

Handed out: 10/06/2017

Due by 4:00 PM EST, 10/14/2017

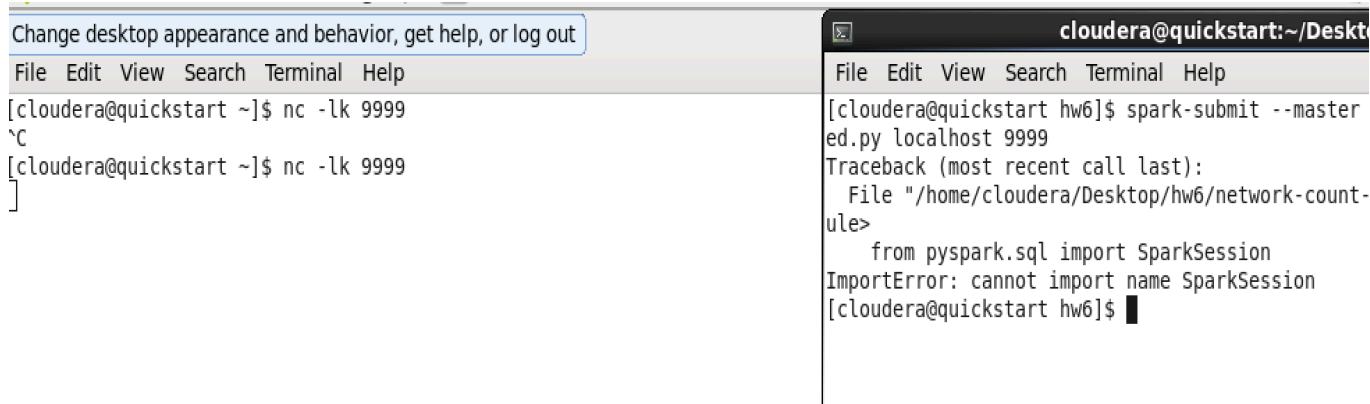
In your solution, please leave the text of every problem as presented here. Add your solution below the problem statement. It is important for us and we will take points if you ignore this request. Please make sure that you provide numeric or textual results of your calculations. If there are no results, we will treat the problem as not addressed. Just providing some code without results will give you 0 point.

Problem 1) Lecture notes contain script `network-count.py` in both Spark Streaming API and Spark Structured Streaming API. Use Linux `nc` (NetCat) utility to demonstrate that scripts work. Run both scripts on your own VM with Spark 2.2 installation. Cloudera VM with Spark 1.6 does not have Spark Structured Streaming API. (20%)

Spark Streaming API – network-count.py

cloudera@quickstart:~/Desktop/hw6	cloudera@quickstart:~/D
<pre>File Edit View Search Terminal Help [cloudera@quickstart hw6]\$ nc -lk 9999 kdsjfsadf sdff f f f f f f sdf asdf 324 wer wer2 3 23</pre>	<pre>File Edit View Search Terminal Help SLF4J: Found binding in [jar:file:/usr/lib/avro/avro-to- oggerBinder.class] SLF4J: See http://www.slf4j.org/codes.html#multiple_bin SLF4J: Actual binding is of type [org.slf4j.impl.Log4jL ----- Time: 2017-10-13 15:55:51 ----- Time: 2017-10-13 15:55:54 ----- (u'sdfff', 1) (u'kdsjfsadf', 1) ----- Time: 2017-10-13 15:55:57 ----- (u'', 1) (u'wer', 1) (u'sdf', 1) (u'324', 1) (u'asdf', 1) (u'f', 6) ----- Time: 2017-10-13 15:56:00 ----- (u'wer2', 1) (u'3', 1)</pre>

Spark Streaming API network-count-structured.py



The image shows two terminal windows side-by-side. The left window is titled 'Change desktop appearance and behavior, get help, or log out' and shows a user running 'nc -lk 9999' twice. The right window is titled 'cloudera@quickstart: ~/Deskt' and shows a user running 'spark-submit --master ed.py localhost 9999', which results in a 'Traceback' error: 'ImportError: cannot import name SparkSession' from 'pyspark.sql'.

```
Change desktop appearance and behavior, get help, or log out
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ nc -lk 9999
^C
[cloudera@quickstart ~]$ nc -lk 9999
]

cloudera@quickstart: ~/Deskt
File Edit View Search Terminal Help
[cloudera@quickstart hw6]$ spark-submit --master
ed.py localhost 9999
Traceback (most recent call last):
  File "/home/cloudera/Desktop/hw6/network-count-
ule>
    from pyspark.sql import SparkSession
ImportError: cannot import name SparkSession
[cloudera@quickstart hw6]$
```

Note to TA: Had issues with starting up **network-count-structured.py** due to not having Spark 2 installed. I know that Spark Session isn't apart of Spark 1.6. I tried to update to Spark 2 VIA a *yum install*, which is what cloudera's website suggested, but it installed 1.6 again. Moved on to the next problem to try to finish the other hw problems.

Problem 2) Expand provide orders.tar.gz file. Also, download shell scripts splitAndSend.original.sh and splitAndSend.sh and the Python script count-buys.py. First run splitAndSend.original.sh and count-buys.py. Record the failure mode of count-buys.py. Simply read the error message produced and tell us what is happening. Then run script splitAndSend.sh and Python program count-buys.py and tell us what the results are. In both cases show use contents of your HDFS directories input, output and staging. The second script splitAndSend.sh is supposed to reduce or eliminate the race condition. You might want to rename HDFS directory output from the first run in order to preserve it's content. In both cases, show the partial contents of your HDFS directories input, output and staging. In the second run, locate an output file named part-00000 that is not empty and show its content to us. Run these experiments on Cloudera VM. You need HDFS for these programs to run.
(30%)

count-buys.py Error

```
Traceback (most recent call last):
  File "/home/cloudera/Desktop/hw6/count-buys.py", line 27, in <module>
    ssc.awaitTermination()
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/streaming/context.py", line 289, in
awaitTermination
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/java_gateway.py", line 813, in
__call__
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/protocol.py", line 308, in
```

```

get_return_value
py4j.protocol.Py4JJavaError: An error occurred while calling o19.awaitTermination.
: org.apache.spark.SparkException: An exception was raised by Python:
Traceback (most recent call last):
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/streaming/util.py", line 65, in call
    r = self.func(t, *rdds)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/streaming/dstream.py", line 260, in
saveAsTextFile
    rdd.saveAsTextFile(path)
  File "/usr/lib/spark/python/lib/pyspark.zip/pyspark/rdd.py", line 1506, in
saveAsTextFile
    keyed._jrdd.map(self.ctx._jvm.BytesToString()).saveAsTextFile(path)
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/java_gateway.py", line 813, in
__call__
    answer, self.gateway_client, self.target_id, self.name)
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/protocol.py", line 308, in
get_return_value
    format(target_id, ".", name), value)
Py4JJavaError: An error occurred while calling o164.saveAsTextFile.
: org.apache.spark.SparkException: Job aborted due to stage failure: Task 1 in stage 9.0
failed 1 times, most recent failure: Lost task 1.0 in stage 9.0 (TID 18, localhost, executor
driver): java.io.FileNotFoundException: File does not exist:
/user/cloudera/input/chunkae._COPYING_

```

splitAndSend.original.sh
hadoop fs -ls /user/cloudera/input/

```

[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/input/
Found 50 items
-rw-r--r-- 1 cloudera cloudera 437626 2017-10-13 18:52 /user/cloudera/input/chunkaa
-rw-r--r-- 1 cloudera cloudera 448647 2017-10-13 18:52 /user/cloudera/input/chunkab
-rw-r--r-- 1 cloudera cloudera 448605 2017-10-13 18:52 /user/cloudera/input/chunkac
-rw-r--r-- 1 cloudera cloudera 448794 2017-10-13 18:52 /user/cloudera/input/chunkad
-rw-r--r-- 1 cloudera cloudera 448624 2017-10-13 18:52 /user/cloudera/input/chunkae
-rw-r--r-- 1 cloudera cloudera 448553 2017-10-13 18:52 /user/cloudera/input/chunkaf
-rw-r--r-- 1 cloudera cloudera 448436 2017-10-13 18:52 /user/cloudera/input/chunkag
-rw-r--r-- 1 cloudera cloudera 448679 2017-10-13 18:52 /user/cloudera/input/chunkah
-rw-r--r-- 1 cloudera cloudera 448424 2017-10-13 18:52 /user/cloudera/input/chunkai
-rw-r--r-- 1 cloudera cloudera 448564 2017-10-13 18:53 /user/cloudera/input/chunkaj
-rw-r--r-- 1 cloudera cloudera 458595 2017-10-13 18:53 /user/cloudera/input/chunkak
-rw-r--r-- 1 cloudera cloudera 458580 2017-10-13 18:53 /user/cloudera/input/chunkal
-rw-r--r-- 1 cloudera cloudera 458605 2017-10-13 18:53
/user/cloudera/input/chunkam
-rw-r--r-- 1 cloudera cloudera 458630 2017-10-13 18:53 /user/cloudera/input/chunkan
-rw-r--r-- 1 cloudera cloudera 458663 2017-10-13 18:53 /user/cloudera/input/chunkao
-rw-r--r-- 1 cloudera cloudera 458540 2017-10-13 18:53 /user/cloudera/input/chunkap

```

-rw-r--r--	1	cloudera	cloudera	458533	2017-10-13	18:53	/user/cloudera/input/chunkaq
-rw-r--r--	1	cloudera	cloudera	458403	2017-10-13	18:53	/user/cloudera/input/chunkar
-rw-r--r--	1	cloudera	cloudera	458470	2017-10-13	18:54	/user/cloudera/input/chunkas
-rw-r--r--	1	cloudera	cloudera	458584	2017-10-13	18:54	/user/cloudera/input/chunkat
-rw-r--r--	1	cloudera	cloudera	458333	2017-10-13	18:54	/user/cloudera/input/chunkau
-rw-r--r--	1	cloudera	cloudera	458600	2017-10-13	18:54	/user/cloudera/input/chunkav
-rw-r--r--	1	cloudera	cloudera	458732	2017-10-13	18:54	/user/cloudera/input/chunkaw
-rw-r--r--	1	cloudera	cloudera	458559	2017-10-13	18:54	/user/cloudera/input/chunkax
-rw-r--r--	1	cloudera	cloudera	458654	2017-10-13	18:54	/user/cloudera/input/chunkay
-rw-r--r--	1	cloudera	cloudera	458533	2017-10-13	18:54	/user/cloudera/input/chunkaz
-rw-r--r--	1	cloudera	cloudera	458535	2017-10-13	18:54	/user/cloudera/input/chunkba
-rw-r--r--	1	cloudera	cloudera	458700	2017-10-13	18:54	/user/cloudera/input/chunkbb
-rw-r--r--	1	cloudera	cloudera	458509	2017-10-13	18:55	/user/cloudera/input/chunkbc
-rw-r--r--	1	cloudera	cloudera	458488	2017-10-13	18:55	/user/cloudera/input/chunkbd
-rw-r--r--	1	cloudera	cloudera	458527	2017-10-13	18:55	/user/cloudera/input/chunkbe
-rw-r--r--	1	cloudera	cloudera	458662	2017-10-13	18:55	/user/cloudera/input/chunkbf
-rw-r--r--	1	cloudera	cloudera	458618	2017-10-13	18:55	/user/cloudera/input/chunkbg
-rw-r--r--	1	cloudera	cloudera	458673	2017-10-13	18:55	/user/cloudera/input/chunkbh
-rw-r--r--	1	cloudera	cloudera	458431	2017-10-13	18:55	/user/cloudera/input/chunkbi
-rw-r--r--	1	cloudera	cloudera	458464	2017-10-13	18:55	/user/cloudera/input/chunkbj
-rw-r--r--	1	cloudera	cloudera	458450	2017-10-13	18:55	/user/cloudera/input/chunkbk
-rw-r--r--	1	cloudera	cloudera	458536	2017-10-13	18:56	/user/cloudera/input/chunkbl
-rw-r--r--	1	cloudera	cloudera	458412	2017-10-13	18:56	/user/cloudera/input/chunkbm
-rw-r--r--	1	cloudera	cloudera	458712	2017-10-13	18:56	/user/cloudera/input/chunkbn
-rw-r--r--	1	cloudera	cloudera	458586	2017-10-13	18:56	/user/cloudera/input/chunkbo
-rw-r--r--	1	cloudera	cloudera	458501	2017-10-13	18:56	/user/cloudera/input/chunkbp
-rw-r--r--	1	cloudera	cloudera	458569	2017-10-13	18:56	/user/cloudera/input/chunkbq
-rw-r--r--	1	cloudera	cloudera	458459	2017-10-13	18:56	/user/cloudera/input/chunkbr
-rw-r--r--	1	cloudera	cloudera	458737	2017-10-13	18:56	/user/cloudera/input/chunkbs
-rw-r--r--	1	cloudera	cloudera	458556	2017-10-13	18:56	/user/cloudera/input/chunkbt
-rw-r--r--	1	cloudera	cloudera	458358	2017-10-13	18:57	/user/cloudera/input/chunkbu
-rw-r--r--	1	cloudera	cloudera	458570	2017-10-13	18:57	/user/cloudera/input/chunkbv
-rw-r--r--	1	cloudera	cloudera	458639	2017-10-13	18:57	

```
/user/cloudera/input/chunkbw
-rw-r--r-- 1 cloudera cloudera 458503 2017-10-13 18:57
/user/cloudera/input/chunkbx
```

```
splitAndSend.original.sh
hadoop fs -ls /user/cloudera/output/
```

```
[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/output/
Found 4 items
drwxr-xr-x - cloudera cloudera 0 2017-10-13 18:52 /user/cloudera/output/output-
1507945923000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 18:52 /user/cloudera/output/output-
1507945932000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 18:52 /user/cloudera/output/output-
1507945941000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 18:52 /user/cloudera/output/output-
1507945950000.txt
```

```
splitAndSend.original.sh
hadoop fs -ls /user/cloudera/staging/
```

```
[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/staging/
```

```
splitAndSend.sh
hadoop fs -ls /user/cloudera/input/
```

```
[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/input
Found 50 items
-rw-r--r-- 1 cloudera cloudera 437626 2017-10-13 19:58 /user/cloudera/input/chunkaa
-rw-r--r-- 1 cloudera cloudera 448647 2017-10-13 19:58 /user/cloudera/input/chunkab
-rw-r--r-- 1 cloudera cloudera 448605 2017-10-13 19:58 /user/cloudera/input/chunkac
-rw-r--r-- 1 cloudera cloudera 448794 2017-10-13 19:58 /user/cloudera/input/chunkad
-rw-r--r-- 1 cloudera cloudera 448624 2017-10-13 19:58 /user/cloudera/input/chunkae
-rw-r--r-- 1 cloudera cloudera 448553 2017-10-13 19:58 /user/cloudera/input/chunkaf
-rw-r--r-- 1 cloudera cloudera 448436 2017-10-13 19:59 /user/cloudera/input/chunkag
-rw-r--r-- 1 cloudera cloudera 448679 2017-10-13 19:59 /user/cloudera/input/chunkah
-rw-r--r-- 1 cloudera cloudera 448424 2017-10-13 19:59 /user/cloudera/input/chunkai
-rw-r--r-- 1 cloudera cloudera 448564 2017-10-13 19:59 /user/cloudera/input/chunkaj
-rw-r--r-- 1 cloudera cloudera 458595 2017-10-13 19:59 /user/cloudera/input/chunkak
-rw-r--r-- 1 cloudera cloudera 458580 2017-10-13 19:59 /user/cloudera/input/chunkal
-rw-r--r-- 1 cloudera cloudera 458605 2017-10-13 19:59
/user/cloudera/input/chunkam
-rw-r--r-- 1 cloudera cloudera 458630 2017-10-13 20:00 /user/cloudera/input/chunkan
-rw-r--r-- 1 cloudera cloudera 458663 2017-10-13 20:00 /user/cloudera/input/chunkao
```

-rw-r--r--	1	cloudera	cloudera	458540	2017-10-13 20:00	/user/cloudera/input/chunkap
-rw-r--r--	1	cloudera	cloudera	458533	2017-10-13 20:00	/user/cloudera/input/chunkaq
-rw-r--r--	1	cloudera	cloudera	458403	2017-10-13 20:00	/user/cloudera/input/chunkar
-rw-r--r--	1	cloudera	cloudera	458470	2017-10-13 20:00	/user/cloudera/input/chunkas
-rw-r--r--	1	cloudera	cloudera	458584	2017-10-13 20:01	/user/cloudera/input/chunkat
-rw-r--r--	1	cloudera	cloudera	458333	2017-10-13 20:01	/user/cloudera/input/chunkau
-rw-r--r--	1	cloudera	cloudera	458600	2017-10-13 20:01	/user/cloudera/input/chunkav
-rw-r--r--	1	cloudera	cloudera	458732	2017-10-13 20:01	/user/cloudera/input/chunkaw
-rw-r--r--	1	cloudera	cloudera	458559	2017-10-13 20:01	/user/cloudera/input/chunkax
-rw-r--r--	1	cloudera	cloudera	458654	2017-10-13 20:01	/user/cloudera/input/chunkay
-rw-r--r--	1	cloudera	cloudera	458533	2017-10-13 20:02	/user/cloudera/input/chunkaz
-rw-r--r--	1	cloudera	cloudera	458535	2017-10-13 20:02	/user/cloudera/input/chunkba
-rw-r--r--	1	cloudera	cloudera	458700	2017-10-13 20:02	/user/cloudera/input/chunkbb
-rw-r--r--	1	cloudera	cloudera	458509	2017-10-13 20:02	/user/cloudera/input/chunkbc
-rw-r--r--	1	cloudera	cloudera	458488	2017-10-13 20:02	/user/cloudera/input/chunkbd
-rw-r--r--	1	cloudera	cloudera	458527	2017-10-13 20:02	/user/cloudera/input/chunkbe
-rw-r--r--	1	cloudera	cloudera	458662	2017-10-13 20:03	/user/cloudera/input/chunkbf
-rw-r--r--	1	cloudera	cloudera	458618	2017-10-13 20:03	/user/cloudera/input/chunkbg
-rw-r--r--	1	cloudera	cloudera	458673	2017-10-13 20:03	/user/cloudera/input/chunkbh
-rw-r--r--	1	cloudera	cloudera	458431	2017-10-13 20:03	/user/cloudera/input/chunkbi
-rw-r--r--	1	cloudera	cloudera	458464	2017-10-13 20:03	/user/cloudera/input/chunkbj
-rw-r--r--	1	cloudera	cloudera	458450	2017-10-13 20:03	/user/cloudera/input/chunkbk
-rw-r--r--	1	cloudera	cloudera	458536	2017-10-13 20:04	/user/cloudera/input/chunkbl
-rw-r--r--	1	cloudera	cloudera	458412	2017-10-13 20:04	/user/cloudera/input/chunkbm
-rw-r--r--	1	cloudera	cloudera	458712	2017-10-13 20:04	/user/cloudera/input/chunkbn
-rw-r--r--	1	cloudera	cloudera	458586	2017-10-13 20:04	/user/cloudera/input/chunkbo
-rw-r--r--	1	cloudera	cloudera	458501	2017-10-13 20:04	/user/cloudera/input/chunkbp
-rw-r--r--	1	cloudera	cloudera	458569	2017-10-13 20:04	/user/cloudera/input/chunkbq
-rw-r--r--	1	cloudera	cloudera	458459	2017-10-13 20:05	/user/cloudera/input/chunkbr
-rw-r--r--	1	cloudera	cloudera	458737	2017-10-13 20:05	/user/cloudera/input/chunkbs
-rw-r--r--	1	cloudera	cloudera	458556	2017-10-13 20:05	/user/cloudera/input/chunkbt
-rw-r--r--	1	cloudera	cloudera	458358	2017-10-13 20:05	/user/cloudera/input/chunkbu
-rw-r--r--	1	cloudera	cloudera	458570	2017-10-13 20:05	/user/cloudera/input/chunkbv

```
-rw-r--r-- 1 cloudera cloudera 458639 2017-10-13 20:05
/user/cloudera/input/chunkbw
-rw-r--r-- 1 cloudera cloudera 458503 2017-10-13 20:06
/user/cloudera/input/chunkbx
```

```
splitAndSend.sh
hadoop fs -ls /user/cloudera/output/
```

```
[[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/output
Found 55 items
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:25
/user/cloudera/output/output-1507951539000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:25
/user/cloudera/output/output-1507951548000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:25
/user/cloudera/output/output-1507951557000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951566000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951575000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951584000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951593000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951602000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:26
/user/cloudera/output/output-1507951611000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951620000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951629000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951638000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951647000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951656000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951665000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:27
/user/cloudera/output/output-1507951674000.txt
drwxr-xr-x - cloudera cloudera 0 2017-10-13 20:28
/user/cloudera/output/output-1507951683000.txt
```

```
splitAndSend.sh
hadoop fs -ls /user/cloudera/staging/
```

```
[cloudera@quickstart hw6]$ hadoop fs -ls /user/cloudera/staging
```

In the second run, locate an output file named `part-00000` that is not empty and show its content to us.

```
[cloudera@quickstart hw6]$ hadoop fs -cat /user/cloudera/output/output-
1507955193000.txt/part-00000
(False, 3940L)
(True, 4050L)
```

Problem 3). In the second run of the previous problem you will notice that many of `part-00000` files in your output directory are empty. Could you explain why. (10%)

There simply was nothing ordered during that timeframe that `part-0000` was written, so nothing will be recorded, and thus the file is empty.

Problem 4) Could you rewrite `count-buys.sh` in Spark Structured Streaming API. If you do that change script `splitAndSend.sh` to move generated chunks from the local files system directory `staging` to local file system directory `input`. Run this experiment on your VM with Spark 2.2. (20%)

Yes, by passing the local argument to `splitAndSend.sh` it will write to the local file system and not HDFS.

I ran out of time to finish, but I understand using `file:///` to get data in `count-buys.sh` and `splitAndSend.sh` to work locally instead of `hdfs:///`.

Problem 5) Examine provided Python program `stateful_wordcount.py`. Make it work as is. If there are errors on the code, fix them. Modify the code so that it outputs the number of words starting with letters `a` and `b`. Demonstrate that modified program work. You should provide several both positive and negative examples. (20%)

Make it work as is. If there are errors on the code, fix them.

```
[cloudera@quickstart hw6]$ nc -lk 9999
```



```
asdasd
asd
112312
3
123
12
1
212
1
212
12
1
21
2
12
13
123
12
31
23
123
sd
c
xcs
dc
sd
sd
sd
sd
```

stateful_wordcount.py

```
-----
Time: 2017-10-14 10:37:09
-----
```

```
-----
Time: 2017-10-14 10:37:10
-----
```

```
(u'112312', 1)
(u'c', 1)
(u'212', 2)
(u'31', 1)
(u'13', 1)
```

(u'asdasd', 1)

(u'1', 3)

(u'3', 1)

(u'12', 4)

(u'21', 1)

...

Time: 2017-10-14 10:37:11

(u'112312', 1)

(u'c', 1)

(u'212', 2)

(u'31', 1)

(u'13', 1)

(u'asdasd', 1)

(u'1', 3)

(u'3', 1)

(u'12', 4)

(u'21', 1)

...

Time: 2017-10-14 10:37:12

(u'112312', 1)

(u'c', 1)

(u'212', 2)

(u'31', 1)

(u'13', 1)

(u'asdasd', 1)

(u'1', 3)

(u'3', 1)

(u'12', 4)

(u'21', 1)

...

Time: 2017-10-14 10:37:13

(u'112312', 1)

(u'c', 1)

(u'212', 2)

(u'31', 1)

(u'13', 1)

(u'asdasd', 1)

```
(u'1', 3)
(u'3', 1)
(u'12', 4)
(u'21', 1)
...
```

Time: 2017-10-14 10:37:14

```
(u'112312', 1)
(u'c', 1)
(u'212', 2)
(u'31', 1)
(u'13', 1)
(u'asdasd', 1)
(u'1', 3)
(u'3', 1)
(u'12', 4)
(u'21', 1)
...
```

Time: 2017-10-14 10:37:15

```
(u'112312', 1)
(u'c', 1)
(u'212', 2)
(u'31', 1)
(u'13', 1)
(u'asdasd', 1)
(u'1', 3)
(u'3', 1)
(u'12', 4)
(u'21', 1)
...
```

Modify the code so that it outputs the number of words starting with letters a and b. Demonstrate that modified program work. You should provide several both positive and negative examples.

CODE CHANGE

```
running_counts = lines.flatMap(lambda line: line.split(" "))\
    .filter(lambda word: word.lower().startswith(('a','b')))\
    .map(lambda word: (word, 1))\
```

```
.updateStateByKey(updateFunc)
```

```
cloudera@quickstart:~/Desktop/hw6
File Edit View Search Terminal Help
[cloudera@quickstart hw6]$ nc -lk 9999
kdsjfkasdf
sdf
f
f
f
f
f
sdf
asdf
324
wer
wer2
3
23

cloudera@quickstart:~/Desktop/hw6
File Edit View Search Terminal Help
SLF4J: Found binding in [jar:file:/usr/lib/avro/avro-tools-1.7.6-cdh5
oggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an e
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
-----
Time: 2017-10-13 15:55:51
-----
Time: 2017-10-13 15:55:54
-----
(u'sdff', 1)
(u'kdsjfkasdf', 1)
-----
Time: 2017-10-13 15:55:57
-----
(u'', 1)
(u'wer', 1)
(u'sdf', 1)
(u'324', 1)
(u'asdf', 1)
(u'f', 6)
-----
Time: 2017-10-13 15:56:00
-----
(u'wer2', 1)
(u'3', 1)

Applications Places System
Browse and run installed applications cloudera@quickstart:~
File Edit View Search Terminal Help
cloudera@quickstart ~]$ nc -lk 9999
asdasd
asb
if
lsdv
/
iv
l
/s
/
;
ivsdvsdv
]
```

You are welcome to implement your solution in any language of your choice.

You are welcome to follow any other instructions and use any other programming or scripting language to accomplish the above goals.

Please, describe every step of your work and present all intermediate and final results in a Word document. Please, copy past text version of all essential command and snippets of results into the Word document. We cannot retype text that is in JPG images. Please, always submit a separate copy of the original, working scripts and/or class files you used

as separate files. Sometimes we need to run your code and retyping is too costly. Please include in your MS Word document only relevant portions of the console output or output files. Sometime either console output or the result file is too long and including it into the MS Word document makes that document too hard to read. PLEASE DO NOT EMBED files into your MS Word document. Please, submit to the class drop box. For issues and comments visit the class Discussion Board. You are not obliged to use Java or Eclipse. You are welcome to use any language and any IDE of your choice.