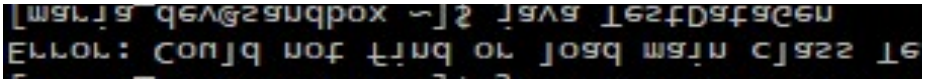
CSP 595 - Assignment 6

Name: Jagruti Vichare
Email ID: jvichare@hawk.iit.edu
CWID: A20378092

Question 1

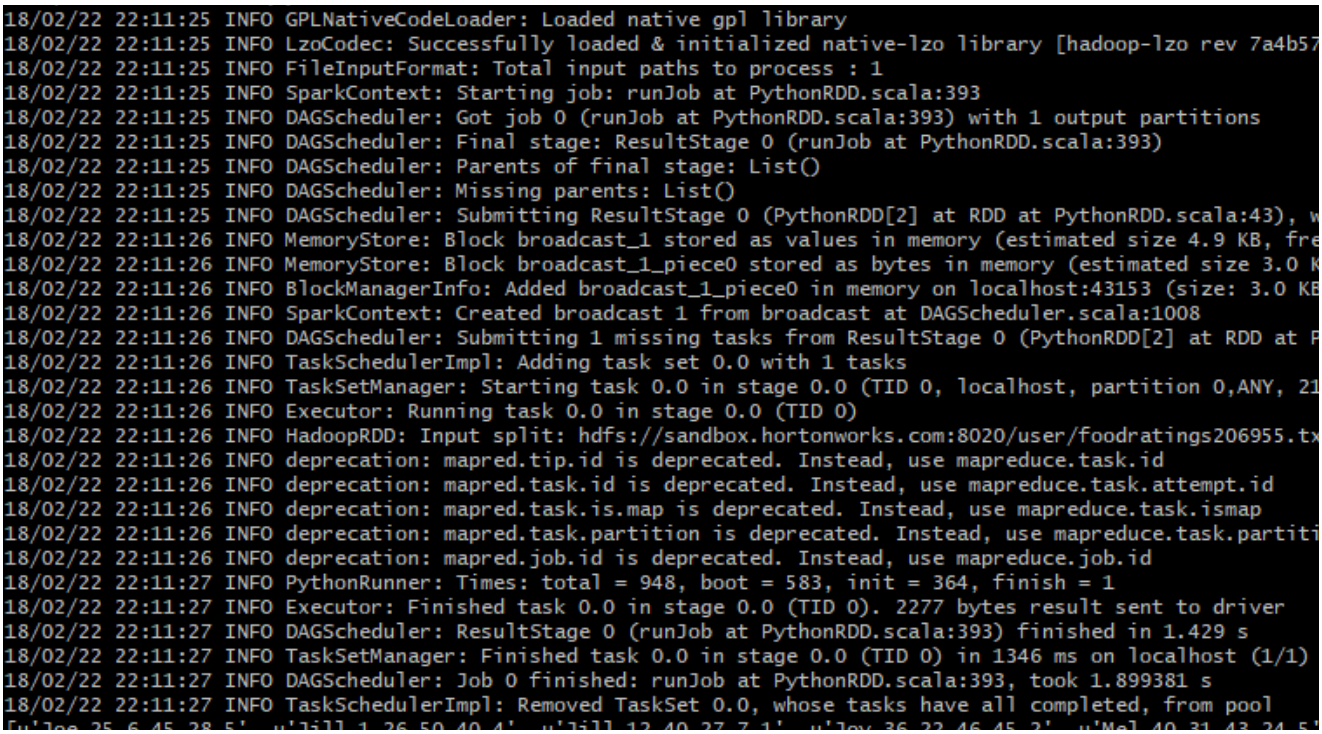


Commands:

ex1RDD=sc.textFile('/user/foodratings206955.txt')
print ex1RDD.take(5)

Output:

[u'Joe,25,6,45,28,5', u'Jill,1,26,50,40,4', u'Jill,12,40,27,7,1', u'Joy,36,22,46,45,2', u'Mel,40,31,43,24,5']



Question 2:

Commands:

ex2RDD = ex1RDD.map(lambda line: line.split(“,”))
print ex2RDD.take(5)

Output:

[[u'Joe', u'25', u'6', u'45', u'28', u'5'], [u'Jill', u'1', u'26', u'50', u'40', u'4'], [u'Jill', u'12', u'40', u'27', u'7', u'1'], [u'Joy', u'36', u'22', u'46', u'45', u'2'], [u'Mel', u'40', u'31', u'43', u'24', u'5']]

```
18/02/22 22:29:55 INFO DAGScheduler: Got job 3 (runJob at PythonRDD.scala:393) with 1 output partitions
18/02/22 22:29:55 INFO DAGScheduler: Final stage: ResultStage 3 (runJob at PythonRDD.scala:393)
18/02/22 22:29:55 INFO DAGScheduler: Parents of final stage: List()
18/02/22 22:29:55 INFO DAGScheduler: Missing parents: List()
18/02/22 22:29:55 INFO DAGScheduler: Submitting ResultStage 3 (PythonRDD[7] at RDD at PythonRDD.scala:43), which has no missing parents
18/02/22 22:29:55 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 5.3 KB, free 771.1 KB)
18/02/22 22:29:55 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 3.4 KB, free 774.5 KB)
18/02/22 22:29:55 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on localhost:43153 (size: 3.4 KB, free: 511.1 MB)
18/02/22 22:29:55 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:1008
18/02/22 22:29:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (PythonRDD[7] at RDD at PythonRDD.scala:43)
18/02/22 22:29:55 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
18/02/22 22:29:55 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, localhost, partition 0,ANY, 2164 bytes)
18/02/22 22:29:55 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)
18/02/22 22:29:55 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/foodratings206955.txt:0+8735
18/02/22 22:29:55 INFO PythonRunner: Times: total = 23, boot = 5, init = 17, finish = 1
18/02/22 22:29:55 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 2494 bytes result sent to driver
18/02/22 22:29:55 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 110 ms on localhost (1/1)
18/02/22 22:29:55 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
18/02/22 22:29:55 INFO DAGScheduler: ResultStage 3 (runJob at PythonRDD.scala:393) finished in 0.131 s
18/02/22 22:29:55 INFO DAGScheduler: Job 3 finished: runJob at PythonRDD.scala:393, took 0.170822 s
[[u'Joe', u'25', u'6', u'45', u'28', u'5'], [u'Jill', u'1', u'26', u'50', u'40', u'4'], [u'Jill', u'12', u'40', u'27', u'7', u'1'], [u'Joy', u'36', u'22', u'46', u'45', u'2'], [u'Mel', u'40',
u'31', u'43', u'24', u'5']]
```

Question 3:

Commands:

ex3RDD = ex2RDD.map(lambda line: [line[0],line[1],int(line[2]),line[3],line[4],line[5]])
print ex3RDD.take(5)

Output:

[[u'Joe', u'25', 6, u'45', u'28', u'5'], [u'Jill', u'1', 26, u'50', u'40', u'4'], [u'Jill', u'12', 40, u'27', u'7', u'1'], [u'Joy', u'36', 22, u'46', u'45', u'2'], [u'Mel', u'40', 31, u'43', u'24', u'5']]

```
18/02/22 23:21:14 INFO SparkContext: Starting job: runJob at PythonRDD.scala:393
18/02/22 23:21:14 INFO DAGScheduler: Got job 9 (runJob at PythonRDD.scala:393) with 1 output partitions
18/02/22 23:21:14 INFO DAGScheduler: Final stage: ResultStage 9 (runJob at PythonRDD.scala:393)
18/02/22 23:21:14 INFO DAGScheduler: Parents of final stage: List()
18/02/22 23:21:14 INFO DAGScheduler: Missing parents: List()
18/02/22 23:21:14 INFO DAGScheduler: Submitting ResultStage 9 (PythonRDD[13] at RDD at PythonRDD.scala:43), which has no missing parents
18/02/22 23:21:14 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 5.6 KB, free 782.6 KB)
18/02/22 23:21:14 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.6 KB, free 786.2 KB)
18/02/22 23:21:14 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:43153 (size: 3.6 KB, free: 511.1 MB)
18/02/22 23:21:14 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1008
18/02/22 23:21:14 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 9 (PythonRDD[13] at RDD at PythonRDD.scala:43)
18/02/22 23:21:14 INFO TaskSchedulerImpl: Adding task set 9.0 with 1 tasks
18/02/22 23:21:14 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 9, localhost, partition 0,ANY, 2164 bytes)
18/02/22 23:21:14 INFO Executor: Running task 0.0 in stage 9.0 (TID 9)
18/02/22 23:21:14 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/foodratings206955.txt:0+8735
18/02/22 23:21:14 INFO PythonRunner: Times: total = 10, boot = 2, init = 8, finish = 0
18/02/22 23:21:14 INFO Executor: Finished task 0.0 in stage 9.0 (TID 9). 2462 bytes result sent to driver
18/02/22 23:21:14 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 9) in 36 ms on localhost (1/1)
18/02/22 23:21:14 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool
18/02/22 23:21:15 INFO DAGScheduler: ResultStage 9 (runJob at PythonRDD.scala:393) finished in 0.040 s
18/02/22 23:21:15 INFO DAGScheduler: Job 9 finished: runJob at PythonRDD.scala:393, took 0.050670 s
[[u'Joe', u'25', 6, u'45', u'28', u'5'], [u'Jill', u'1', 26, u'50', u'40', u'4'], [u'Jill', u'12', 40, u'27', u'7', u'1'], [u'Joy', u'36', 22, u'46', u'45', u'2'], [u'Mel', u'40', 31, u'43', u
'24', u'5']]
```

Question 4

Commands:

ex4RDD = ex3RDD.filter(lambda line: line[2]<25)
print ex4RDD.take(5)

Output:

[[u'Joe', u'25', 6, u'45', u'28', u'5'], [u'Joy', u'36', 22, u'46', u'45', u'2'], [u'Jill', u'38', 12, u'16', u'42', u'3'], [u'Mel', u'2', 14, u'26', u'40', u'5'], [u'Joy', u'8', 17, u'25', u'47', u'2']]

```
18/02/22 23:25:44 INFO SparkContext: Starting job: runJob at PythonRDD.scala:393
18/02/22 23:25:44 INFO DAGScheduler: Got job 10 (runJob at PythonRDD.scala:393) with 1 output partitions
18/02/22 23:25:44 INFO DAGScheduler: Final stage: ResultStage 10 (runJob at PythonRDD.scala:393)
18/02/22 23:25:44 INFO DAGScheduler: Parents of final stage: List()
18/02/22 23:25:44 INFO DAGScheduler: Missing parents: List()
18/02/22 23:25:44 INFO DAGScheduler: Submitting ResultStage 10 (PythonRDD[14] at RDD at PythonRDD.scala:43), which has no missing parents
18/02/22 23:25:44 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 5.9 KB, free 792.1 KB)
18/02/22 23:25:44 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 3.8 KB, free 795.8 KB)
18/02/22 23:25:44 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on localhost:43153 (size: 3.8 KB, free: 511.0 MB)
18/02/22 23:25:44 INFO SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:1008
18/02/22 23:25:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 10 (PythonRDD[14] at RDD at PythonRDD.scala:43)
18/02/22 23:25:44 INFO TaskSchedulerImpl: Adding task set 10.0 with 1 tasks
18/02/22 23:25:44 INFO TaskSetManager: Starting task 0.0 in stage 10.0 (TID 10, localhost, partition 0,ANY, 2164 bytes)
18/02/22 23:25:44 INFO Executor: Running task 0.0 in stage 10.0 (TID 10)
18/02/22 23:25:44 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/foodratings206955.txt:0+8735
18/02/22 23:25:44 INFO PythonRunner: Times: total = 11, boot = 4, init = 7, finish = 0
18/02/22 23:25:44 INFO Executor: Finished task 0.0 in stage 10.0 (TID 10). 2461 bytes result sent to driver
18/02/22 23:25:44 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 10) in 111 ms on localhost (1/1)
18/02/22 23:25:44 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
18/02/22 23:25:44 INFO DAGScheduler: ResultStage 10 (runJob at PythonRDD.scala:393) finished in 0.117 s
18/02/22 23:25:44 INFO DAGScheduler: Job 10 finished: runJob at PythonRDD.scala:393, took 0.138023 s
[[u'Joe', u'25', 6, u'45', u'28', u'5'], [u'Joy', u'36', 22, u'46', u'45', u'2'], [u'Jill', u'38', 12, u'16', u'42', u'3'], [u'Mel', u'2', 14, u'26', u'40', u'5'], [u'Joy', u'8', 17, u'25', u'
47', u'2']]
```

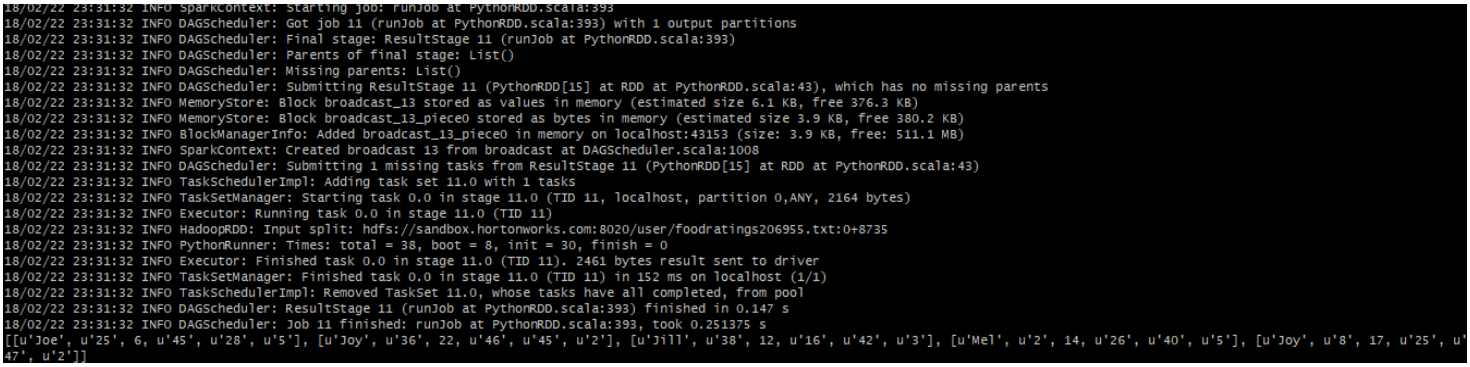
Question 5

Commands:

ex5RDD = ex4RDD.filter(lambda line: (line[0],line))
print ex5RDD.take(5)

Output:

[[u'Joe', u'25', 6, u'45', u'28', u'5'], [u'Joy', u'36', 22, u'46', u'45', u'2'], [u'Jill', u'38', 12, u'16', u'42', u'3'], [u'Mel', u'2', 14, u'26', u'40', u'5'], [u'Joy', u'8', 17, u'25', u'47', u'2']]



Question 6

Commands:

ex6RDD = ex5RDD.sortByKey(True,1)
print ex6RDD.take(5)

Output:

[[u'Jill', u'38', 12, u'16', u'42', u'3'], [u'Jill', u'31', 18, u'18', u'25', u'4'], [u'Jill', u'38', 5, u'11', u'33', u'2'], [u'Jill', u'36', 10, u'26', u'18', u'3'], [u'Jill', u'4', 13, u'6', u'47', u'4']]