ie
SCHOOL OF
HUMAN SCIENCES
& TECHNOLOGY

# STREAM PROCESSING & REAL TIME ANALYTICS

## INDIVIDUAL ASSIGNMENT: Analyzing Twitter with Spark Streaming

By: Qing Loh

1. Fill the missing parts in the Spark Streaming application. (the TODO parts)

   TODO 1: ssc.checkpoint("checkpoint")

   TODO 2: tweets = ssc.socketTextStream("localhost", 9009)

   TODO 3: words = tweets.flatMap(lambda line: line.split(" "))

   TODO 4:

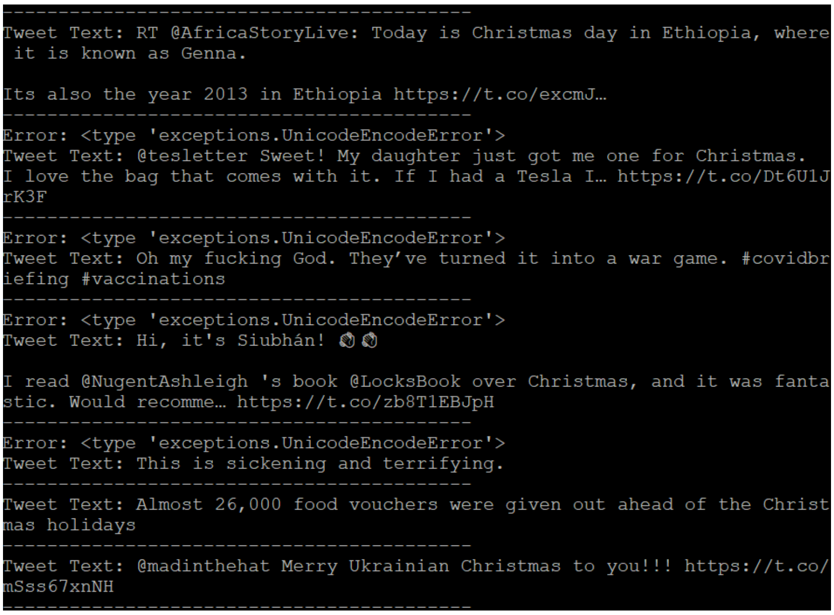
   ```
   def updateFunc(new_values, last_sum):
       return sum(new_values) + (last_sum or 0)
   tags_totals = hashtags.updateStateByKey(updateFunc)
   ```

2. Connect to the edge machine and upload the files or create them with a text editor. Please note that you need to change the application port from 9009 to a different one to avoid conflicts with running programs from other colleagues.

   See notebook codes.

3. In the query from twitter_app.py instead of getting the tweets of the location we have, use another location and word instead of #.
   For example, you can use Madrid bounding box and track Tweets that has Madrid word instead of #. Coordinates of Madrid Bounding Box: -3.7834,40.3735,-3.6233,40.4702 Execute the end-to-end example: (i) Start the python application and then (2) start the pyspark application. Copy and paste some output lines of this example.

   query_data = [('language', 'en'), ('locations', '0.489,51.2867,0.236,51.686'),('track','christmas')]

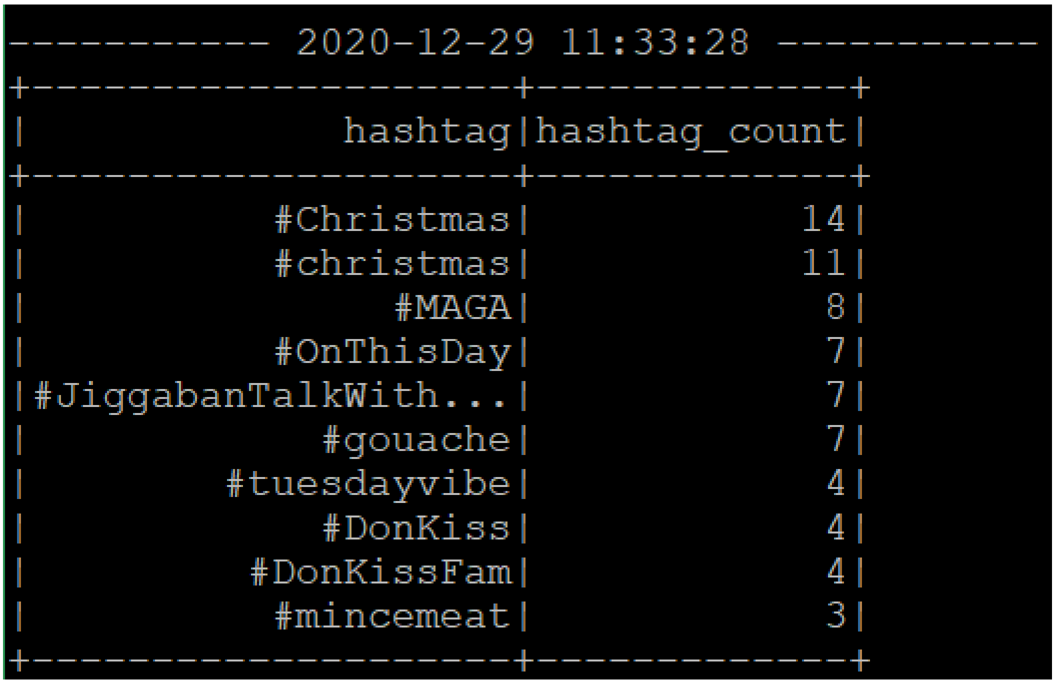


```
------------------------------------------
Tweet Text: RT @AfricaStoryLive: Today is Christmas day in Ethiopia, where
 it is known as Genna.

Its also the year 2013 in Ethiopia https://t.co/excmJ…
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: @tesletter Sweet! My daughter just got me one for Christmas.
I love the bag that comes with it. If I had a Tesla I… https://t.co/Dt6U1J
rK3F
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: Oh my fucking God. They’ve turned it into a war game. #covidbr
iefing #vaccinations
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: Hi, it's Siubhán! 👋👋

I read @NugentAshleigh 's book @LocksBook over Christmas, and it was fanta
stic. Would recomme… https://t.co/zb8T1EBJpH
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: This is sickening and terrifying.
------------------------------------------
Tweet Text: Almost 26,000 food vouchers were given out ahead of the Christ
mas holidays
------------------------------------------
Tweet Text: @madinthehat Merry Ukrainian Christmas to you!!! https://t.co/
mSss67xnNH
------------------------------------------
```

4. When the top10 elements are computed, copy the output and paste in your submission.



```
----------- 2020-12-29 11:33:28 -----------
+--------------------+-------------+
|             hashtag|hashtag_count|
+--------------------+-------------+
|          #Christmas|           14|
|          #christmas|           11|
|               #MAGA|            8|
|          #OnThisDay|            7|
|#JiggabanTalkWith...|            7|
|            #gouache|            7|
|        #tuesdayvibe|            4|
|            #DonKiss|            4|
|         #DonKissFam|            4|
|          #mincemeat|            3|
+--------------------+-------------+
```

## Bonus points

1. Use your own Access credentials. Tip: You need to create an application from your twitter account.

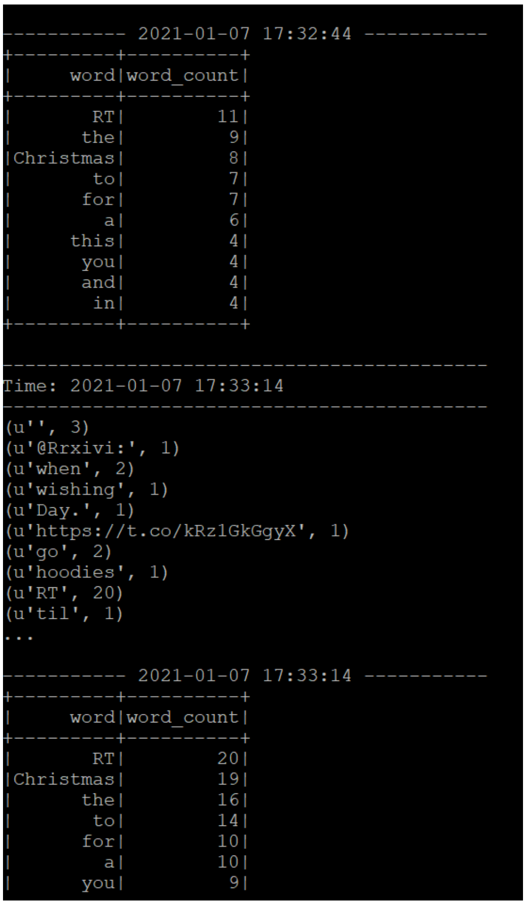
   ACCESS_TOKEN = '1343553375695216640-wnzw62iEmRMjxfLNoZAkPAkwjh0nJ6'

   ACCESS_SECRET = 'ynHVRKAS0Tr4zvnr47L1lkbR2TxEq619cbU7ZiEUbgDeM'

   CONSUMER_KEY = 'MC07tJC8bT237UFzxYwal9Yte'

   CONSUMER_SECRET = 'CtndQG4syKwpGC09OHsQnDSH8pUy66rypw5EsGN7JD64d8H7EP'

2. Instead of computing the top10 elements with Spark SQL, change the code to obtain the Top10 words (not only hashtags) using a moving window of 10 minutes every 30 seconds. Copy & paste the result.



```
----------- 2021-01-07 17:32:44 -----------
+---------+----------+
|     word|word_count|
+---------+----------+
|       RT|        11|
|      the|         9|
|Christmas|         8|
|       to|         7|
|      for|         7|
|        a|         6|
|     this|         4|
|      you|         4|
|      and|         4|
|       in|         4|
+---------+----------+

-------------------------------------------
Time: 2021-01-07 17:33:14
-------------------------------------------
(u'', 3)
(u'@Rrxivi:', 1)
(u'when', 2)
(u'wishing', 1)
(u'Day.', 1)
(u'https://t.co/kRz1GkGgyX', 1)
(u'go', 2)
(u'hoodies', 1)
(u'RT', 20)
(u'til', 1)
...

----------- 2021-01-07 17:33:14 -----------
+---------+----------+
|     word|word_count|
+---------+----------+
|       RT|        20|
|Christmas|        19|
|      the|        16|
|       to|        14|
|      for|        10|
|        a|        10|
|      you|         9|
```

3. Connect Spark Streaming with Kafka.

   Was not able to test out the code below as cluster did not allow me to create a topic but code to connect to Kafka should be as shown below:

   KafkaUtils.createDirectStream(ssc, ['topic'], {"bootstrap.servers":“localhost:9009"})