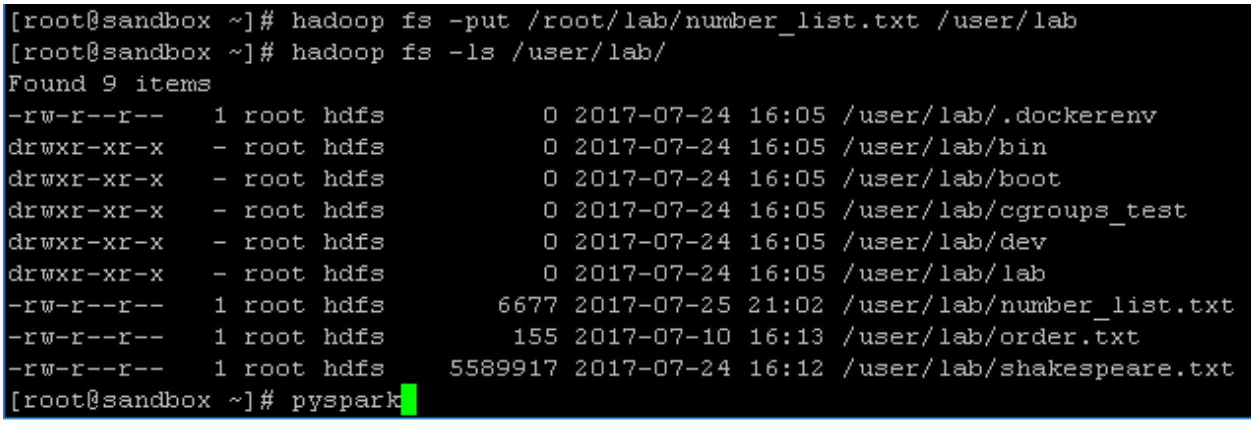
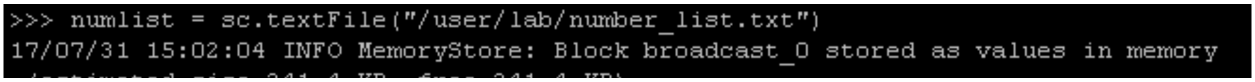
Assignment 4

1. ODD/EVEN NUMBER (30 pts)

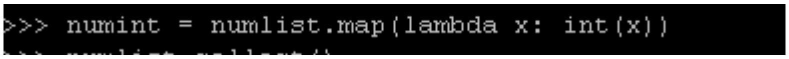


```
[root@sandbox ~]# hadoop fs -put /root/lab/number_list.txt /user/lab
[root@sandbox ~]# hadoop fs -ls /user/lab/
Found 9 items
-rw-r--r--   1 root hdfs          0 2017-07-24 16:05 /user/lab/.dockerenv
drwxr-xr-x   - root hdfs          0 2017-07-24 16:05 /user/lab/bin
drwxr-xr-x   - root hdfs          0 2017-07-24 16:05 /user/lab/boot
drwxr-xr-x   - root hdfs          0 2017-07-24 16:05 /user/lab/cgroups_test
drwxr-xr-x   - root hdfs          0 2017-07-24 16:05 /user/lab/dev
drwxr-xr-x   - root hdfs          0 2017-07-24 16:05 /user/lab/lab
-rw-r--r--   1 root hdfs       6677 2017-07-25 21:02 /user/lab/number_list.txt
-rw-r--r--   1 root hdfs        155 2017-07-10 16:13 /user/lab/order.txt
-rw-r--r--   1 root hdfs    5589917 2017-07-24 16:12 /user/lab/shakespeare.txt
[root@sandbox ~]# pyspark
```

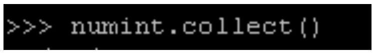


```
>>> numlist = sc.textFile("/user/lab/number_list.txt")
17/07/31 15:02:04 INFO MemoryStore: Block broadcast_0 stored as values in memory
```

convert the numbers to int



```
>>> numint = numlist.map(lambda x: int(x))
```

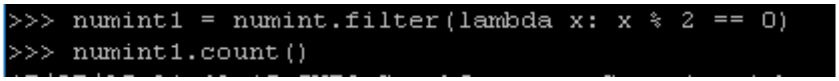


```
>>> numint.collect()
```

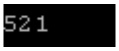
Input: number_list.txt (a list of 1000 integers)

Output: Count the number of odd numbers and even numbers in the file

Even numbers.

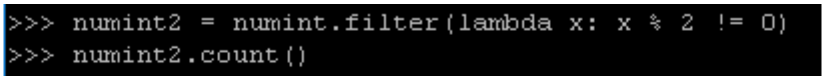


```
>>> numint1 = numint.filter(lambda x: x % 2 == 0)
>>> numint1.count()
```

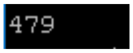


```
521
```

Odd numbers

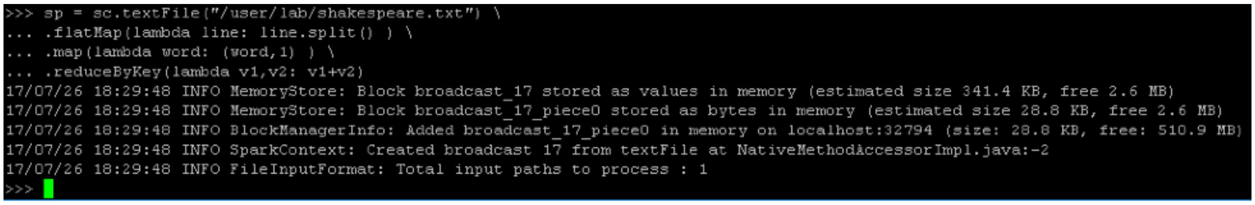


```
>>> numint2 = numint.filter(lambda x: x % 2 != 0)
>>> numint2.count()
```



```
479
```

2. Top K and bottom K words (30 pts)

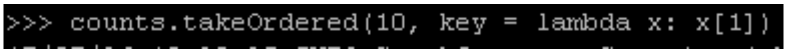


```
>>> sp = sc.textFile("/user/lab/shakespeare.txt") \
... .flatMap(lambda line: line.split() ) \
... .map(lambda word: (word,1) ) \
... .reduceByKey(lambda v1,v2: v1+v2)
17/07/26 18:29:48 INFO MemoryStore: Block broadcast_17 stored as values in memory (estimated size 341.4 KB, free 2.6 MB)
17/07/26 18:29:48 INFO MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimated size 28.8 KB, free 2.6 MB)
17/07/26 18:29:48 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on localhost:32794 (size: 28.8 KB, free: 510.9 MB)
17/07/26 18:29:48 INFO SparkContext: Created broadcast 17 from textFile at NativeMethodAccessorImpl.java:-2
17/07/26 18:29:48 INFO FileInputFormat: Total input paths to process : 1
>>> 
```

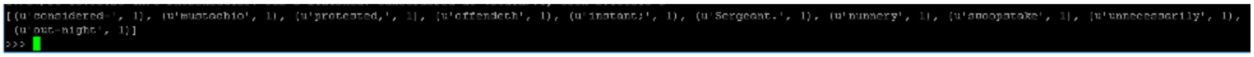
Input: shakespeare.txt

Output: 10 words with the highest count and 10 words with lowest count

Bottom

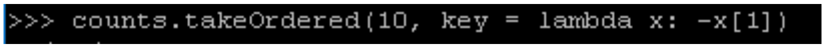


```
>>> counts.takeOrdered(10, key = lambda x: x[1])
```

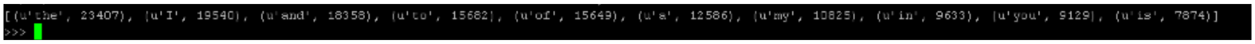


```
[(u'considered-', 1), (u'mustachio', 1), (u'protested,', 1), (u'offendeth', 1), (u'instant;', 1), (u'Sergeant.', 1), (u'nunnery', 1), (u'sweepstake', 1), (u'unnecessarily', 1),
 (u'out-night', 1)]
>>> 
```

Top



```
>>> counts.takeOrdered(10, key = lambda x: -x[1])
```

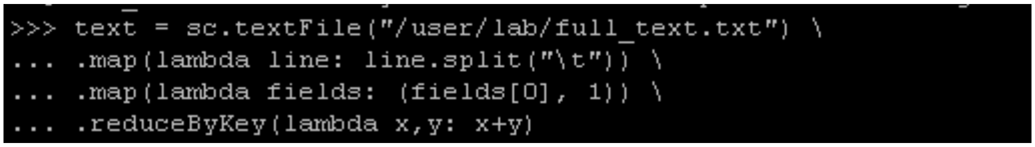


```
[(u'the', 23407), (u'I', 19540), (u'and', 18358), (u'to', 15682), (u'of', 15649), (u'a', 12586), (u'my', 10825), (u'in', 9633), (u'you', 9129), (u'is', 7874)]
>>> 
```

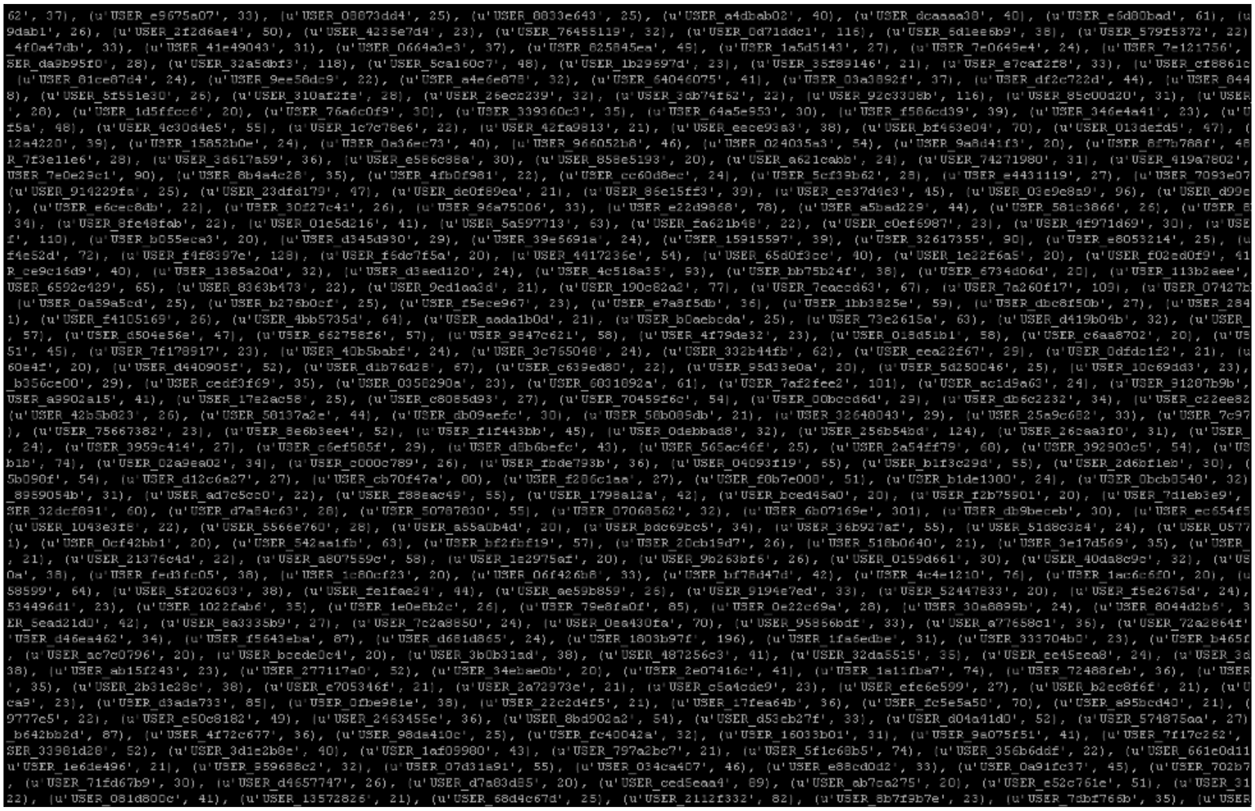
3. Group and Count (40 pts)

Input: full_text_txt

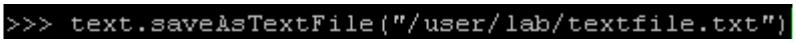
Output: Count the number of tweets for each user_id and save the results in a text file.



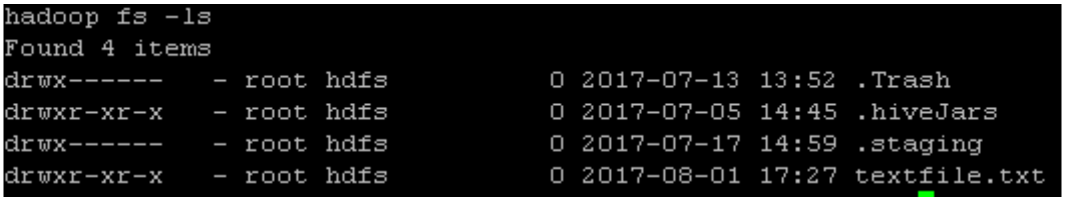
```
>>> text = sc.textFile("/user/lab/full_text.txt") \
... .map(lambda line: line.split("\t")) \
... .map(lambda fields: (fields[0], 1)) \
... .reduceByKey(lambda x,y: x+y)
```



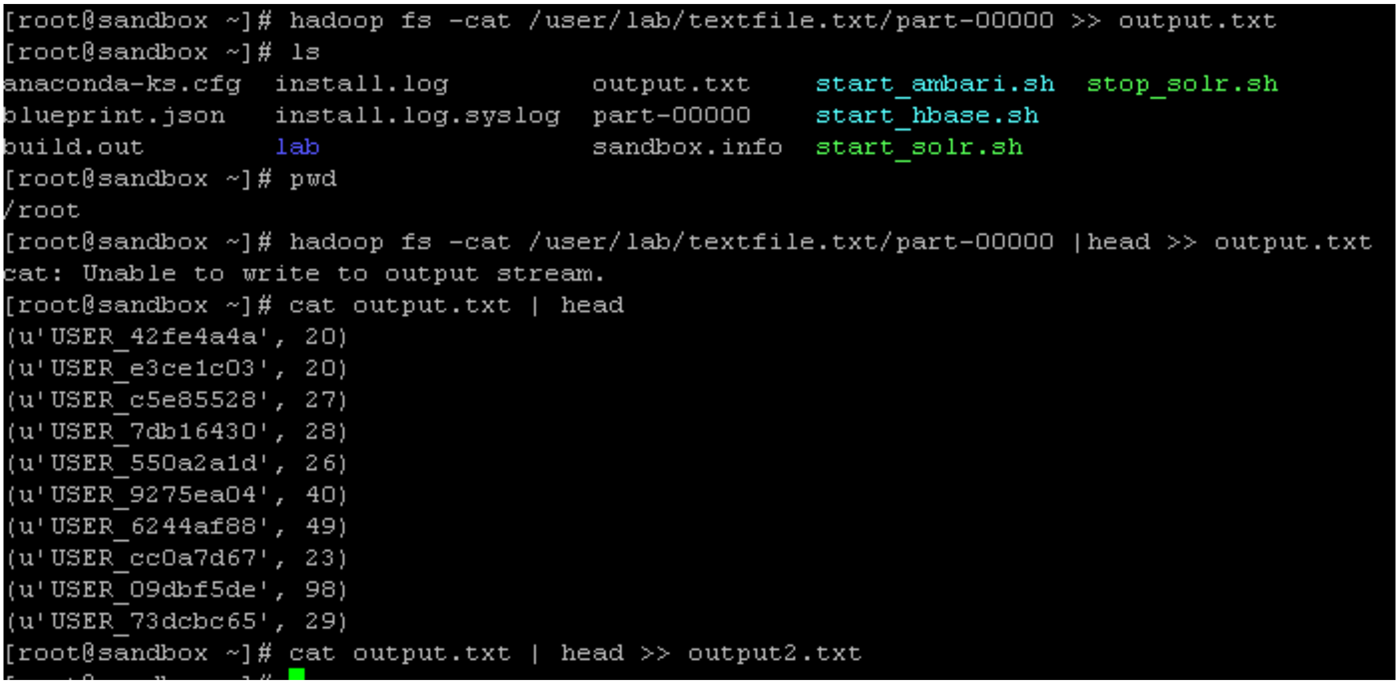
```
62', 37), (u'USER_e9675a07', 33), (u'USER_08873dd4', 25), (u'USER_0833e643', 25), (u'USER_a4dbab02', 40), (u'USER_dcaaaa38', 40), (u'USER_e6d80bad', 61), (u
9dab1', 26), (u'USER_2f2d6ae4', 50), (u'USER_4235e7d4', 23), (u'USER_76455119', 32), (u'USER_0d71ddc1', 116), (u'USER_6d1ee6b9', 38), (u'USER_579f5372', 22)
_4f0a47db', 33), (u'USER_41e49043', 31), (u'USER_0664a3e3', 37), (u'USER_825845ea', 49), (u'USER_1a5d5143', 27), (u'USER_7e0649e4', 24), (u'USER_7e121756',
SER_da9b95f0', 28), (u'USER_32a5dbf3', 118), (u'USER_5ca160c7', 48), (u'USER_1b29697d', 23), (u'USER_35f89146', 21), (u'USER_e7caf2f8', 33), (u'USER_cf8861c
 (u'USER_81ce87d4', 24), (u'USER_9ee58dc9', 22), (u'USER_a4e6e878', 32), (u'USER_64046075', 41), (u'USER_03a3892f', 37), (u'USER_df2c722d', 44), (u'USER_844
8), (u'USER_5f551e30', 26), (u'USER_310af2fe', 28), (u'USER_26ecb239', 32), (u'USER_3db74f62', 22), (u'USER_92c3308b', 116), (u'USER_85c00d20', 31), (u'USER
', 28), (u'USER_1d5ffcc6', 20), (u'USER_76a6c0f9', 30), (u'USER_339360c3', 35), (u'USER_64a5e953', 30), (u'USER_f586cd39', 39), (u'USER_346e4a41', 23), (u'U
f5a', 48), (u'USER_4c30d4e5', 55), (u'USER_1c7c78e6', 22), (u'USER_42fa9813', 21), (u'USER_eece93a3', 38), (u'USER_bf463e04', 70), (u'USER_013defd5', 47), (
12a4220', 39), (u'USER_15852b0e', 24), (u'USER_0a36ec73', 40), (u'USER_966052b8', 46), (u'USER_024035a3', 54), (u'USER_9a8d41f3', 20), (u'USER_8f7b788f', 48
R_7f3e11e6', 28), (u'USER_3d617a59', 36), (u'USER_e586c88a', 30), (u'USER_858e5193', 20), (u'USER_a621cabb', 24), (u'USER_74271980', 31), (u'USER_419a7802',
USER_7e0e29c1', 90), (u'USER_8b4a4c28', 35), (u'USER_4fb0f981', 22), (u'USER_cc60d8ec', 24), (u'USER_5cf39b62', 28), (u'USER_e4431119', 27), (u'USER_7093e07
(u'USER_914229fa', 25), (u'USER_23dfd179', 47), (u'USER_de0f89ea', 21), (u'USER_86e15ff3', 39), (u'USER_ee37d4e3', 45), (u'USER_03e9e8a9', 96), (u'USER_d99e
), (u'USER_e6cec8db', 22), (u'USER_30f27c41', 26), (u'USER_96a75006', 33), (u'USER_e22d9868', 78), (u'USER_a5bad229', 44), (u'USER_581c3866', 26), (u'USER_8
 34), (u'USER_8fe48fab', 22), (u'USER_01e5d216', 41), (u'USER_5a597713', 63), (u'USER_fa621b48', 22), (u'USER_c0ef6987', 23), (u'USER_4f971d69', 30), (u'USE
f', 110), (u'USER_b055eca3', 20), (u'USER_d345d930', 29), (u'USER_39e5691e', 24), (u'USER_15915597', 39), (u'USER_32617355', 90), (u'USER_e8053214', 25), (u
f4e52d', 72), (u'USER_f4f8397e', 128), (u'USER_f6dc7f5a', 20), (u'USER_4417236e', 54), (u'USER_65d0f3cc', 40), (u'USER_1e22f6a5', 20), (u'USER_f02ed0f9', 41
R_ce9c16d9', 40), (u'USER_1385a20d', 32), (u'USER_d3aed120', 24), (u'USER_4c518a35', 93), (u'USER_bb75b24f', 38), (u'USER_6734d06d', 20), (u'USER_113b2aee',
USER_6592c429', 65), (u'USER_8363b473', 22), (u'USER_9cd1aa3d', 21), (u'USER_190c82a2', 77), (u'USER_7eaccd63', 67), (u'USER_7a260f17', 109), (u'USER_07427b
 (u'USER_0a59a5cd', 25), (u'USER_b276b0cf', 25), (u'USER_f5ece967', 23), (u'USER_e7a8f5db', 36), (u'USER_1bb3825e', 59), (u'USER_dbc8f50b', 27), (u'USER_284
1), (u'USER_f4105169', 26), (u'USER_4bb5735d', 64), (u'USER_aada1b0d', 21), (u'USER_b0acbcda', 25), (u'USER_73e2615a', 63), (u'USER_d419b04b', 32), (u'USER_
, 57), (u'USER_d504e56e', 47), (u'USER_662758f6', 57), (u'USER_9847c621', 58), (u'USER_4f79de32', 23), (u'USER_018d51b1', 58), (u'USER_c6aa8702', 20), (u'US
51', 45), (u'USER_7f178917', 23), (u'USER_40b5babf', 24), (u'USER_3c765048', 24), (u'USER_332b44fb', 62), (u'USER_eea22f67', 29), (u'USER_0dfdc1f2', 21), (u
60e4f', 20), (u'USER_d440905f', 52), (u'USER_d1b76d28', 67), (u'USER_c639ed80', 22), (u'USER_95d33e0a', 20), (u'USER_5d250046', 25), (u'USER_10c69dd3', 23),
_b356ce00', 29), (u'USER_cedf3f69', 35), (u'USER_0358290a', 23), (u'USER_6031092a', 61), (u'USER_7af2fee2', 101), (u'USER_ac1d9a63', 24), (u'USER_91207b9b',
USER_a9902a15', 41), (u'USER_17e2ac58', 25), (u'USER_c8085d93', 27), (u'USER_70459f6c', 54), (u'USER_00bccd6d', 29), (u'USER_db6c2232', 34), (u'USER_c22ee82
(u'USER_42b5b823', 26), (u'USER_50137a2e', 44), (u'USER_db09aefc', 30), (u'USER_58b089db', 21), (u'USER_32648043', 29), (u'USER_25a9c682', 33), (u'USER_7c97
), (u'USER_75667382', 23), (u'USER_8e6b3ee4', 52), (u'USER_f1f443bb', 45), (u'USER_0debbad8', 32), (u'USER_256b54bd', 124), (u'USER_26caa3f0', 31), (u'USER_
, 24), (u'USER_3959c414', 27), (u'USER_c6ef585f', 29), (u'USER_d8b6befc', 43), (u'USER_565ac46f', 25), (u'USER_2a54ff79', 60), (u'USER_392903c5', 54), (u'US
b1b', 74), (u'USER_02a9ea02', 34), (u'USER_c000c789', 26), (u'USER_fbde793b', 36), (u'USER_04093f19', 65), (u'USER_b1f3c29d', 55), (u'USER_2d6bf1eb', 30), (
5b090f', 54), (u'USER_d12c6a27', 27), (u'USER_cb70f47a', 80), (u'USER_f286c1ae', 27), (u'USER_f0b7e008', 51), (u'USER_b1de1300', 24), (u'USER_0bcb8540', 32)
_8959054b', 31), (u'USER_ad7c5cc0', 22), (u'USER_f88eac49', 55), (u'USER_1798a12a', 42), (u'USER_bced45a0', 20), (u'USER_f2b75901', 20), (u'USER_7d1eb3e9',
SER_32dcf891', 60), (u'USER_d7a84c63', 28), (u'USER_50787830', 55), (u'USER_07068562', 32), (u'USER_6b07169e', 301), (u'USER_db9beceb', 30), (u'USER_ec654f5
(u'USER_1043e3f8', 22), (u'USER_5566e760', 28), (u'USER_a55a0b4d', 20), (u'USER_bdc69bc5', 34), (u'USER_36b927af', 55), (u'USER_51d8c3b4', 24), (u'USER_0577
1), (u'USER_0cf42bb1', 20), (u'USER_542aa1fb', 63), (u'USER_bf2fbf19', 57), (u'USER_20cb19d7', 26), (u'USER_518b0640', 21), (u'USER_3e17d569', 35), (u'USER_
, 21), (u'USER_21376c4d', 22), (u'USER_a807559c', 58), (u'USER_1e2975af', 20), (u'USER_9b263bf6', 26), (u'USER_0159d661', 30), (u'USER_40da8c9c', 32), (u'US
0a', 38), (u'USER_fed3fc05', 38), (u'USER_1c80cf23', 20), (u'USER_06f426b8', 33), (u'USER_bf78d47d', 42), (u'USER_4c4e1210', 76), (u'USER_1ac6c6f0', 20), (u
58599', 64), (u'USER_5f202603', 38), (u'USER_fe1fae24', 44), (u'USER_ae59b859', 26), (u'USER_9194e7ed', 33), (u'USER_52447833', 20), (u'USER_f5e2675d', 24),
534496d1', 23), (u'USER_1022fab6', 35), (u'USER_1e0e8b2c', 26), (u'USER_79e8fa0f', 85), (u'USER_0e22c69a', 28), (u'USER_30a8899b', 24), (u'USER_8044d2b6', 3
ER_5ead21d0', 42), (u'USER_8a3335b9', 27), (u'USER_7c2a8850', 24), (u'USER_0ea430fa', 70), (u'USER_95856bdf', 33), (u'USER_a77658c1', 36), (u'USER_72a2864f'
'USER_d46ea462', 34), (u'USER_f5643eba', 87), (u'USER_d681d865', 24), (u'USER_1803b97f', 196), (u'USER_1fa6edbe', 31), (u'USER_333704b0', 23), (u'USER_b465f
, (u'USER_ac7c0796', 20), (u'USER_bcede0c4', 20), (u'USER_3b0b31ad', 38), (u'USER_487256c3', 41), (u'USER_32da5515', 35), (u'USER_ee45eea8', 24), (u'USER_3d
38), (u'USER_ab15f243', 23), (u'USER_277117a0', 52), (u'USER_34ebae0b', 20), (u'USER_2e07416c', 41), (u'USER_1a11fba7', 74), (u'USER_72488feb', 36), (u'USER
', 35), (u'USER_2b31e28c', 38), (u'USER_e705346f', 21), (u'USER_2a72973e', 21), (u'USER_c5a4cde9', 23), (u'USER_efe6e599', 27), (u'USER_b2ec8f6f', 21), (u'U
ca9', 23), (u'USER_d3ada733', 85), (u'USER_0fbe981e', 38), (u'USER_22c2d4f5', 21), (u'USER_17fea64b', 36), (u'USER_fc5e5a50', 70), (u'USER_a95bcd40', 21), (
9777c5', 22), (u'USER_e50c8182', 49), (u'USER_2463455c', 36), (u'USER_8bd902a2', 54), (u'USER_d53cb27f', 33), (u'USER_d04a41d0', 52), (u'USER_574875aa', 27)
_b642bb2d', 87), (u'USER_4f72c677', 36), (u'USER_98da410c', 25), (u'USER_fc40042a', 32), (u'USER_16033b01', 31), (u'USER_9a075f51', 41), (u'USER_7f17c262',
SER_33981d28', 52), (u'USER_3d1e2b8e', 40), (u'USER_1af09980', 43), (u'USER_797a2bc7', 21), (u'USER_5f1c68b5', 74), (u'USER_356b6ddf', 22), (u'USER_661e0d11
u'USER_1e6de496', 21), (u'USER_959688c2', 32), (u'USER_07d31a91', 55), (u'USER_034ca407', 46), (u'USER_e88cd0d2', 33), (u'USER_0a91fc37', 45), (u'USER_702b7
, (u'USER_71fd67b9', 30), (u'USER_d4657747', 26), (u'USER_d7a83d85', 20), (u'USER_ced5eaa4', 89), (u'USER_ab7ca275', 20), (u'USER_e52c761e', 51), (u'USER_31
22), (u'USER_081d800c', 41), (u'USER_13572826', 21), (u'USER_68d4c67d', 25), (u'USER_2112f332', 82), (u'USER_8b7f9b7e', 23), (u'USER_7dbf766b', 35), (u'USER
```



```
>>> text.saveAsTextFile("/user/lab/textfile.txt")
```



```
hadoop fs -ls
Found 4 items
drwx------   - root hdfs          0 2017-07-13 13:52 .Trash
drwxr-xr-x   - root hdfs          0 2017-07-05 14:45 .hiveJars
drwx------   - root hdfs          0 2017-07-17 14:59 .staging
drwxr-xr-x   - root hdfs          0 2017-08-01 17:27 textfile.txt
```



```
[root@sandbox ~]# hadoop fs -cat /user/lab/textfile.txt/part-00000 >> output.txt
[root@sandbox ~]# ls
anaconda-ks.cfg  install.log          output.txt     start_ambari.sh  stop_solr.sh
blueprint.json   install.log.syslog   part-00000     start_hbase.sh
build.out        lab                  sandbox.info   start_solr.sh
[root@sandbox ~]# pwd
/root
[root@sandbox ~]# hadoop fs -cat /user/lab/textfile.txt/part-00000 |head >> output.txt
cat: Unable to write to output stream.
[root@sandbox ~]# cat output.txt | head
(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_6244af88', 49)
(u'USER_cc0a7d67', 23)
(u'USER_09dbf5de', 98)
(u'USER_73dcbc65', 29)
[root@sandbox ~]# cat output.txt | head >> output2.txt
[root@sandbox ~]#
```