Paul Ycay

500709618

CIND 719

Instructor: Sebnem Kuzulugil

Fall 2019

Assignment 3 – Spark

Download the input files from Resources/Spark Resources section of the course page and upload to your VM.
Copy the files to /user/lab/ in the HDFS.
If you decide to use the file on your local system instead of HDFS, please state this in your submission file.

First, go into Resources/Spark in Course shell, and download all files. Transfer it to VM. Make sure you have a working /user/lab folder in HDFS (check Lab 2 instructions). Transfer files to hdfs from VM by going into putty, and typing:

[@sandbox]# hadoop fs -put /home/lab/shakespeare.txt /user/lab ---make sure you have a /home/lab folder by creating a directory under home in VM, then changing directory to lab by typing cd /home/lab in putty

```
[@sandbox ]# hadoop fs -put /home/lab/wordCount.py /user/lab
```

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

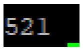
```
[@sandbox ]# hadoop fs -put /home/lab/dept_salary.txt
```

```
[@sandbox ]# hadoop fs -put /home/lab/full_text.txt /user/lab
```

1. ODD/EVEN NUMBER (30 pts)
(Hint: Note that you are reading the file as text and need to convert the numbers to int())

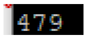
Input: number_list.txt (a list of 1000 integers)
Output: Count the number of odd numbers and even numbers in the file

```
>>> number_list = sc.textFile(“/user/lab/number_list.txt”) ---read the file from directory
>>> evens = number_list.map(lambda x: int(x) % 2 == 0)
>>> even.sum() ---count even numbers
```



```
521
```

```
>>> odds = number_list.map(lambda x: int(x) % 2 != 0)
>>> odds.sum() --- count odd numbers
```



```
479
```

```
2. Top K and bottom K words (30 pts)
(Hint: Search and use takeOrdered() method)


Input: shakespeare.txt
Output: 10 words with the highest count and 10 words with lowest count
```

>>>shakespeare_count = sc.textFile(“/user/lab/shakespeare.txt”) \ --- hit return

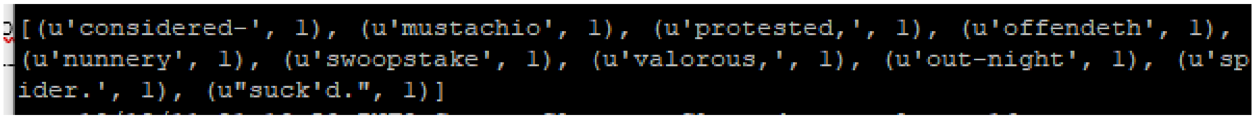
>>>.flatMap(lambda line: line.lower().split()) \ --- return

>>>.map(lambda word: (word, 1)) \ --- return

>>>.reduceByKey(lambda a, b: a+b) --- return

--- Output 10 words with the lowest count

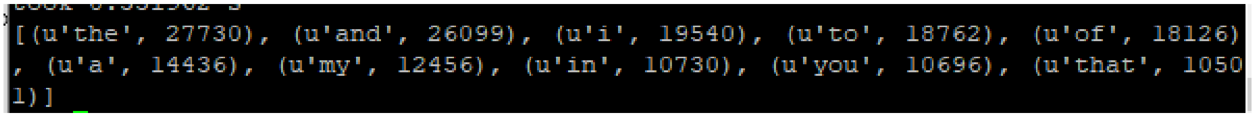
>>> shakespeare_count.takeOrdered(10, lambda x: x[1])



```
[(u'considered-', 1), (u'mustachio', 1), (u'protested,', 1), (u'offendeth', 1),
(u'nunnery', 1), (u'swoopstake', 1), (u'valorous,', 1), (u'out-night', 1), (u'sp
ider.', 1), (u"suck'd.", 1)]
```

--- output 10 words with the highest count

>>> shakespeare_count.takeOrdered(10, lambda x: -x[1])



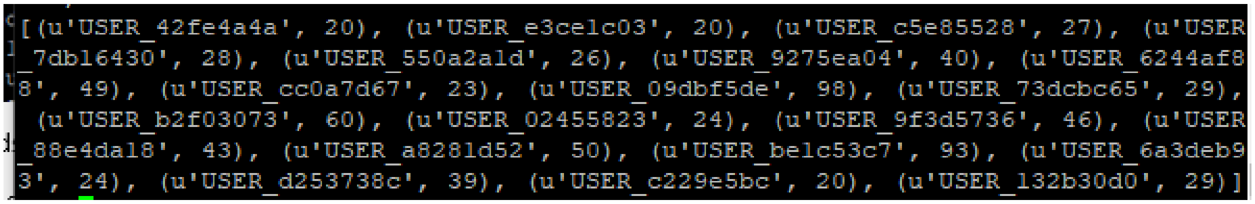
```
[(u'the', 27730), (u'and', 26099), (u'i', 19540), (u'to', 18762), (u'of', 18126)
, (u'a', 14436), (u'my', 12456), (u'in', 10730), (u'you', 10696), (u'that', 1050
1)]
```

```
3. Group and Count (40 pts)

Input: fulltext_txt
Output: Count the number of tweets for each user_id and save the results in a
text file.

SUBMIT YOUR SCRIPT AND THE OUTPUT OF YOUR SCRIPT.
```

>>> tweets = sc.textFile(“/user/lab/full_text.txt”)\ ---hit return
>>>.map(lambda line: (line.split(‘\t’)[0], 1))\ --- return
>>>.reduceByKey(lambda a, b: a+b)

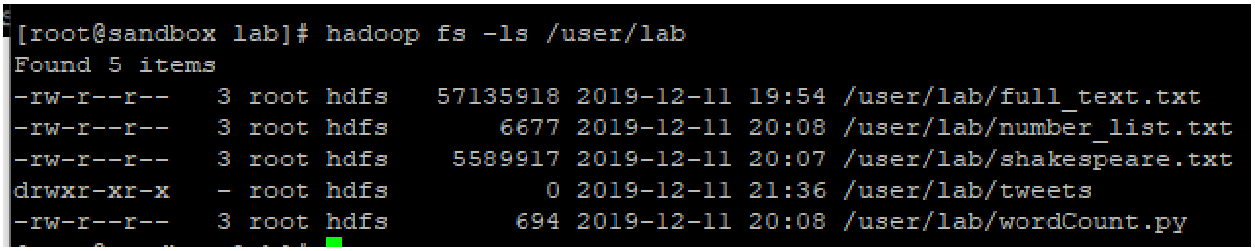
--- Extract tweet number’s in first 20 users

>>>tweets.take(20)



```
[(u'USER_42fe4a4a', 20), (u'USER_e3ce1c03', 20), (u'USER_c5e85528', 27), (u'USER
_7db16430', 28), (u'USER_550a2a1d', 26), (u'USER_9275ea04', 40), (u'USER_6244af8
8', 49), (u'USER_cc0a7d67', 23), (u'USER_09dbf5de', 98), (u'USER_73dcbc65', 29),
 (u'USER_b2f03073', 60), (u'USER_02455823', 24), (u'USER_9f3d5736', 46), (u'USER
_88e4da18', 43), (u'USER_a8281d52', 50), (u'USER_be1c53c7', 93), (u'USER_6a3deb9
3', 24), (u'USER_d253738c', 39), (u'USER_c229e5bc', 20), (u'USER_132b30d0', 29)]
```

--- Save script and output
>>>tweets.saveAsTextFile(“/user/lab/tweets”)
[@sandbox]# hadoop fs -ls /user/lab



```
[root@sandbox lab]# hadoop fs -ls /user/lab
Found 5 items
-rw-r--r--   3 root hdfs   57135918 2019-12-11 19:54 /user/lab/full_text.txt
-rw-r--r--   3 root hdfs       6677 2019-12-11 20:08 /user/lab/number_list.txt
-rw-r--r--   3 root hdfs    5589917 2019-12-11 20:07 /user/lab/shakespeare.txt
drwxr-xr-x   - root hdfs          0 2019-12-11 21:36 /user/lab/tweets
-rw-r--r--   3 root hdfs        694 2019-12-11 20:08 /user/lab/wordCount.py
```