```
OpenJDK 64-Bit Server VM (build 25.432-b09, mixed mode)
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ echo -e "hello world hello\nspark is awesome\nthis is a test spark world\nhello spark world\nhello" > sample.txt
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ nano word_count.py
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ spark-submit word_count.py
25/05/20 11:39:52 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
25/05/20 11:39:52 INFO SparkContext: Running Spark version 3.0.3
25/05/20 11:39:52 INFO ResourceUtils: ==============================================================
25/05/20 11:39:52 INFO ResourceUtils: Resources for spark.driver:

25/05/20 11:39:52 INFO ResourceUtils: ==============================================================
25/05/20 11:39:52 INFO SparkContext: Submitted application: SimpleWordCount
25/05/20 11:39:53 INFO SecurityManager: Changing view acls to: bmscecse
25/05/20 11:39:53 INFO SecurityManager: Changing modify acls to: bmscecse
25/05/20 11:39:53 INFO SecurityManager: Changing view acls groups to:
25/05/20 11:39:53 INFO SecurityManager: Changing modify acls groups to:
25/05/20 11:39:53 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(bmscecse); groups with view permissions: Set(); users  with modify permissions: Set(bmscecse); groups with modify permissions: Set()
25/05/20 11:39:53 INFO Utils: Successfully started service 'sparkDriver' on port 34445.
25/05/20 11:39:53 INFO SparkEnv: Registering MapOutputTracker
25/05/20 11:39:53 INFO SparkEnv: Registering BlockManagerMaster
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
25/05/20 11:39:53 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
25/05/20 11:39:53 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-ede53b67-2dce-41cc-ae20-7cc9c45cde11
25/05/20 11:39:53 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
25/05/20 11:39:53 INFO SparkEnv: Registering OutputCommitCoordinator
25/05/20 11:39:53 INFO Utils: Successfully started service 'SparkUI' on port 4040.
25/05/20 11:39:53 INFO SparkUI: Bound SparkUI to 127.0.0.1, and started at http://localhost:4040
25/05/20 11:39:53 INFO Executor: Starting executor ID driver on host localhost
25/05/20 11:39:53 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 34187.
25/05/20 11:39:53 INFO NettyBlockTransferService: Server created on localhost:34187
25/05/20 11:39:53 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
25/05/20 11:39:53 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO BlockManagerMasterEndpoint: Registering block manager localhost:34187 with 366.3 MiB RAM, BlockManagerId(driver, localhost, 34187, None)
```

```
25/05/20 11:39:53 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, localhost, 34187, None)
25/05/20 11:39:53 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 241.7 KiB, free 366.1 MiB)
25/05/20 11:39:53 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.4 KiB, free 366.0 MiB)
25/05/20 11:39:53 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:34187 (size: 23.4 KiB, free: 366.3 MiB)
25/05/20 11:39:53 INFO SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
25/05/20 11:39:54 INFO FileInputFormat: Total input paths to process : 1
25/05/20 11:39:54 INFO SparkContext: Starting job: collect at /home/bmscecse/word_count.py:16
25/05/20 11:39:54 INFO DAGScheduler: Registering RDD 3 (reduceByKey at /home/bmscecse/word_count.py:12) as input to shuffle 0
25/05/20 11:39:54 INFO DAGScheduler: Got job 0 (collect at /home/bmscecse/word_count.py:16) with 1 output partitions
25/05/20 11:39:54 INFO DAGScheduler: Final stage: ResultStage 1 (collect at /home/bmscecse/word_count.py:16)
25/05/20 11:39:54 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
25/05/20 11:39:54 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)
25/05/20 11:39:54 INFO DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/bmscecse/word_count.py:12), which has no missing parents
25/05/20 11:39:54 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 11.5 KiB, free 366.0 MiB)
25/05/20 11:39:54 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 7.0 KiB, free 366.0 MiB)
25/05/20 11:39:54 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on localhost:34187 (size: 7.0 KiB, free: 366.3 MiB)
25/05/20 11:39:54 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1223
25/05/20 11:39:54 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/bmscecse/word_count.py:12) (first 15 tasks are for partitions Vector(0))
25/05/20 11:39:54 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
25/05/20 11:39:54 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 7360 bytes)
25/05/20 11:39:54 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
25/05/20 11:39:54 INFO HadoopRDD: Input split: file:/home/bmscecse/sample.txt:0+86
25/05/20 11:39:54 INFO PythonRunner: Times: total = 212, boot = 166, init = 46, finish = 0
25/05/20 11:39:55 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1803 bytes result sent to driver
25/05/20 11:39:55 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 837 ms on localhost (executor driver) (1/1)
25/05/20 11:39:55 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
25/05/20 11:39:55 INFO PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 58787
25/05/20 11:39:55 INFO DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/bmscecse/word_count.py:12) finished in 0.905 s
25/05/20 11:39:55 INFO DAGScheduler: looking for newly runnable stages
25/05/20 11:39:55 INFO DAGScheduler: running: Set()
25/05/20 11:39:55 INFO DAGScheduler: waiting: Set(ResultStage 1)
25/05/20 11:39:55 INFO DAGScheduler: failed: Set()
25/05/20 11:39:55 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:16), which has no miss
```

```
25/05/20 11:39:55 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:16), which has no miss
ing parents
25/05/20 11:39:55 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 8.7 KiB, free 366.0 MiB)
25/05/20 11:39:55 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 5.2 KiB, free 366.0 MiB)
25/05/20 11:39:55 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:34187 (size: 5.2 KiB, free: 366.3 MiB)
25/05/20 11:39:55 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1223
25/05/20 11:39:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at collect at /home/bmscecse/word_count.py:1
6) (first 15 tasks are for partitions Vector(0))
25/05/20 11:39:55 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
25/05/20 11:39:55 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, NODE_LOCAL, 7143 bytes
)
25/05/20 11:39:55 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
25/05/20 11:39:55 INFO ShuffleBlockFetcherIterator: Getting 1 (156.0 B) non-empty blocks including 1 (156.0 B) local and 0 (0.0 B) host-local
 and 0 (0.0 B) remote blocks
25/05/20 11:39:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
25/05/20 11:39:55 INFO PythonRunner: Times: total = 42, boot = -518, init = 560, finish = 0
25/05/20 11:39:55 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 1767 bytes result sent to driver
25/05/20 11:39:55 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 67 ms on localhost (executor driver) (1/1)
25/05/20 11:39:55 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
25/05/20 11:39:55 INFO DAGScheduler: ResultStage 1 (collect at /home/bmscecse/word_count.py:16) finished in 0.073 s
25/05/20 11:39:55 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
25/05/20 11:39:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
25/05/20 11:39:55 INFO DAGScheduler: Job 0 finished: collect at /home/bmscecse/word_count.py:16, took 1.006372 s
25/05/20 11:39:55 INFO SparkUI: Stopped Spark web UI at http://localhost:4040
25/05/20 11:39:55 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
25/05/20 11:39:55 INFO MemoryStore: MemoryStore cleared
25/05/20 11:39:55 INFO BlockManager: BlockManager stopped
25/05/20 11:39:55 INFO BlockManagerMaster: BlockManagerMaster stopped
25/05/20 11:39:55 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
25/05/20 11:39:55 INFO SparkContext: Successfully stopped SparkContext
25/05/20 11:39:56 INFO ShutdownHookManager: Shutdown hook called
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-aa3fe98f-0602-4d46-9aa7-736cc067dbc4
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-46383acd-daf9-43ba-9b6c-337b1f545cc0
25/05/20 11:39:56 INFO ShutdownHookManager: Deleting directory /tmp/spark-46383acd-daf9-43ba-9b6c-337b1f545cc0/pyspark-bab4954c-8903-4094-94e
c-b517c5aef7db
```

```
  GNU nano 6.2                        word_count.py
from pyspark import SparkContext

# Initialize SparkContext
sc = SparkContext("local", "SimpleWordCount")

# Read the file (use the temporary sample.txt)
rdd = sc.textFile("sample.txt")  # Use the sample file created above

# Count words
counts = (rdd.flatMap(lambda line: line.split())          # Split each li>
             .map(lambda word: (word, 1))                 # Map each word >
             .reduceByKey(lambda a, b: a + b)            # Reduce by key t>
             .filter(lambda x: x[1] > 4))                # Filter words wi>

# Show result
for word, count in counts.collect():
    print(word, count)

# Stop the SparkContext
sc.stop()




                        [ Read 20 lines ]
^G Help       ^O Write Out  ^W Where Is   ^K Cut        ^T Execute
^X Exit       ^R Read File  ^\ Replace    ^U Paste      ^J Justify
```