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)

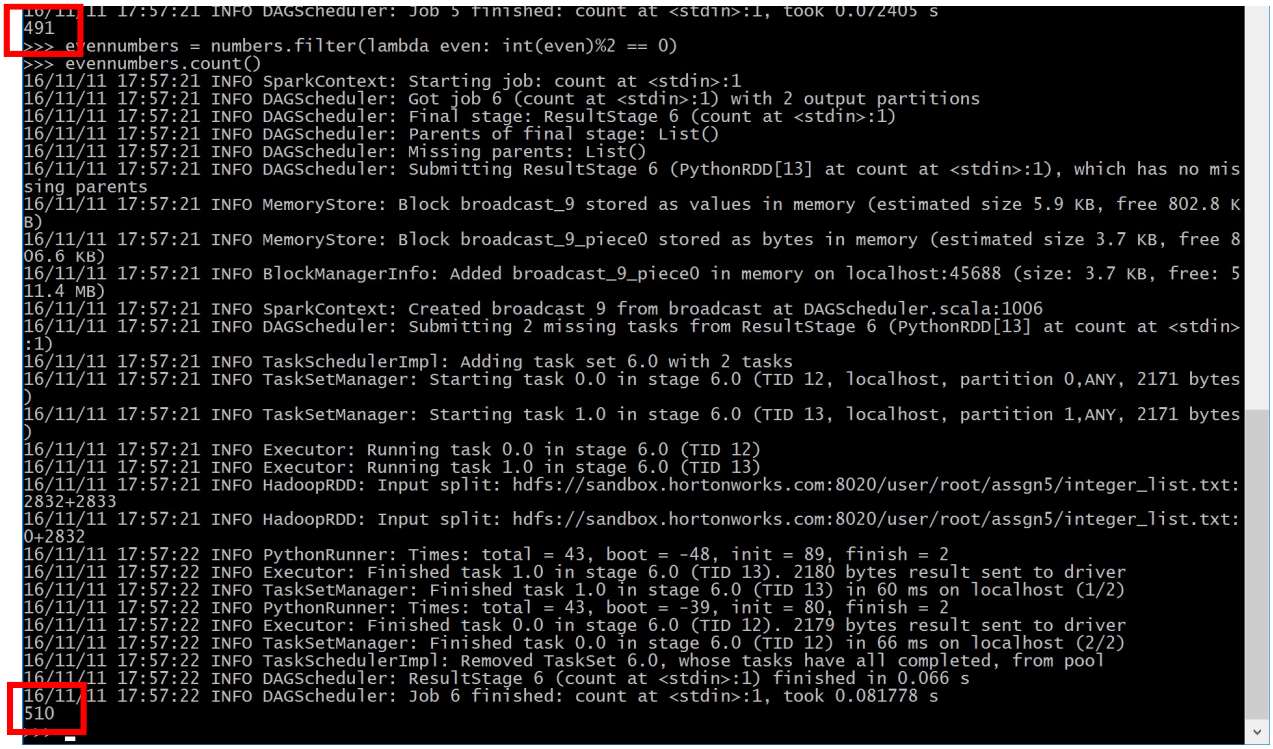
- Input: Department[SPACE]Salary
    1. Input file is "dept_salary" that you downloaded from D2L and uploaded to VirtualBox.
    2. Copy the file to "/user/root" in the hdfs
- Output: Department Name and Salary Sum

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.

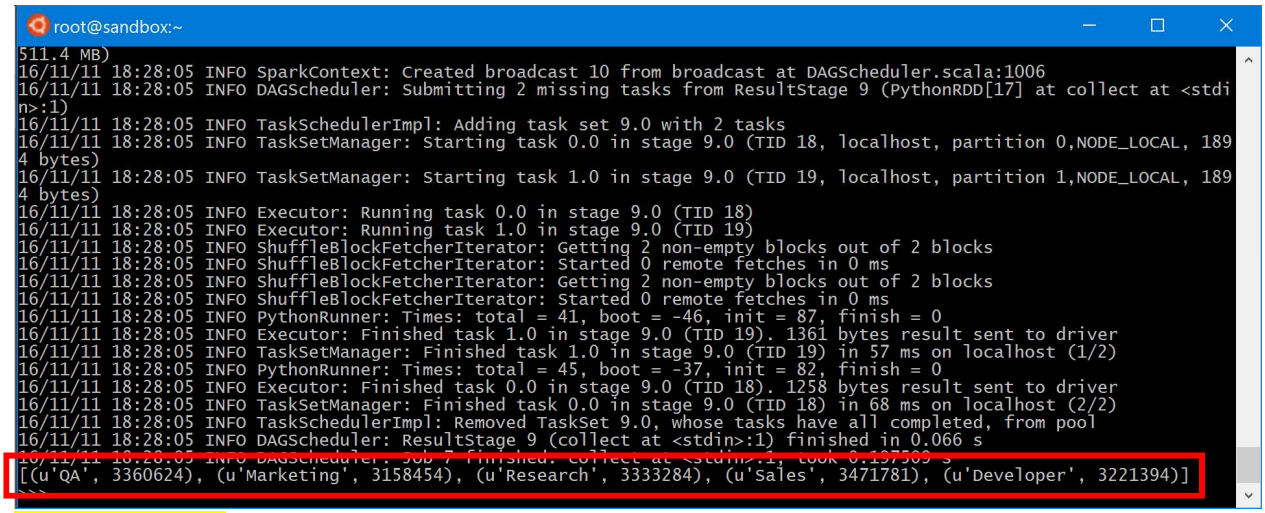
Resolution

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



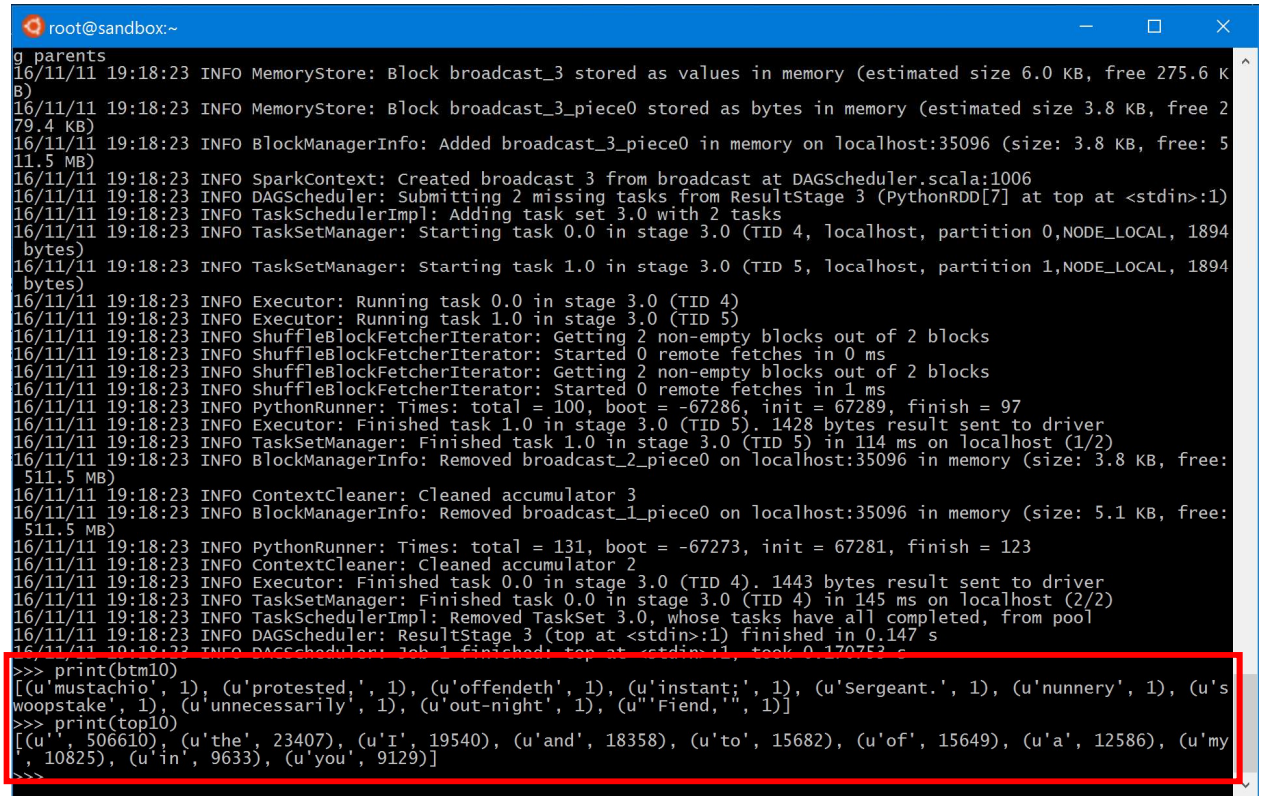
491 odd numbers, 510 even numbers

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)
```



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)
```



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
‘nunnery’, 1
‘swoopstake’, 1
‘unnecessarily’, 1
‘out-night’, 1
‘Fiend’, 1