IE SCHOOL OF HUMAN SCIENCES AND TECHNOLOGY

STREAMING FROM TWITTER

Individual Assignment

by Jonas Hellevang

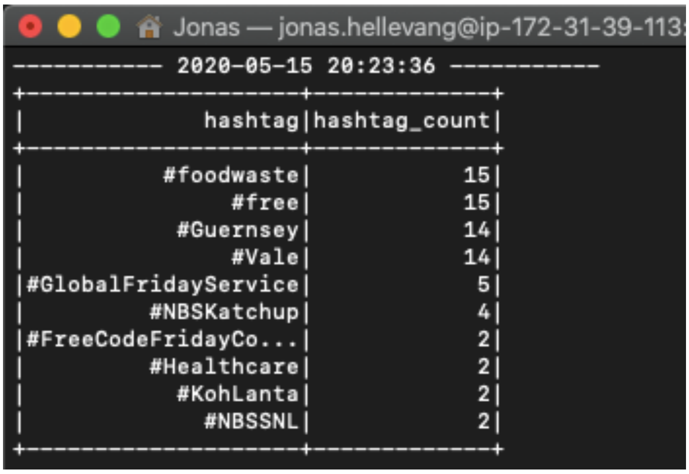
Prof. Federico Castanedo Sotela

Step 1:

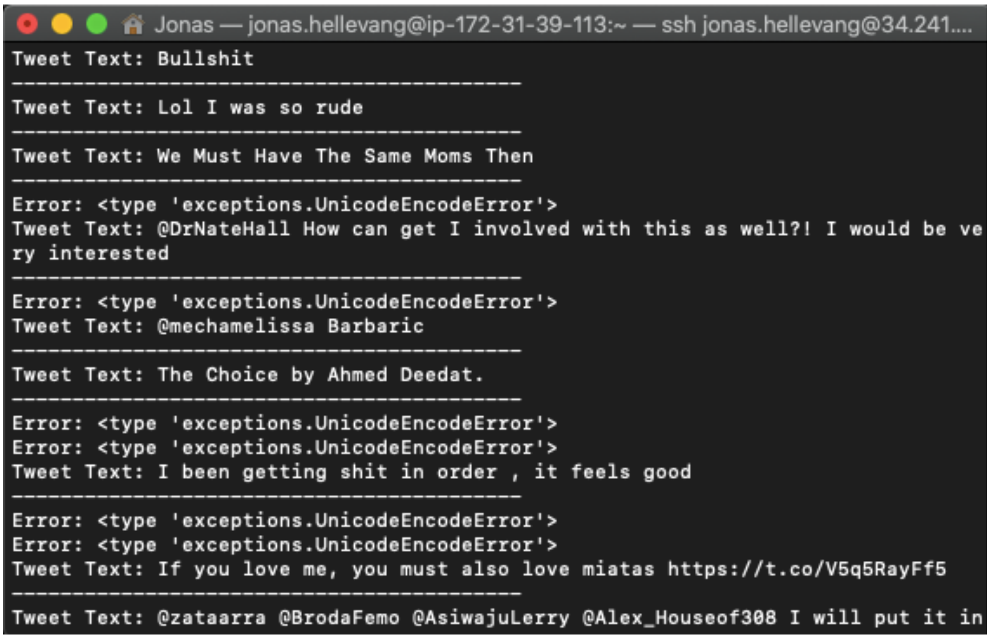
```
# setting a checkpoint to allow RDD recovery
ssc.checkpoint("Checkpoint_Jonas")

# read data from port 1994
dataStream = ssc.socketTextStream("localhost", 1994)

# split each tweet into words
words = dataStream.flatMap(lambda line: line.split(" "))

# adding the count of each hashtag to its last count
tags_totals = hashtags.updateStateByKey(aggregate_tags_count)
```

Files: “twitter_app.py” and “spark_streaming_twitter.py”

Step 2:



```
----------- 2020-05-15 20:23:36 -----------
+--------------------+-------------+
|             hashtag|hashtag_count|
+--------------------+-------------+
|          #foodwaste|           15|
|               #free|           15|
|           #Guernsey|           14|
|               #Vale|           14|
|#GlobalFridayService|            5|
|         #NBSKatchup|            4|
|#FreeCodeFridayCo...|            2|
|         #Healthcare|            2|
|           #KohLanta|            2|
|             #NBSSNL|            2|
+--------------------+-------------+
```


Jonas — jonas.hellevang@ip-172-31-39-113:~ — ssh jonas.hellevang@34.241....

```
Tweet Text: Bullshit
------------------------------------------
Tweet Text: Lol I was so rude
------------------------------------------
Tweet Text: We Must Have The Same Moms Then
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: @DrNateHall How can get I involved with this as well?! I would be ve
ry interested
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: @mechamelissa Barbaric
------------------------------------------
Tweet Text: The Choice by Ahmed Deedat.
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: I been getting shit in order , it feels good
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: If you love me, you must also love miatas https://t.co/V5q5RayFf5
------------------------------------------
Tweet Text: @zataarra @BrodaFemo @AsiwajuLerry @Alex_Houseof308 I will put it in
```

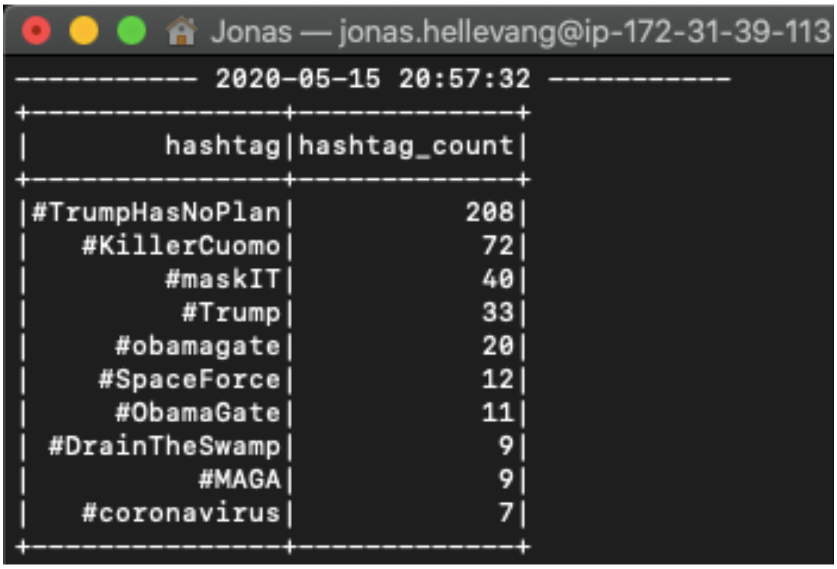
Step 3:

For this exercise I decided to do tweets with the word trump, as well as see what Texas has to say about him. I don’t think Texas is a very big fan of Trump tonight!

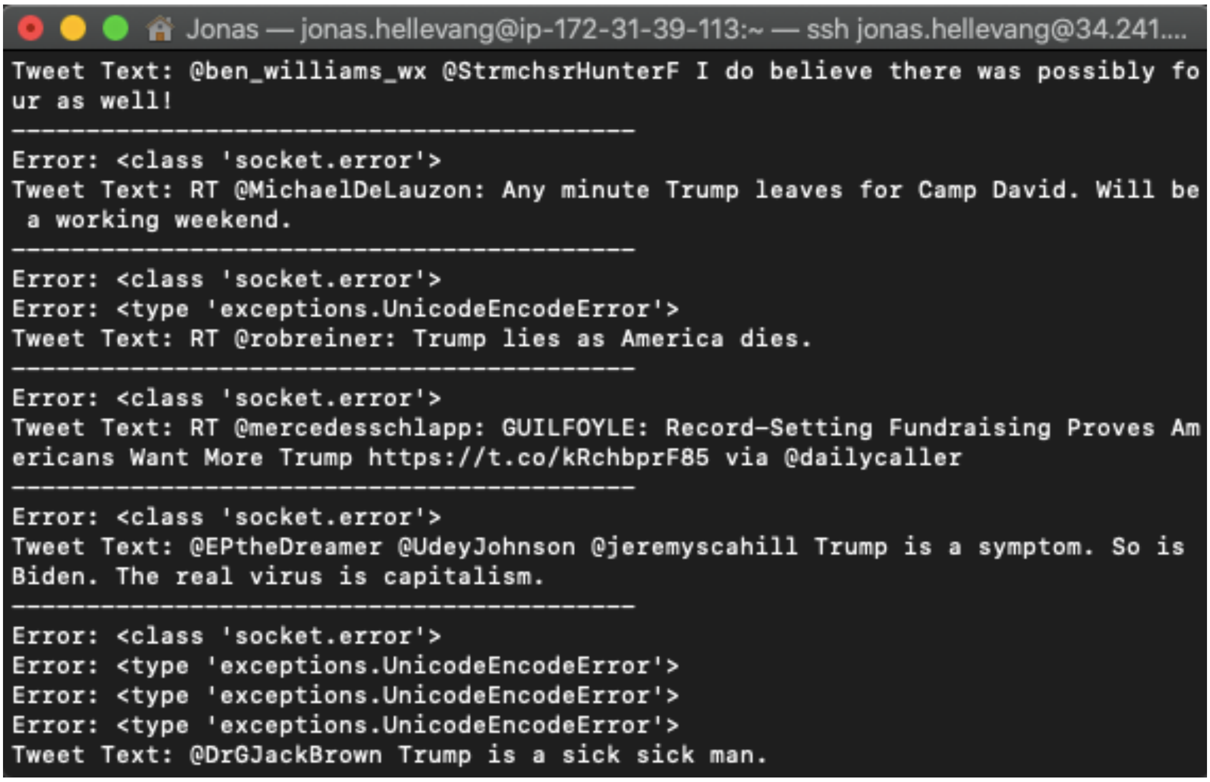
This is the code I used for the bounding box of Texas and tracking Trump:

query_data = [('language', 'en'), ('locations','-106.645646,25.837377,-93.508292,36.500704'),('track','trump')]

Files: “twitter_app_1.py” and “spark_streaming_twitter.py”



```
----------- 2020-05-15 20:57:32 -----------
+----------------+--------------+
|         hashtag|hashtag_count|
+----------------+--------------+
|#TrumpHasNoPlan|           208|
|   #KillerCuomo|            72|
|         #maskIT|            40|
|          #Trump|            33|
|      #obamagate|            20|
|     #SpaceForce|            12|
|      #ObamaGate|            11|
|  #DrainTheSwamp|             9|
|           #MAGA|             9|
|    #coronavirus|             7|
+----------------+--------------+
```

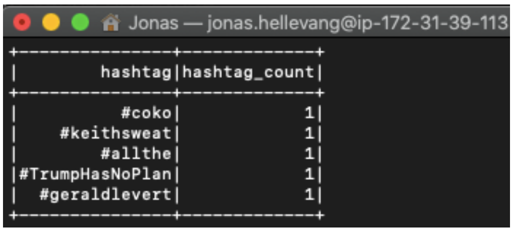

Jonas — jonas.hellevang@ip-172-31-39-113:~ — ssh jonas.hellevang@34.241....

```
Tweet Text: @ben_williams_wx @StrmchsrHunterF I do believe there was possibly fo
ur as well!
------------------------------------------
Error: <class 'socket.error'>
Tweet Text: RT @MichaelDeLauzon: Any minute Trump leaves for Camp David. Will be
 a working weekend.
------------------------------------------
Error: <class 'socket.error'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: RT @robreiner: Trump lies as America dies.
------------------------------------------
Error: <class 'socket.error'>
Tweet Text: RT @mercedesschlapp: GUILFOYLE: Record-Setting Fundraising Proves Am
ericans Want More Trump https://t.co/kRchbprF85 via @dailycaller
------------------------------------------
Error: <class 'socket.error'>
Tweet Text: @EPtheDreamer @UdeyJohnson @jeremyscahill Trump is a symptom. So is
Biden. The real virus is capitalism.
------------------------------------------
Error: <class 'socket.error'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: @DrGJackBrown Trump is a sick sick man.
```

Step 4:

See above as the question is the same as step 3 which says: “Copy and paste some output lines of this example”. If you wanted to see a different output, I have here included an output from before it reached 10, as it isn’t fully counting to 10 from the very beginning:

Files: “twitter_app_1.py” and “spark_streaming_twitter.py”

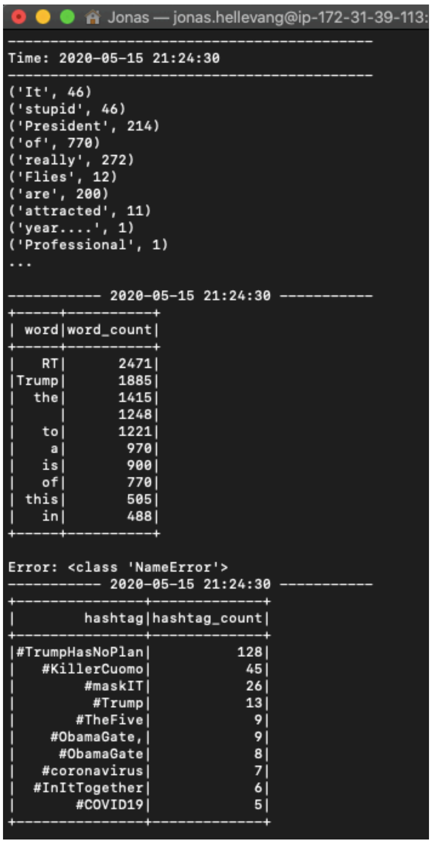


```
+---------------+-------------+
|        hashtag|hashtag_count|
+---------------+-------------+
|          #coko|            1|
|    #keithsweat|            1|
|        #allthe|            1|
|#TrumpHasNoPlan|            1|
|  #geraldlevert|            1|
+---------------+-------------+
```

Step 5: Top10 elements and Top10 w/ moving window of 10 min every 30 seconds.

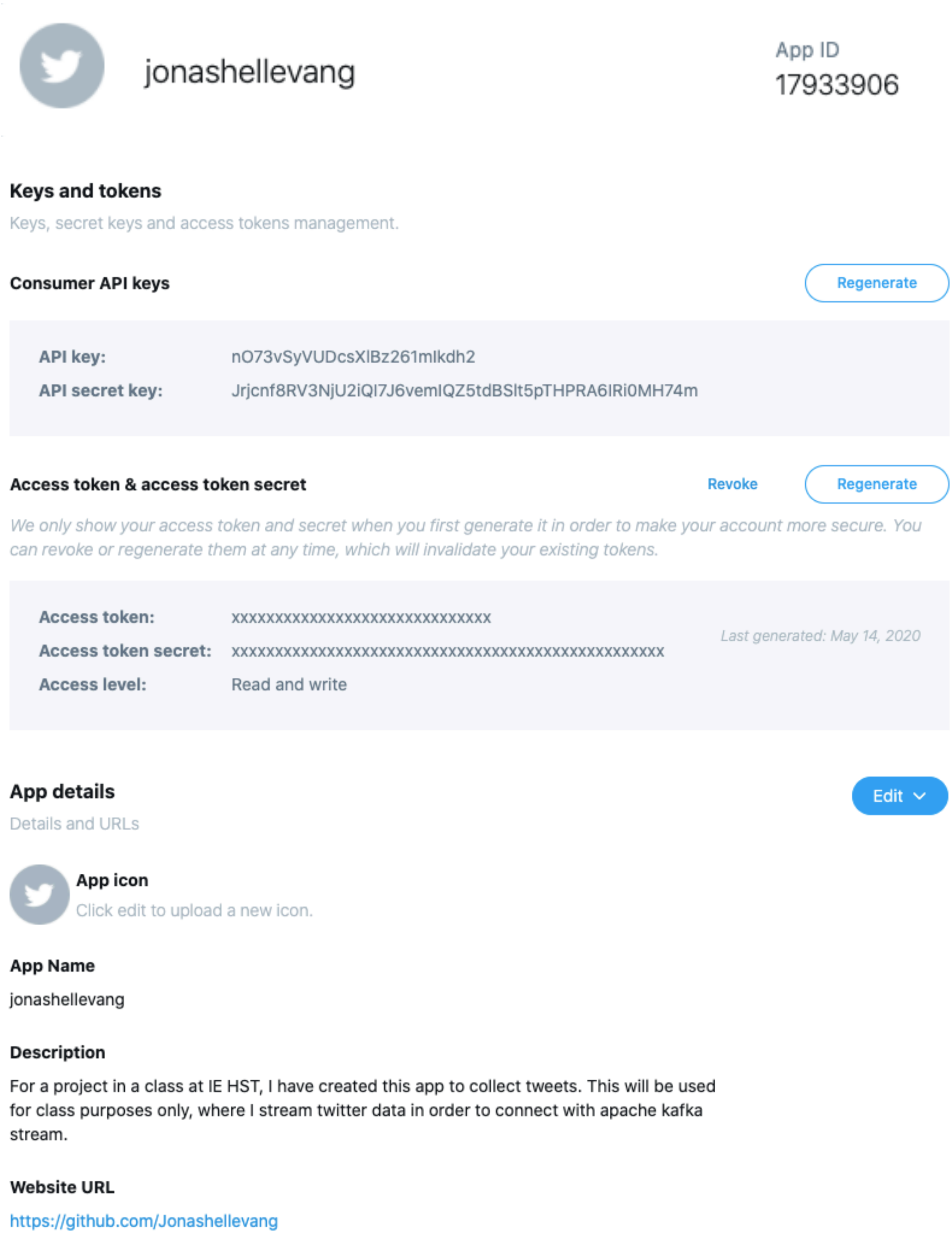
Here I interpreted that I will combine the code and do both top10 elements and the top 10 words. I checked earlier for this output and was not planning to use Trump for either of these exercises, but I thought it was too funny that the top trending hashtag of Trump is #TrumpHasNoPlan. Below is the output of this, still using Trump as tracking word and Texas as location:

Files: “twitter_app_1.py” and “spark_streaming_twitter_1.py”



```
-------------------------------------------
Time: 2020-05-15 21:24:30
-------------------------------------------
('It', 46)
('stupid', 46)
('President', 214)
('of', 770)
('really', 272)
('Flies', 12)
('are', 200)
('attracted', 11)
('year....', 1)
('Professional', 1)
...

----------- 2020-05-15 21:24:30 -----------
+-----+----------+
| word|word_count|
+-----+----------+
|   RT|      2471|
|Trump|      1885|
|  the|      1415|
|     |      1248|
|   to|      1221|
|    a|       970|
|   is|       900|
|   of|       770|
| this|       505|
|   in|       488|
+-----+----------+

Error: <class 'NameError'>
----------- 2020-05-15 21:24:30 -----------
+---------------+-------------+
|        hashtag|hashtag_count|
+---------------+-------------+
|#TrumpHasNoPlan|          128|
|   #KillerCuomo|           45|
|        #maskIT|           26|
|         #Trump|           13|
|       #TheFive|            9|
|    #ObamaGate,|            9|
|     #ObamaGate|            8|
|   #coronavirus|            7|
|  #InItTogether|            6|
|       #COVID19|            5|
+---------------+-------------+
```

Bonus 1 – Using my own Access credentials:



jonashellevang

App ID
17933906

Keys and tokens

Keys, secret keys and access tokens management.

Consumer API keys

Regenerate

API key: nO73vSyVUDcsXlBz261mlkdh2

API secret key: Jrjcnf8RV3NjU2iQI7J6vemIQZ5tdBSlt5pTHPRA6IRi0MH74m

Access token & access token secret

Revoke

Regenerate

We only show your access token and secret when you first generate it in order to make your account more secure. You can revoke or regenerate them at any time, which will invalidate your existing tokens.

Access token: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Access token secret: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Access level: Read and write

Last generated: May 14, 2020

App details

Details and URLs

Edit

App icon

Click edit to upload a new icon.

App Name

jonashellevang

Description

For a project in a class at IE HST, I have created this app to collect tweets. This will be used for class purposes only, where I stream twitter data in order to connect with apache kafka stream.

Website URL

https://github.com/Jonashellevang

Bonus 2 – Flajolet-Martin Algorithm:

I tried a couple things for this one, but if I have correctly understood the exercise, doing a top hashtag does not make sense using Flajolet-Martin. Flajolet-Martin takes the distinct values, and I hope that from this exercise you wanted us to understand this use case for it.

Bonus 3 – Connecting Spark Stream with Kafka:

Here I'm still using Texas and Trump, but only doing hashtags.

Files: "twitter_app_2.py" and "spark_streaming_twitter_2.py"

"twitter_app_2.py" changes:

```
def send_tweets_to_spark(http_resp, producer, topic):
   for line in http_resp.iter_lines():
      try:
         full_tweet = json.loads(line)
         tweet_text = full_tweet['text']
         print("Tweet Text: " + tweet_text)
         print ("------------------------------------------")
         producer.send(topic,str(tweet_text))
      except:
         e = sys.exc_info()[0]
         print("Error: %s" % e)

topic = "Jonas_Hellevang"
producer = KafkaProducer(bootstrap_servers= "data3:6667")

print("Connected... Starting getting tweets.")
resp = get_tweets()
send_tweets_to_spark(resp, producer, topic)
```

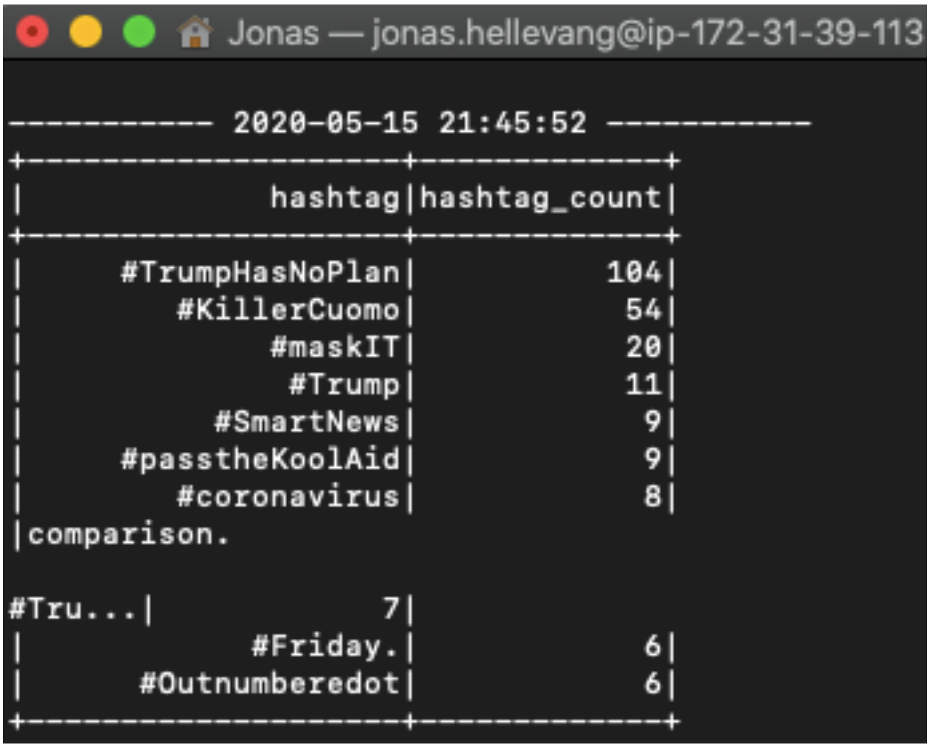
"spark_streaming_twitter_2.py" changes:

```
# read data from Kafka Topic
topic= "Jonas_Hellevang"
brokers= "data3:6667"
dataStream = KafkaUtils.createDirectStream(ssc, [topic],{"metadata.broker.list": brokers})
parsed = dataStream.map(lambda x: x[1])
```

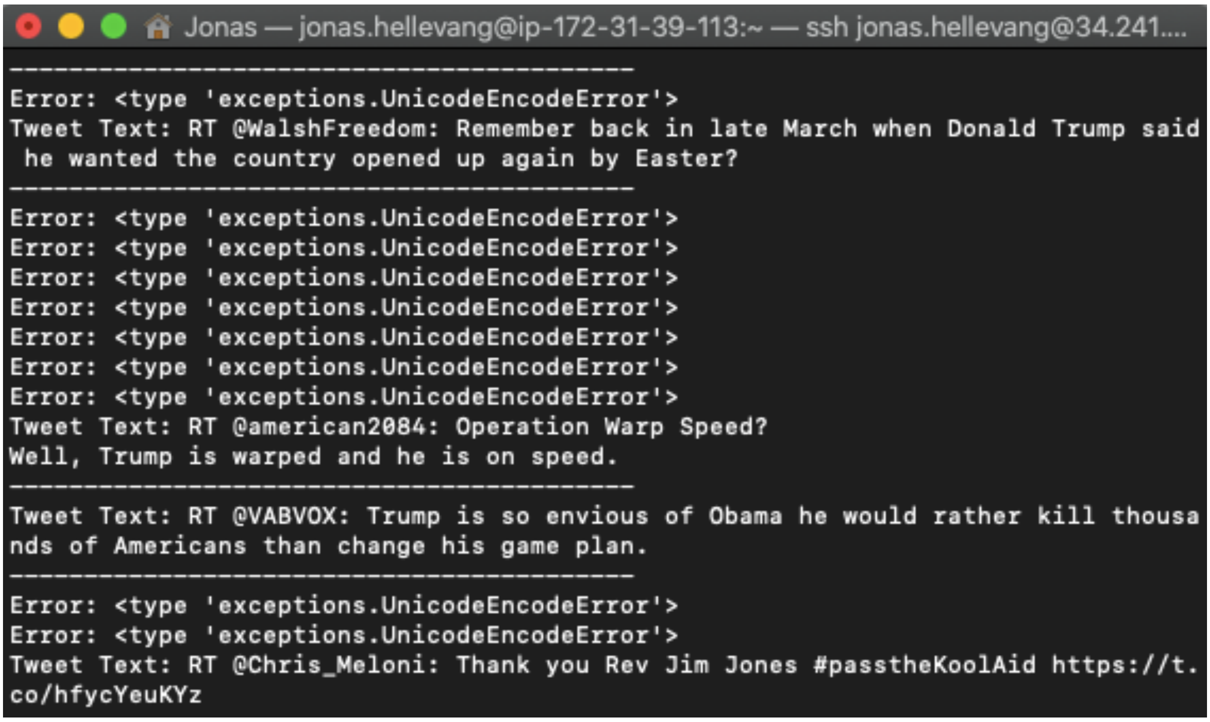
to be able to consume from kafka topic, I run this code:

```
/usr/hdp/current/spark-client/bin/spark-submit --packages org.apache.spark:spark-streaming-kafka_2.10:1.6.3 spark_streaming_twitter_2.py
```

From running the code for both consumer and producer, I get this output:



```
----------- 2020-05-15 21:45:52 -----------
+--------------------+-------------+
|             hashtag|hashtag_count|
+--------------------+-------------+
|     #TrumpHasNoPlan|          104|
|        #KillerCuomo|           54|
|             #maskIT|           20|
|              #Trump|           11|
|          #SmartNews|            9|
|     #passtheKoolAid|            9|
|        #coronavirus|            8|
|comparison.

#Tru...|            7|
|            #Friday.|            6|
|      #Outnumberedot|            6|
+--------------------+-------------+
```


Jonas — jonas.hellevang@ip-172-31-39-113:~ — ssh jonas.hellevang@34.241....

```
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: RT @WalshFreedom: Remember back in late March when Donald Trump said
 he wanted the country opened up again by Easter?
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: RT @american2084: Operation Warp Speed?
Well, Trump is warped and he is on speed.
------------------------------------------
Tweet Text: RT @VABVOX: Trump is so envious of Obama he would rather kill thousa
nds of Americans than change his game plan.
------------------------------------------
Error: <type 'exceptions.UnicodeEncodeError'>
Error: <type 'exceptions.UnicodeEncodeError'>
Tweet Text: RT @Chris_Meloni: Thank you Rev Jim Jones #passtheKoolAid https://t.
co/hfycYeuKYz
```