

Management of Big Data and Tools – DS8003 – Fall 2016

Assignment 05

NAJLIS, BERNARDO - Student Number #500744793

Submission:

Submit a single doc/pdf file that has python code (with comments) and a English description of what your code is doing. You can either perform the commands on "pyspark" console or type commands into python file and use "spark-submit"

Submit into Assignment folder titled "Assignment 5: Spark"

1. ODD/EVEN NUMBER (3)

- Input: From a file containing a list of numbers
 1. Input file is "integer_list" that you downloaded from D2L and uploaded to VirtualBox.
 2. Copy the file to "/user/root" in the hdfs
- Output: Count the number of odd numbers and even numbers

2. SALARY SUM PER DEPARTMENT (3)

- Input: Department[SPACE]Salary
 1. Input file is "dept_salary" that you downloaded from D2L and uploaded to VirtualBox.
 2. Copy the file to "/user/root" in the hdfs
- Output: Department Name and Salary Sum

3. Top K and bottom K words (4)

- Input: shakespeare_100.txt
- Output: 10 words with most count and 10 words with least count
- Simply limit by 10 in ascending and descending order of count. You do not have to worry about edge-cases where the 10th word and 11th word have the same count.

Resolution

1. numbers =
sc.textFile("hdfs://sandbox.hortonworks.com:8020/user/root/assgn5/integer_list.txt")
odddnumbers = numbers.filter(lambda odd: int(odd)%2 == 1)
odddnumbers.count()
evennumbers = numbers.filter(lambda even: int(even)%2 == 0)
evennumbers.count()

```
16/11/11 17:57:21 INFO DAGScheduler: Job 5 finished: count at <stdin>:1, took 0.072405 s
491
>>> evennumbers = numbers.filter(lambda even: int(even)%2 == 0)
>>> evennumbers.count()
16/11/11 17:57:21 INFO SparkContext: Starting job: count at <stdin>:1
16/11/11 17:57:21 INFO DAGScheduler: Got job 6 (count at <stdin>:1) with 2 output partitions
16/11/11 17:57:21 INFO DAGScheduler: Final stage: ResultStage 6 (count at <stdin>:1)
16/11/11 17:57:21 INFO DAGScheduler: Parents of final stage: List()
16/11/11 17:57:21 INFO DAGScheduler: Missing parents: List()
16/11/11 17:57:21 INFO DAGScheduler: Submitting ResultStage 6 (PythonRDD[13] at count at <stdin>:1), which has no missing parents
16/11/11 17:57:21 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 5.9 KB, free 802.8 KB)
16/11/11 17:57:21 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 3.7 KB, free 806.6 KB)
16/11/11 17:57:21 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:45688 (size: 3.7 KB, free: 511.4 MB)
16/11/11 17:57:21 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1006
16/11/11 17:57:21 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 6 (PythonRDD[13] at count at <stdin>:1)
16/11/11 17:57:21 INFO TaskSchedulerImpl: Adding task set 6.0 with 2 tasks
16/11/11 17:57:21 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 12, localhost, partition 0,ANY, 2171 bytes)
16/11/11 17:57:21 INFO TaskSetManager: Starting task 1.0 in stage 6.0 (TID 13, localhost, partition 1,ANY, 2171 bytes)
16/11/11 17:57:21 INFO Executor: Running task 0.0 in stage 6.0 (TID 12)
16/11/11 17:57:21 INFO Executor: Running task 1.0 in stage 6.0 (TID 13)
16/11/11 17:57:21 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/root/assgn5/integer_list.txt:2832+2833
16/11/11 17:57:21 INFO HadoopRDD: Input split: hdfs://sandbox.hortonworks.com:8020/user/root/assgn5/integer_list.txt:0+2832
16/11/11 17:57:22 INFO PythonRunner: Times: total = 43, boot = -48, init = 89, finish = 2
16/11/11 17:57:22 INFO Executor: Finished task 1.0 in stage 6.0 (TID 13). 2180 bytes result sent to driver
16/11/11 17:57:22 INFO TaskSetManager: Finished task 1.0 in stage 6.0 (TID 13) in 60 ms on localhost (1/2)
16/11/11 17:57:22 INFO PythonRunner: Times: total = 43, boot = -39, init = 80, finish = 2
16/11/11 17:57:22 INFO Executor: Finished task 0.0 in stage 6.0 (TID 12). 2179 bytes result sent to driver
16/11/11 17:57:22 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 12) in 66 ms on localhost (2/2)
16/11/11 17:57:22 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
16/11/11 17:57:22 INFO DAGScheduler: ResultStage 6 (count at <stdin>:1) finished in 0.066 s
16/11/11 17:57:22 INFO DAGScheduler: Job 6 finished: count at <stdin>:1, took 0.081778 s
510
```

491 odd numbers, 510 even numbers

2. salaries =
sc.textFile("hdfs://sandbox.hortonworks.com:8020/user/root/assgn5/dept_salary.txt")
mapped_salaries = salaries.map(lambda line: line.split(" "))
grouped_salaries = mapped_salaries.map(lambda fields: (fields[0], int(fields[1])))
aggregated_salaries = grouped_salaries.reduceByKey(lambda x, y: x + y)

```
root@sandbox:~  
511.4 MB)  
16/11/11 18:28:05 INFO SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:1006  
16/11/11 18:28:05 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 9 (PythonRDD[17] at collect at <stdin>:1)  
16/11/11 18:28:05 INFO TaskSchedulerImpl: Adding task set 9.0 with 2 tasks  
16/11/11 18:28:05 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 18, localhost, partition 0,NODE_LOCAL, 189  
4 bytes)  
16/11/11 18:28:05 INFO TaskSetManager: Starting task 1.0 in stage 9.0 (TID 19, localhost, partition 1,NODE_LOCAL, 189  
4 bytes)  
16/11/11 18:28:05 INFO Executor: Running task 0.0 in stage 9.0 (TID 18)  
16/11/11 18:28:05 INFO Executor: Running task 1.0 in stage 9.0 (TID 19)  
16/11/11 18:28:05 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks  
16/11/11 18:28:05 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms  
16/11/11 18:28:05 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks  
16/11/11 18:28:05 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms  
16/11/11 18:28:05 INFO PythonRunner: Times: total = 41, boot = -46, init = 87, finish = 0  
16/11/11 18:28:05 INFO Executor: Finished task 1.0 in stage 9.0 (TID 19). 1361 bytes result sent to driver  
16/11/11 18:28:05 INFO TaskSetManager: Finished task 1.0 in stage 9.0 (TID 19) in 57 ms on localhost (1/2)  
16/11/11 18:28:05 INFO PythonRunner: Times: total = 45, boot = -37, init = 82, finish = 0  
16/11/11 18:28:05 INFO Executor: Finished task 0.0 in stage 9.0 (TID 18). 1258 bytes result sent to driver  
16/11/11 18:28:05 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 18) in 68 ms on localhost (2/2)  
16/11/11 18:28:05 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have all completed, from pool  
16/11/11 18:28:05 INFO DAGScheduler: ResultStage 9 (collect at <stdin>:1) finished in 0.066 s  
16/11/11 18:28:05 INFO DAGScheduler: Job 7 finished. Collect at <stdin>:1, took 0.137565 s  
[(u'QA', 3360624), (u'Marketing', 3158454), (u'Research', 3333284), (u'Sales', 3471781), (u'Developer', 3221394)]
```

```
QA - 3360624  
Marketing - 3158454  
Research - 3333284  
Sales - 3471781  
Developer - 3221394
```

3. `document =`
`sc.textFile("hdfs://sandbox.hortonworks.com:8020/user/root/assgn5`
`/shakespeare_100.txt")`
`words = document.flatMap(lambda w: w.split(" "))`
`wcount = words.map(lambda w: (w, 1))`
`cntbyword = wcount.reduceByKey(lambda x, y : x + y)`
`top10 = cntbyword.top(10, key=lambda i: i[1])`
`print(top10)`
`btm10 = cntbyword.top(10, key=lambda i: -i[1])`
`print(btm10)`

```

root@sandbox:~
g parents
16/11/11 19:18:23 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 6.0 KB, free 275.6 KB)
16/11/11 19:18:23 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 3.8 KB, free 279.4 KB)
16/11/11 19:18:23 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:35096 (size: 3.8 KB, free: 511.5 MB)
16/11/11 19:18:23 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1006
16/11/11 19:18:23 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 3 (PythonRDD[7] at top at <stdin>:1)
16/11/11 19:18:23 INFO TaskSchedulerImpl: Adding task set 3.0 with 2 tasks
16/11/11 19:18:23 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 4, localhost, partition 0,NODE_LOCAL, 1894 bytes)
16/11/11 19:18:23 INFO TaskSetManager: Starting task 1.0 in stage 3.0 (TID 5, localhost, partition 1,NODE_LOCAL, 1894 bytes)
16/11/11 19:18:23 INFO Executor: Running task 0.0 in stage 3.0 (TID 4)
16/11/11 19:18:23 INFO Executor: Running task 1.0 in stage 3.0 (TID 5)
16/11/11 19:18:23 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/11/11 19:18:23 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/11/11 19:18:23 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks out of 2 blocks
16/11/11 19:18:23 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/11/11 19:18:23 INFO PythonRunner: Times: total = 100, boot = -67286, init = 67289, finish = 97
16/11/11 19:18:23 INFO Executor: Finished task 1.0 in stage 3.0 (TID 5). 1428 bytes result sent to driver
16/11/11 19:18:23 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 5) in 114 ms on localhost (1/2)
16/11/11 19:18:23 INFO BlockManagerInfo: Removed broadcast_2_piece0 on localhost:35096 in memory (size: 3.8 KB, free: 511.5 MB)
16/11/11 19:18:23 INFO ContextCleaner: Cleaned accumulator 3
16/11/11 19:18:23 INFO BlockManagerInfo: Removed broadcast_1_piece0 on localhost:35096 in memory (size: 5.1 KB, free: 511.5 MB)
16/11/11 19:18:23 INFO PythonRunner: Times: total = 131, boot = -67273, init = 67281, finish = 123
16/11/11 19:18:23 INFO ContextCleaner: Cleaned accumulator 2
16/11/11 19:18:23 INFO Executor: Finished task 0.0 in stage 3.0 (TID 4). 1443 bytes result sent to driver
16/11/11 19:18:23 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 4) in 145 ms on localhost (2/2)
16/11/11 19:18:23 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/11/11 19:18:23 INFO DAGScheduler: ResultStage 3 (top at <stdin>:1) finished in 0.147 s
16/11/11 19:18:23 INFO DAGScheduler: Job 1 finished: top at <stdin>:1, took 0.170753 s

>>> print(btm10)
[(u'mustachio', 1), (u'protested', 1), (u'offendeth', 1), (u'instant', 1), (u'Sergeant.', 1), (u'unnery', 1), (u'swoopstake', 1), (u'unneccessarily', 1), (u'out-night', 1), (u'Fiend,', 1)]
>>> print(top10)
[(u'', 506610), (u'the', 23407), (u'I', 19450), (u'and', 18358), (u'to', 15682), (u'of', 15649), (u'a', 12586), (u'my', 10825), (u'in', 9633), (u'you', 9129)]
>>>

```

Top 10 words

```

'', 506610
'the', 23407
'I', 19450
'and', 18358
'to', 15682
'of', 15649
'a', 12586
'my', 10825
'in', 9633
'you', 9129

```

Bottom 10 words

```

'mustachio', 1
'protested', 1
'offendeth', 1
'instant', 1
'Sergeant', 1
'unnery', 1
'swoopstake', 1
'unneccessarily', 1
'out-night', 1
'Fiend', 1

```