W205 summer 2017 section 2 – Lab10 - K Iwasaki

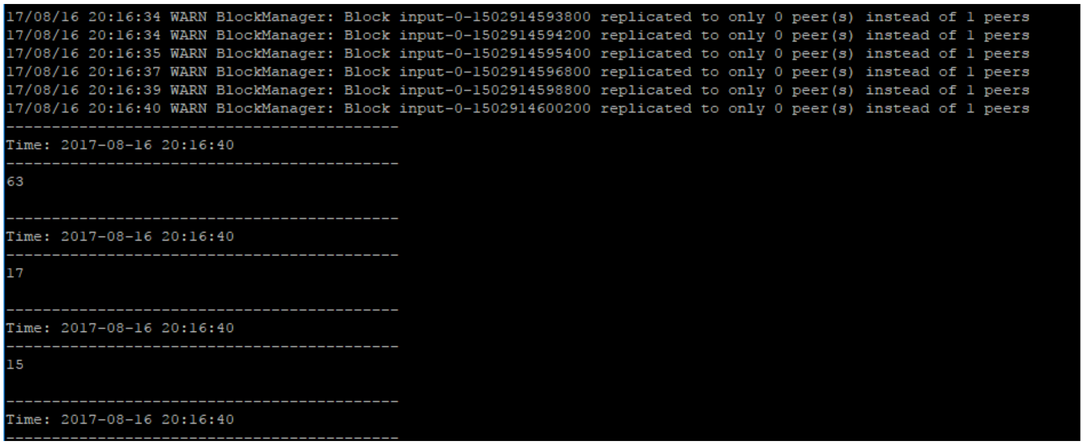
SUBMISSION 1: *Print only words with a length > 5 characters. Submit the pyspark code*

```
MASTER=local[2] pyspark
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
ssc = StreamingContext(sc, 1)
lines= ssc.textFileStream("file:///tmp/datastreams")
uclines = lines.map(lambda word: word.upper())
uclines = uclines.filter(lambda word: len(word) > 5)
uclines.pprint()
ssc.start()
ssc.stop()
```

SUBMISSION 2: *Change the code so that you save the venue components to a text file. Submit you code.*

```
MASTER=local[2] pyspark
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
ssc = StreamingContext(sc, 1)
lines= ssc.textFileStream("file:///tmp/datastreams")
slines = lines.flatMap(lambda x: [ j['venue'] for j in json.loads('['+x+']') if
'venue' in j] )
cnt=slines.count()
cnt.pprint()
slines.pprint()
slines.saveAsTextFiles("file:///tmp/venues.txt")
ssc.start()
ssc.stop()
```

SUBMISSION 4a: *Provide a screenshot of the running Spark Streaming application that shows that the CountByWindow indeed provides an sum of the counts from the 3 latest batches.*

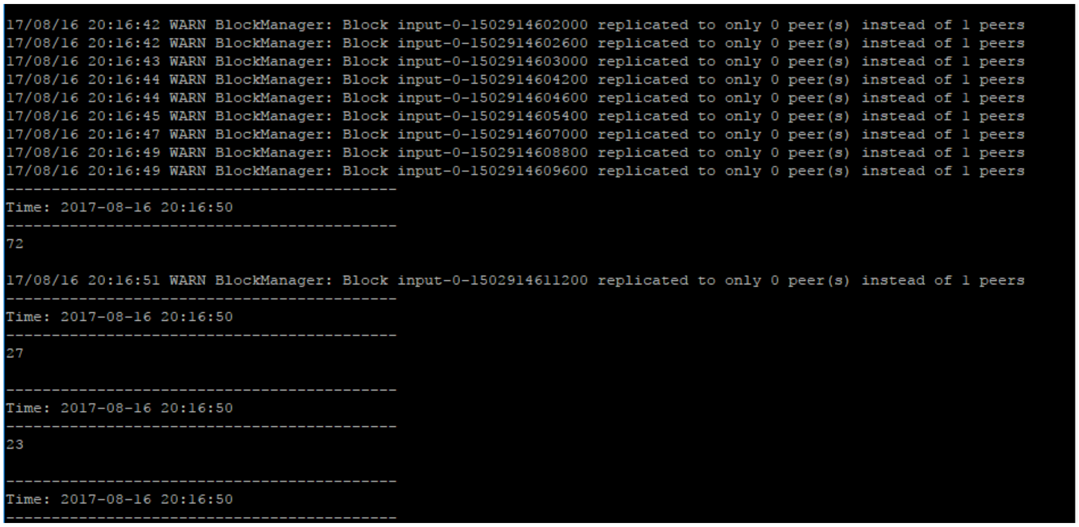


```
17/08/16 20:16:34 WARN BlockManager: Block input-0-1502914593800 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:34 WARN BlockManager: Block input-0-1502914594200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:35 WARN BlockManager: Block input-0-1502914595400 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:37 WARN BlockManager: Block input-0-1502914596800 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:39 WARN BlockManager: Block input-0-1502914598800 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:40 WARN BlockManager: Block input-0-1502914600200 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:16:40
-------------------------------------------
63

-------------------------------------------
Time: 2017-08-16 20:16:40
-------------------------------------------
17

-------------------------------------------
Time: 2017-08-16 20:16:40
-------------------------------------------
15

-------------------------------------------
Time: 2017-08-16 20:16:40
-------------------------------------------
```

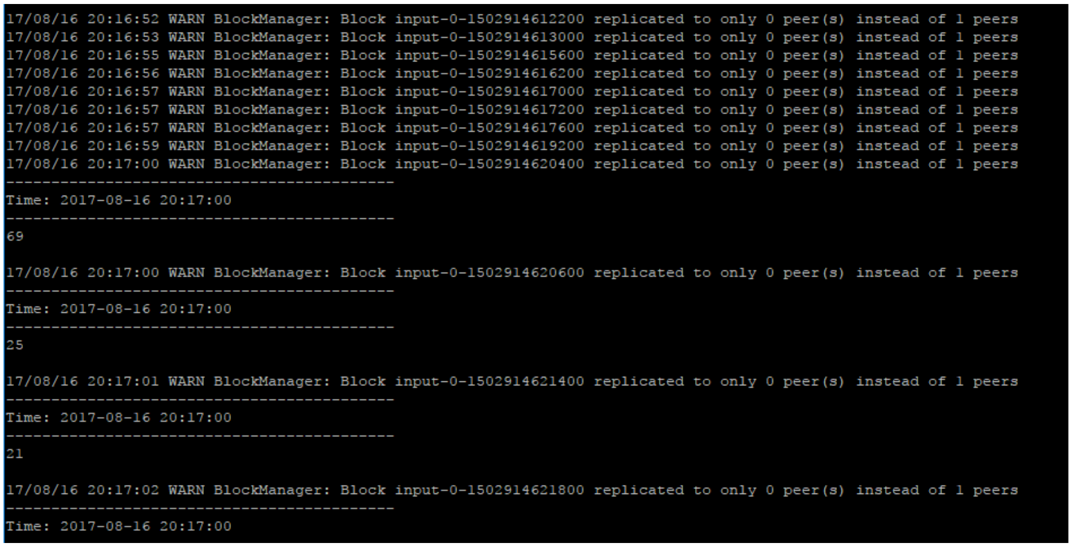


```
17/08/16 20:16:42 WARN BlockManager: Block input-0-1502914602000 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:42 WARN BlockManager: Block input-0-1502914602600 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:43 WARN BlockManager: Block input-0-1502914603000 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:44 WARN BlockManager: Block input-0-1502914604200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:44 WARN BlockManager: Block input-0-1502914604600 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:45 WARN BlockManager: Block input-0-1502914605400 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:47 WARN BlockManager: Block input-0-1502914607000 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:49 WARN BlockManager: Block input-0-1502914608800 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:49 WARN BlockManager: Block input-0-1502914609600 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:16:50
-------------------------------------------
72

17/08/16 20:16:51 WARN BlockManager: Block input-0-1502914611200 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:16:50
-------------------------------------------
27

-------------------------------------------
Time: 2017-08-16 20:16:50
-------------------------------------------
23

-------------------------------------------
Time: 2017-08-16 20:16:50
-------------------------------------------
```



```
17/08/16 20:16:52 WARN BlockManager: Block input-0-1502914612200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:53 WARN BlockManager: Block input-0-1502914613000 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:55 WARN BlockManager: Block input-0-1502914615600 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:56 WARN BlockManager: Block input-0-1502914616200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:57 WARN BlockManager: Block input-0-1502914617000 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:57 WARN BlockManager: Block input-0-1502914617200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:57 WARN BlockManager: Block input-0-1502914617600 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:16:59 WARN BlockManager: Block input-0-1502914619200 replicated to only 0 peer(s) instead of 1 peers
17/08/16 20:17:00 WARN BlockManager: Block input-0-1502914620400 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:17:00
-------------------------------------------
69

17/08/16 20:17:00 WARN BlockManager: Block input-0-1502914620600 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:17:00
-------------------------------------------
25

17/08/16 20:17:01 WARN BlockManager: Block input-0-1502914621400 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:17:00
-------------------------------------------
21

17/08/16 20:17:02 WARN BlockManager: Block input-0-1502914621800 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2017-08-16 20:17:00
```

SUBMISSION 4b: *Also explain what the difference is between having 10 sec batches with a 30 sec sliding window and a 30 second batch length.*

See the two columns: Batch: 30 sec and Window: 30 sec, Batch 10 sec for the difference

Time (base t)	**Events** Some random words stream	**Batch: 10 sec**	**Batch: 30 sec**	**Window: 30 sec, Batch: 10 sec**
T	W1	W1		
T+10	W2	W2		
T+20	W3	W3	W1 W2 W3	W1 W2 W3
T+30	W4	W4		W2 W3 W4
T+40	W5	W5		W3 W4 W5

T+50	W6	W6	W4 W5 W6	W4 W5 W6
T+60	W7	W7		W5 W6 W7