1)CREATE AN ACCOUNT IN TWEETER APP. Get app

api\_key = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

api\_secret = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

access\_token = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

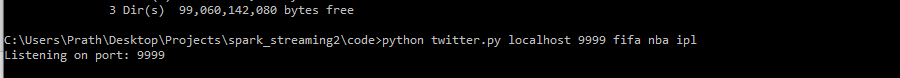
access\_token\_secret = "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"

2)WRITE PROGRAM TO CONNECT TO TWITTER STREAMING API USING TWEEPY.

3) RUN THE PYTHON PROGRAM.

**Twitter.py localhost 9999 fifa nba ipl**

**Twitter.py localhost 9999 fifa nba corona**



4) WRITE SPARK PROGRAM TO FILTER AND PROCESS DATA (Count hashtag in twitter to find the overall trend.)

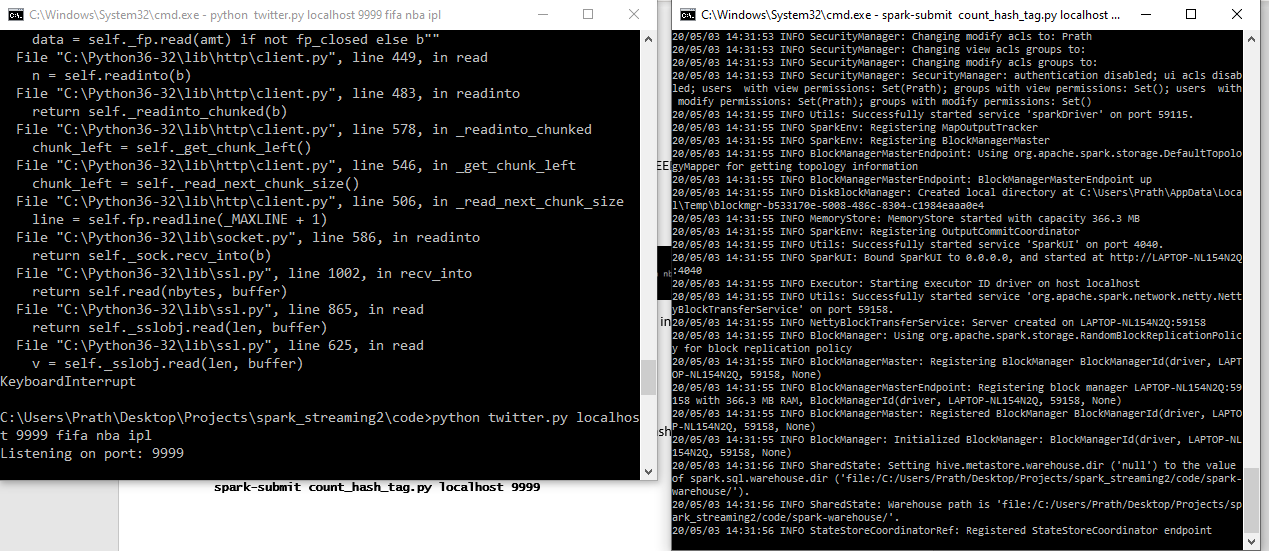
4.1) Need to connect to a host and port.

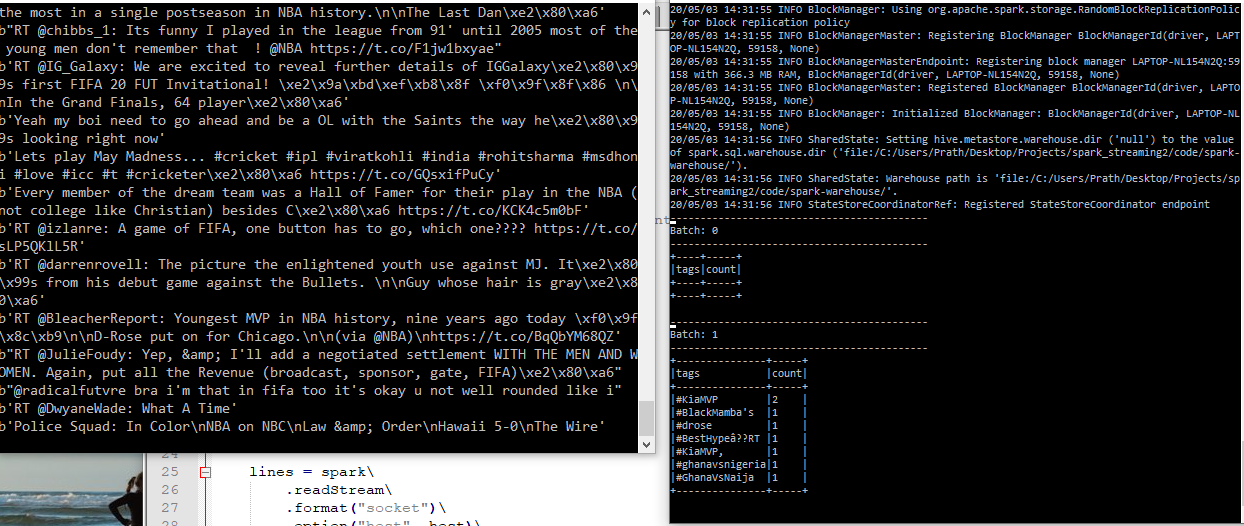
4.2) Use spark session to read streaming data

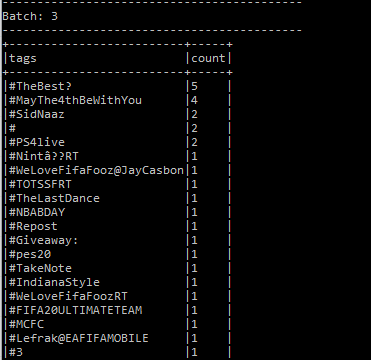
4.3) Extract the message and split the message to get individual hash tag

5) Submit the spark job

**spark-submit count\_hash\_tag.py localhost 9999**



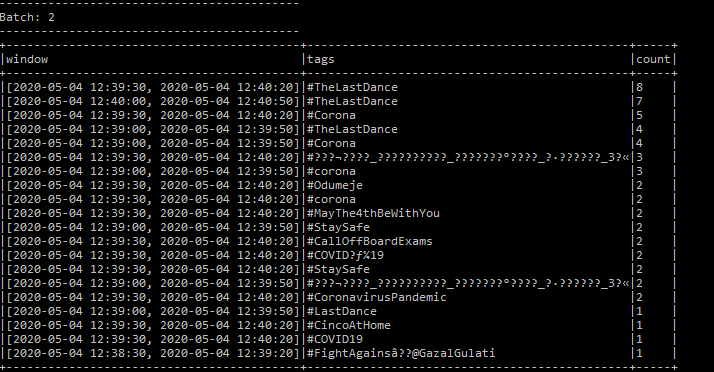




6) Include window function to see the trending tweets for every 50 seconds, with sliding window of 30 seconds.

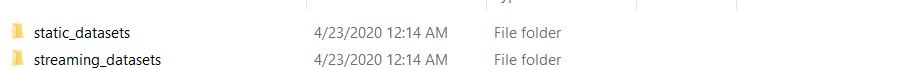
**spark-submit count\_hash\_tag\_window.py localhost 9999**

**Output**

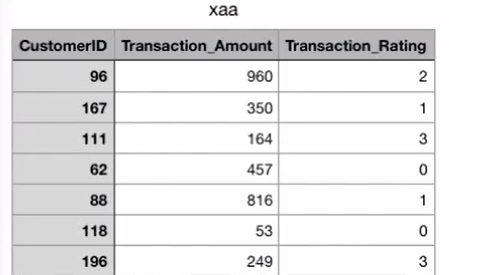


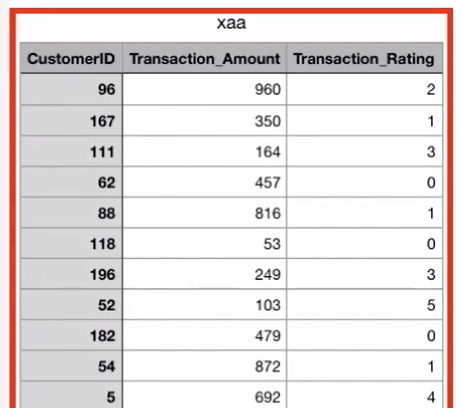
7) Join with batch and streaming data

Within the dataset folder we have



This contains customer specific data transactions done by the customers.

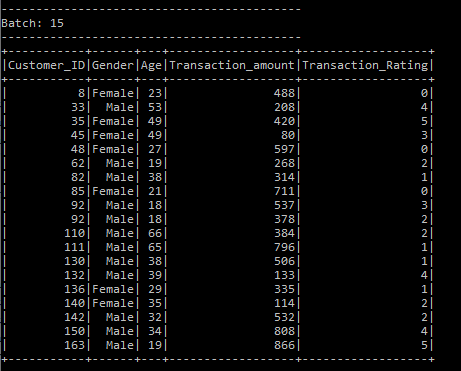




Run:

**spark-submit join\_batch\_streaming.py localhost 9999**

Output



8) GET THE AVERAGE TRANSACTION AMOUNT PER GENDER

**spark-submit join\_batch\_aggregate.py localhost 9999**