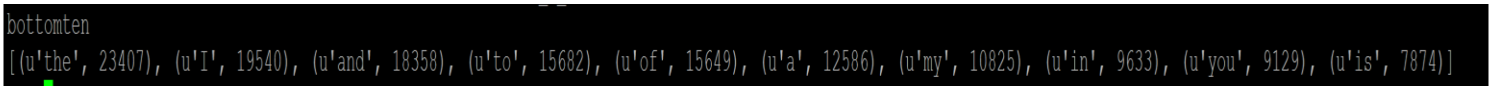


```
>>> wordreduce = wordWithCount.reduceByKey(lambda v1,v2: v1+v2)
```

The following shows the 10 words with the highest count (although the variable name indicates otherwise):

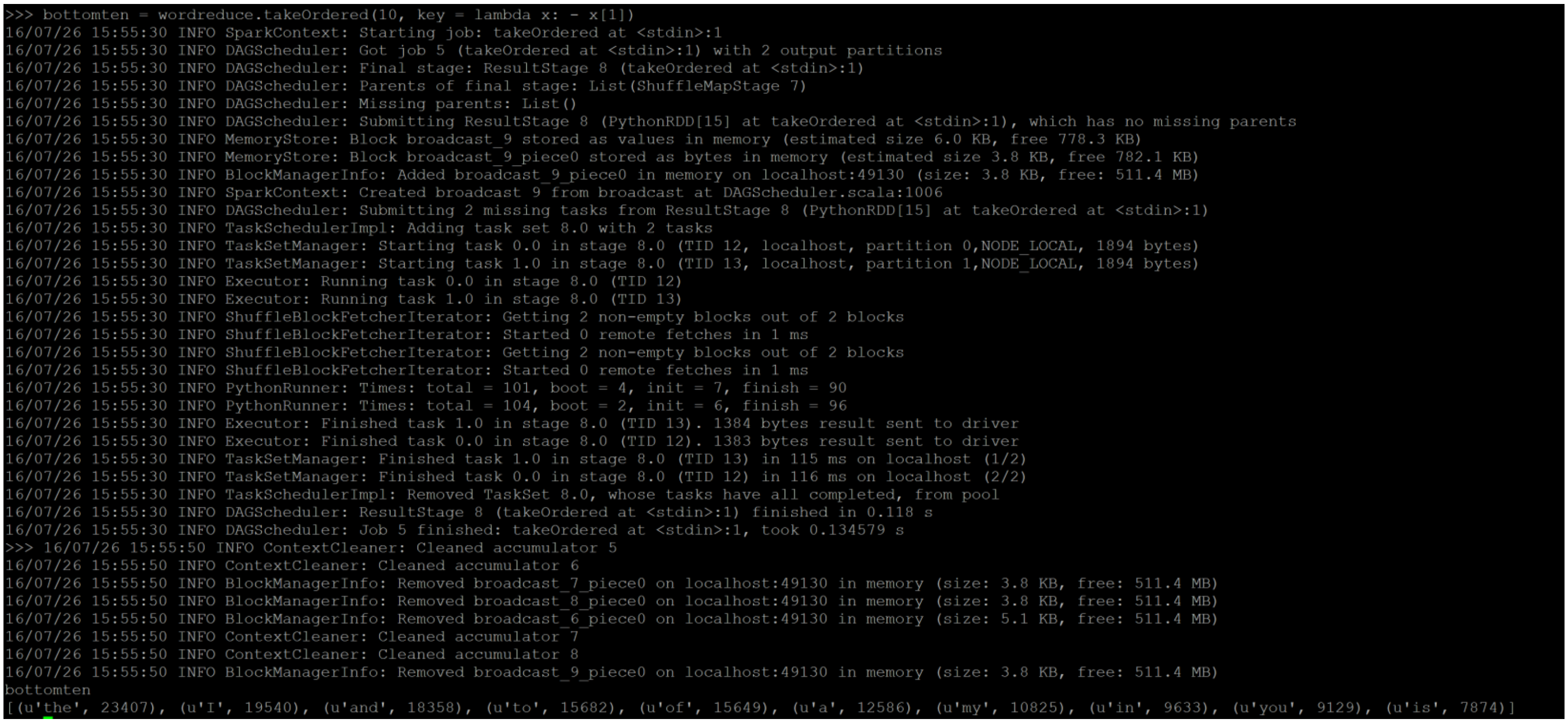


```
>>> bottomten = wordreduce.takeOrdered(10, key = lambda x: - x[1])
```



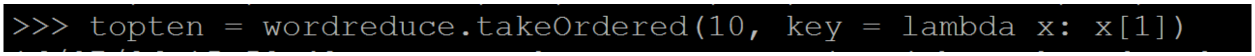
```
bottomten
[(u'the', 23407), (u'I', 19540), (u'and', 18358), (u'to', 15682), (u'of', 15649), (u'a', 12586), (u'my', 10825), (u'in', 9633), (u'you', 9129), (u'is', 7874)]
```

The following (below) shows the process that pyspark carried out and shows the answer as mentioned above:

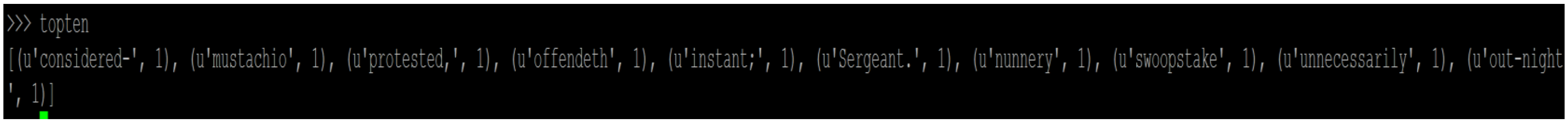


```
>>> bottomten = wordreduce.takeOrdered(10, key = lambda x: - x[1])
16/07/26 15:55:30 INFO SparkContext: Starting job: takeOrdered at <stdin>:1
16/07/26 15:55:30 INFO DAGScheduler: Got job 5 (takeOrdered at <stdin>:1) with 2 output partitions
16/07/26 15:55:30 INFO DAGScheduler: Final stage: ResultStage 8 (takeOrdered at <stdin>:1)
16/07/26 15:55:30 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 7)
16/07/26 15:55:30 INFO DAGScheduler: Missing parents: List()
16/07/26 15:55:30 INFO DAGScheduler: Submitting ResultStage 8 (PythonRDD[15] at takeOrdered at <stdin>:1), which has no missing parents
16/07/26 15:55:30 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 6.0 KB, free 778.3 KB)
16/07/26 15:55:30 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 3.8 KB, free 782.1 KB)
16/07/26 15:55:30 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:49130 (size: 3.8 KB, free: 511.4 MB)
16/07/26 15:55:30 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1006
16/07/26 15:55:30 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 8 (PythonRDD[15] at takeOrdered at <stdin>:1)
16/07/26 15:55:30 INFO TaskSchedulerImpl: Adding task set 8.0 with 2 tasks
16/07/26 15:55:30 INFO TaskSetManager: Starting task 0.0 in stage 8.0 (TID 12, localhost, partition 0,NODE_LOCAL, 1894 bytes)
16/07/26 15:55:30 INFO TaskSetManager: Starting task 1.0 in stage 8.0 (TID 13, localhost, partition 1,NODE_LOCAL, 1894 bytes)
16/07/26 15:55:30 INFO Executor: Running task 0.0 in stage 8.0 (TID 12)
16/07/26 15:55:30 INFO Executor: Running task 1.0 in stage 8.0 (TID 13)
16/07/26 15:55:30 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/07/26 15:55:30 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/07/26 15:55:30 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/07/26 15:55:30 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/07/26 15:55:30 INFO PythonRunner: Times: total = 101, boot = 4, init = 7, finish = 90
16/07/26 15:55:30 INFO PythonRunner: Times: total = 104, boot = 2, init = 6, finish = 96
16/07/26 15:55:30 INFO Executor: Finished task 1.0 in stage 8.0 (TID 13). 1384 bytes result sent to driver
16/07/26 15:55:30 INFO Executor: Finished task 0.0 in stage 8.0 (TID 12). 1383 bytes result sent to driver
16/07/26 15:55:30 INFO TaskSetManager: Finished task 1.0 in stage 8.0 (TID 13) in 115 ms on localhost (1/2)
16/07/26 15:55:30 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 12) in 116 ms on localhost (2/2)
16/07/26 15:55:30 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
16/07/26 15:55:30 INFO DAGScheduler: ResultStage 8 (takeOrdered at <stdin>:1) finished in 0.118 s
16/07/26 15:55:30 INFO DAGScheduler: Job 5 finished: takeOrdered at <stdin>:1, took 0.134579 s
>>> 16/07/26 15:55:50 INFO ContextCleaner: Cleaned accumulator 5
16/07/26 15:55:50 INFO ContextCleaner: Cleaned accumulator 6
16/07/26 15:55:50 INFO BlockManagerInfo: Removed broadcast_7_piece0 on localhost:49130 in memory (size: 3.8 KB, free: 511.4 MB)
16/07/26 15:55:50 INFO BlockManagerInfo: Removed broadcast_8_piece0 on localhost:49130 in memory (size: 3.8 KB, free: 511.4 MB)
16/07/26 15:55:50 INFO BlockManagerInfo: Removed broadcast_6_piece0 on localhost:49130 in memory (size: 5.1 KB, free: 511.4 MB)
16/07/26 15:55:50 INFO ContextCleaner: Cleaned accumulator 7
16/07/26 15:55:50 INFO ContextCleaner: Cleaned accumulator 8
16/07/26 15:55:50 INFO BlockManagerInfo: Removed broadcast_9_piece0 on localhost:49130 in memory (size: 3.8 KB, free: 511.4 MB)
bottomten
[(u'the', 23407), (u'I', 19540), (u'and', 18358), (u'to', 15682), (u'of', 15649), (u'a', 12586), (u'my', 10825), (u'in', 9633), (u'you', 9129), (u'is', 7874)]
```

The following shows the 10 words with the lowest count (although the variable name indicates otherwise).

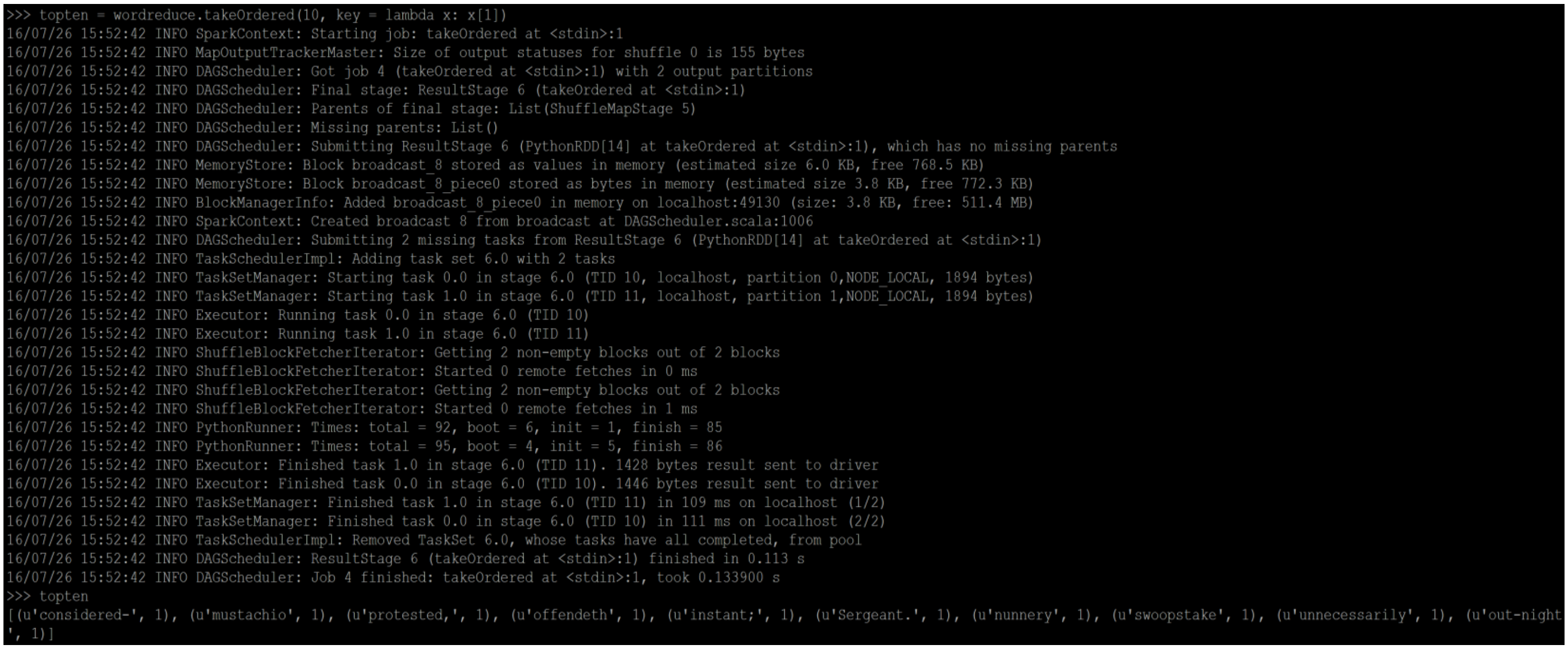


```
>>> topten = wordreduce.takeOrdered(10, key = lambda x: x[1])
```



```
>>> topten
[(u'considered-', 1), (u'mustachio', 1), (u'protested,', 1), (u'offendeth', 1), (u'instant;', 1), (u'Sergeant.', 1), (u'nunnery', 1), (u'swoopstake', 1), (u'unnecessarily', 1), (u'out-night
', 1)]
```

The following (below) shows the process that pyspark carried out and shows the answer as mentioned above:



```
>>> topten = wordreduce.takeOrdered(10, key = lambda x: x[1])
16/07/26 15:52:42 INFO SparkContext: Starting job: takeOrdered at <stdin>:1
16/07/26 15:52:42 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 155 bytes
16/07/26 15:52:42 INFO DAGScheduler: Got job 4 (takeOrdered at <stdin>:1) with 2 output partitions
16/07/26 15:52:42 INFO DAGScheduler: Final stage: ResultStage 6 (takeOrdered at <stdin>:1)
16/07/26 15:52:42 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 5)
16/07/26 15:52:42 INFO DAGScheduler: Missing parents: List()
16/07/26 15:52:42 INFO DAGScheduler: Submitting ResultStage 6 (PythonRDD[14] at takeOrdered at <stdin>:1), which has no missing parents
16/07/26 15:52:42 INFO MemoryStore: Block broadcast_8 stored as values in memory (estimated size 6.0 KB, free 768.5 KB)
16/07/26 15:52:42 INFO MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 3.8 KB, free 772.3 KB)
16/07/26 15:52:42 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on localhost:49130 (size: 3.8 KB, free: 511.4 MB)
16/07/26 15:52:42 INFO SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:1006
16/07/26 15:52:42 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 6 (PythonRDD[14] at takeOrdered at <stdin>:1)
16/07/26 15:52:42 INFO TaskSchedulerImpl: Adding task set 6.0 with 2 tasks
16/07/26 15:52:42 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 10, localhost, partition 0,NODE_LOCAL, 1894 bytes)
16/07/26 15:52:42 INFO TaskSetManager: Starting task 1.0 in stage 6.0 (TID 11, localhost, partition 1,NODE_LOCAL, 1894 bytes)
16/07/26 15:52:42 INFO Executor: Running task 0.0 in stage 6.0 (TID 10)
16/07/26 15:52:42 INFO Executor: Running task 1.0 in stage 6.0 (TID 11)
16/07/26 15:52:42 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/07/26 15:52:42 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/07/26 15:52:42 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/07/26 15:52:42 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/07/26 15:52:42 INFO PythonRunner: Times: total = 92, boot = 6, init = 1, finish = 85
16/07/26 15:52:42 INFO PythonRunner: Times: total = 95, boot = 4, init = 5, finish = 86
16/07/26 15:52:42 INFO Executor: Finished task 1.0 in stage 6.0 (TID 11). 1428 bytes result sent to driver
16/07/26 15:52:42 INFO Executor: Finished task 0.0 in stage 6.0 (TID 10). 1446 bytes result sent to driver
16/07/26 15:52:42 INFO TaskSetManager: Finished task 1.0 in stage 6.0 (TID 11) in 109 ms on localhost (1/2)
16/07/26 15:52:42 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 10) in 111 ms on localhost (2/2)
16/07/26 15:52:42 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
16/07/26 15:52:42 INFO DAGScheduler: ResultStage 6 (takeOrdered at <stdin>:1) finished in 0.113 s
16/07/26 15:52:42 INFO DAGScheduler: Job 4 finished: takeOrdered at <stdin>:1, took 0.133900 s
>>> topten
[(u'considered-', 1), (u'mustachio', 1), (u'protested,', 1), (u'offendeth', 1), (u'instant;', 1), (u'Sergeant.', 1), (u'nunnery', 1), (u'swoopstake', 1), (u'unnecessarily', 1), (u'out-night
', 1)]
```