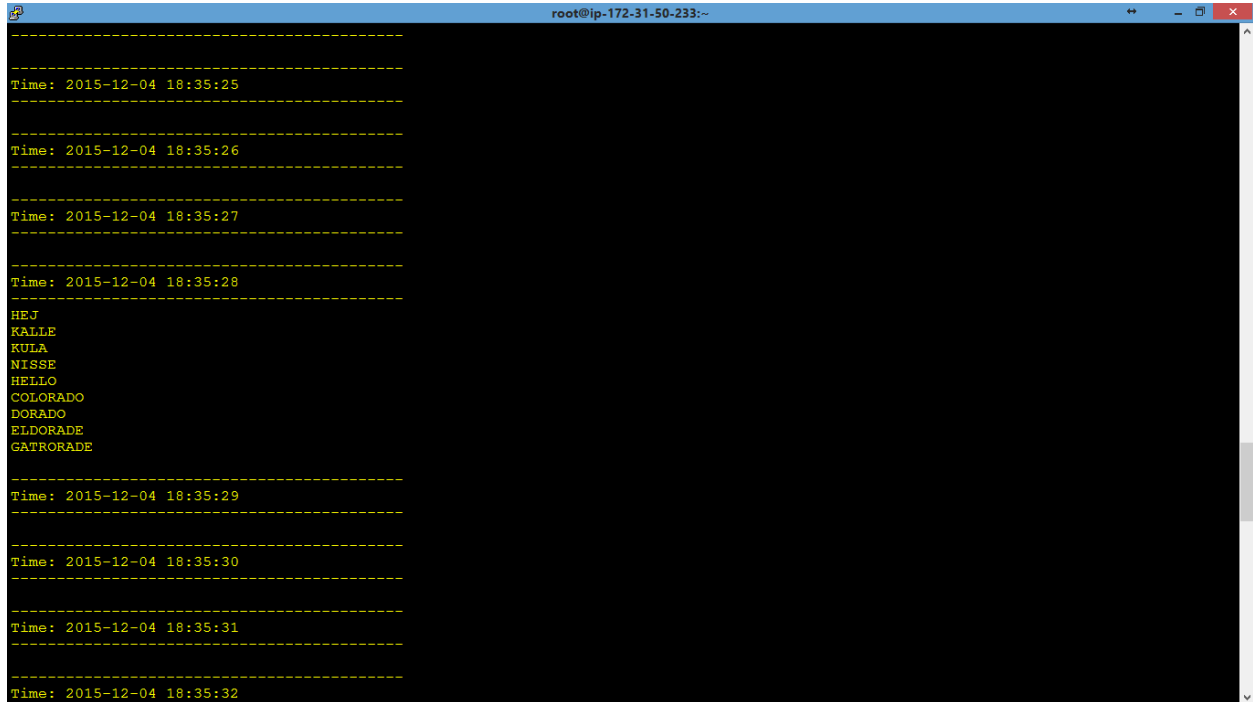


MIDS W205 Lab # 12 Lab Spark Streaming Introduction

Alan Wang

SUBMISSION 1: Provide a screenshot of the output from the Spark Streaming process



The screenshot shows a terminal window with a black background and yellow text. The window title bar indicates the user is root at ip-172-31-50-233. The output consists of several timestamped lines, each preceded and followed by dashed lines. Between the timestamps at 18:35:28 and 18:35:29, there is a list of words: HEJ, KALLE, KULA, NISSE, HELLO, COLORADO, DORADO, ELDORADE, and GATRORADE.

```
-----  
Time: 2015-12-04 18:35:25  
-----  
  
-----  
Time: 2015-12-04 18:35:26  
-----  
  
-----  
Time: 2015-12-04 18:35:27  
-----  
  
-----  
Time: 2015-12-04 18:35:28  
-----  
HEJ  
KALLE  
KULA  
NISSE  
HELLO  
COLORADO  
DORADO  
ELDORADE  
GATRORADE  
-----  
Time: 2015-12-04 18:35:29  
-----  
  
-----  
Time: 2015-12-04 18:35:30  
-----  
  
-----  
Time: 2015-12-04 18:35:31  
-----  
  
-----  
Time: 2015-12-04 18:35:32  
-----
```

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

\$vi lab12.s2.py

Also I attached the py file in the same 'lab12' folder

```
root@ip-172-31-50-233:~  
#MASTER=local[2] pyspark  
from pyspark import SparkContext  
from pyspark.streaming import StreamingContext  
import json  
ssc = StreamingContext(sc, 5)  
lines=ssc.textFileStream("file:///tmp/datastreams").flatMap(lambda x: [ j['venue'] fo  
r j in json.loads([''+x+']) if 'venue' in j] ).saveAsTextFiles('file:///tmp/output/o  
uts')  
#the file is saved as a textfile in potentially parts as  
#/tmp/output/outs-1449421755000/part-00000 and so on  
ssc.start()  
█
```

Saved Text File:

\$more /tmp/output/outs-1449421755000/part-00000

```
root@ip-172-31-50-233:/tmp/output/outs-1449421755000  
{u'lat': 37.790005, u'venue_id': 21741962, u'lon': -122.397354, u'venue_name': u'Couchbase San Fra  
ncisco'}  
{u'lat': 1.300357, u'venue_id': 17525192, u'lon': 103.839043, u'venue_name': u'Food Republic at Le  
vel 5'}  
{u'lat': -28.001623, u'venue_id': 1087923, u'lon': 153.416496, u'venue_name': u'Gold Coast Art Cen  
tre'}  
{u'lat': 46.807144, u'venue_id': 24186719, u'lon': -71.238655, u'venue_name': u'Kalimera'}  
{u'lat': 42.622475, u'venue_id': 5728532, u'lon': -71.361862, u'venue_name': u'Panera Bread'}  
{u'lat': 45.515067, u'venue_id': 899456, u'lon': -122.678453, u'venue_name': u'Portland City Hall'  
}  
{u'lat': 47.65704, u'venue_id': 19107532, u'lon': -122.344719, u'venue_name': u'Little Heart Space  
'}  
{u'lat': 33.048553, u'venue_id': 4448112, u'lon': -96.82901, u'venue_name': u'Cinemark West Plano  
and XD'}  
{u'lat': 33.216774, u'venue_id': 23939042, u'lon': -117.218994, u'venue_name': u'First Lutheran Ch  
urch'}  
{u'lat': -28.001623, u'venue_id': 1087923, u'lon': 153.416496, u'venue_name': u'Gold Coast Art Cen  
tre'}  
{u'lat': 45.515067, u'venue_id': 899456, u'lon': -122.678453, u'venue_name': u'Portland City Hall'  
}  
{u'lat': 35.098927, u'venue_id': 269007, u'lon': -106.480415, u'venue_name': u'Embudo Trailhead'}  
{u'lat': 40.744094, u'venue_id': 883795, u'lon': -73.987724, u'venue_name': u'230 Fifth Rooftop'}  
{u'lat': 47.61945, u'venue_id': 23949899, u'lon': -122.324974, u'venue_name': u'Starbucks'}  
--More-- (8%)
```

SUBMISSION 3: Provide a screenshot showing the running Spark Streaming application.

```
root@ip-172-31-50-233:~/lab12
{u'lat': 36.740242, u'venue_id': 24002879, u'lon': -119.78109, u'venue_name': u'El Torito'}
{u'lat': 40.778038, u'venue_id': 24140787, u'lon': -73.98098, u'venue_name': u'Triad Theater'}
{u'lat': 42.52779, u'venue_id': 7214802, u'lon': -71.227333, u'venue_name': u'Community Congregational Church '}
{u'lat': 39.764874, u'venue_id': 934792, u'lon': -86.161437, u'venue_name': u'Tilt'}
...
15/12/06 17:48:14 WARN BlockManager: Block input-0-1449424094600 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:17 WARN BlockManager: Block input-0-1449424097000 replicated to only 0 peer(s) instead of 1 peers
-----
Time: 2015-12-06 17:48:20
-----
10
-----
Time: 2015-12-06 17:48:20
-----
10
-----
15/12/06 17:48:21 WARN BlockManager: Block input-0-1449424101000 replicated to only 0 peer(s) instead of 1 peers
-----
Time: 2015-12-06 17:48:20
-----
{u'lat': 37.337513, u'venue_id': 660886, u'lon': -122.04075, u'venue_name': u'Homestead Lanes'}
{u'lat': 42.317829, u'venue_id': 9426332, u'lon': -72.6334, u'venue_name': u'The Academy of Music'}
{u'lat': 40.754635, u'venue_id': 24233328, u'lon': -73.845627, u'venue_name': u'Mets - Willets Point Train Station'}
{u'lat': 26.928852, u'venue_id': 997970, u'lon': -82.044839, u'venue_name': u'Beef O'Brady's"}
{u'lat': 44.549026, u'venue_id': 24239706, u'lon': 10.939021, u'venue_name': u'Castelnuovo Rangone'}
{u'lat': 40.621346, u'venue_id': 23524059, u'lon': -74.02697, u'venue_name': u'Outside the main entrance of Century 21, near the red awn
ing and street clock')}
{u'lat': 51.226822, u'venue_id': 4084742, u'lon': 6.77321, u'venue_name': u'McLaughlins Irish Pub'}
{u'lat': 33.74831, u'venue_id': 23592369, u'lon': -84.391113, u'venue_name': u'La casa di LaRue'}
{u'lat': 43.081066, u'venue_id': 24243639, u'lon': -73.783417, u'venue_name': u'Saratoga public library- H. Dutcher Community Room'}
{u'lat': 37.788273, u'venue_id': 24221498, u'lon': -122.420532, u'venue_name': u'Mayes Restaurant & Nightclub'}
15/12/06 17:48:22 WARN BlockManager: Block input-0-1449424101800 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:22 WARN BlockManager: Block input-0-1449424102000 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:24 WARN BlockManager: Block input-0-1449424104200 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:28 WARN BlockManager: Block input-0-1449424107800 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:28 WARN BlockManager: Block input-0-1449424108400 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:29 WARN BlockManager: Block input-0-1449424108800 replicated to only 0 peer(s) instead of 1 peers
15/12/06 17:48:29 WARN BlockManager: Block input-0-1449424109400 replicated to only 0 peer(s) instead of 1 peers
-----
Time: 2015-12-06 17:48:30
-----
26
```

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. See example screenshot below.

Screen shot with the running Spark streaming, but can't show a window of 3 because of the warning messages:

```
root@ip-172-31-50-233:~/lab12
47
-----
Time: 2015-12-06 18:27:50
13
-----
Time: 2015-12-06 18:27:50
7
-----
Time: 2015-12-06 18:27:50
(u'lat': 40.723244, u'venue_id': 24175206, u'lon': -74.006424, u'venue_name': u'Adelphi University Manhattan Center, 75 Varick Street, 2nd floor. State-issued photo ID required for admission to the building.')}
(u'lat': 29.794413, u'venue_id': 19112172, u'lon': -95.394051, u'venue_name': u'Reagan High School')}
(u'lat': 52.138832, u'venue_id': 23535697, u'lon': -0.458717, u'venue_name': u'the REC 6 Rothsay Gardens '}
(u'lat': 42.138123, u'venue_id': 4669162, u'lon': -87.995934, u'venue_name': u'Arlington Lanes LLC')}
(u'lat': 33.523117, u'venue_id': 16851342, u'lon': -117.158798, u'venue_name': u'Karl Strauss Brewery')}
(u'lat': 40.761105, u'venue_id': 20839562, u'lon': -73.97565, u'venue_name': u'MoDA')}
(u'lat': 44.978698, u'venue_id': 23980552, u'lon': -93.261116, u'venue_name': u'Crooked Pint Ale House')}

15/12/06 18:28:13 WARN BlockManager: Block input-0-1449426492800 replicated to only 0 peer(s) instead of 1 peers
-----
Time: 2015-12-06 18:28:00
67
-----
Time: 2015-12-06 18:28:00
38
-----
Time: 2015-12-06 18:28:00
29
-----
Time: 2015-12-06 18:28:00
(u'lat': 42.138123, u'venue_id': 4669162, u'lon': -87.995934, u'venue_name': u'Arlington Lanes LLC')}
(u'lat': 28.635307, u'venue_id': 16003372, u'lon': 77.22496, u'venue_name': u'cannaught place')}
(u'lat': 46.20245, u'venue_id': 23877436, u'lon': 6.13114, u'venue_name': u'L'Ethno Bar')}
(u'lat': 33.70089, u'venue_id': 18330822, u'lon': -78.927345, u'venue_name': u'Cinemark At Myrtle Beach')}
(u'lat': 33.922909, u'venue_id': 23705743, u'lon': -84.379799, u'venue_name': u'Steve's Live Music')}
(u'lat': 33.803177, u'venue_id': 23820949, u'lon': -84.393074, u'venue_name': u'Macquarium')}
(u'lat': 51.278728, u'venue_id': 12339222, u'lon': 1.081372, u'venue_name': u'Cafe Mauresque')}
(u'lat': 29.735094, u'venue_id': 23753432, u'lon': -95.39064, u'venue_name': u'The Iron Yard')}
(u'lat': 38.011555, u'venue_id': 17453102, u'lon': -121.324074, u'venue_name': u'Market Tavern')}
(u'lat': 42.98753, u'venue_id': 23870170, u'lon': -81.24939, u'venue_name': u'Campus Creative')}
...

15/12/06 18:28:14 WARN BlockManager: Block input-0-1449426494400 replicated to only 0 peer(s) instead of 1 peers
-----
Time: 2015-12-06 18:28:10
51
-----
Time: 2015-12-06 18:28:10
-----
Time: 2015-12-06 18:28:10
-----
Time: 2015-12-06 18:28:10
```

Now, I redirect the output to a file and virtually weeded out the warning messages, and was able to catch a 3-batch window as below: $15 + 9 + 14 = 38$ from 18:23:40, 18:23:50, and 18:24:00 respectively

```
root@ip-172-31-50-233:~/lab12
---
Time: 2015-12-06 18:23:40
53
---
Time: 2015-12-06 18:23:40
15
---
Time: 2015-12-06 18:23:40
9
---
Time: 2015-12-06 18:23:40
(u'lat': 26.09639, u'venue_id': 17954022, u'lon': -80.123146, u'venue_name': u'Broward Convention Center')
(u'lat': 29.716854, u'venue_id': 23576196, u'lon': -95.414925, u'venue_name': u'Plate & Bottle')
(u'lat': 0, u'venue_id': 22690182, u'lon': 0, u'venue_name': u'Downtown soccer meet up')
(u'lat': 38.944088, u'venue_id': 23574622, u'lon': -77.320999, u'venue_name': u'Bechtel Conference Center @ ASCE')
(u'lat': 51.224712, u'venue_id': 24090032, u'lon': 1.401505, u'venue_name': u'The Astor Community Theatre')
(u'lat': 45.503258, u'venue_id': 8378742, u'lon': -122.639793, u'venue_name': u'Clinton Street Theater')
(u'lat': 33.899609, u'venue_id': 24038830, u'lon': -117.885567, u'venue_name': u'Craig Regional Park')
(u'lat': 0, u'venue_id': 17364142, u'lon': 0, u'venue_name': u'Temperance Ave Park n Ride')
(u'lat': 26.16, u'venue_id': 23910511, u'lon': -80.309998, u'venue_name': u'Lesters Diner')
---
Time: 2015-12-06 18:23:50
39
---
Time: 2015-12-06 18:23:50
9
---
Time: 2015-12-06 18:23:50
5
---
Time: 2015-12-06 18:23:50
(u'lat': 43.043221, u'venue_id': 23794462, u'lon': -76.051208, u'venue_name': u'erie canal')
(u'lat': 52.599319, u'venue_id': 14265832, u'lon': -0.270522, u'venue_name': u'AMF Bowling')
(u'lat': 45.498764, u'venue_id': 22438332, u'lon': -122.766884, u'venue_name': u'leher pub')
(u'lat': 35.95612, u'venue_id': 24244169, u'lon': -83.926315, u'venue_name': u'Cox Auditorium - Alumni Memorial Building')
(u'lat': 43.699821, u'venue_id': 6305612, u'lon': -79.707977, u'venue_name': u'Mayfield United Church - Back Hall')
---
Time: 2015-12-06 18:24:00
38
---
Time: 2015-12-06 18:24:00
14
---
Time: 2015-12-06 18:24:00
14
---
Time: 2015-12-06 18:24:00
(u'lat': 40.860889, u'venue_id': 24081154, u'lon': -74.372223, u'venue_name': u'Pure Restaurant & Lounge')
--More-- (614)
```

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.

For 10-sec batch + 30 sliding window, each window covers 3 batches, and for 30-sec batch + 30 sliding window each window covers only 1 batch.