**Final project report**

**Twitter Streaming Using Apache Spark socket**

Student Name: Nuha Ibraheem

Course name: cloud computing

There is two ways to work with streaming using spark we can use a distributer like **kafa**, other way is using socket, here is my steps using socket for Twitter API streaming analysis, I used twitter API to get the popular tweets for specific topic and the related hashtags to this topic.

**USING SOCKET**

1. Get my keys to access to twitter from twitter.developer

2. Write my code

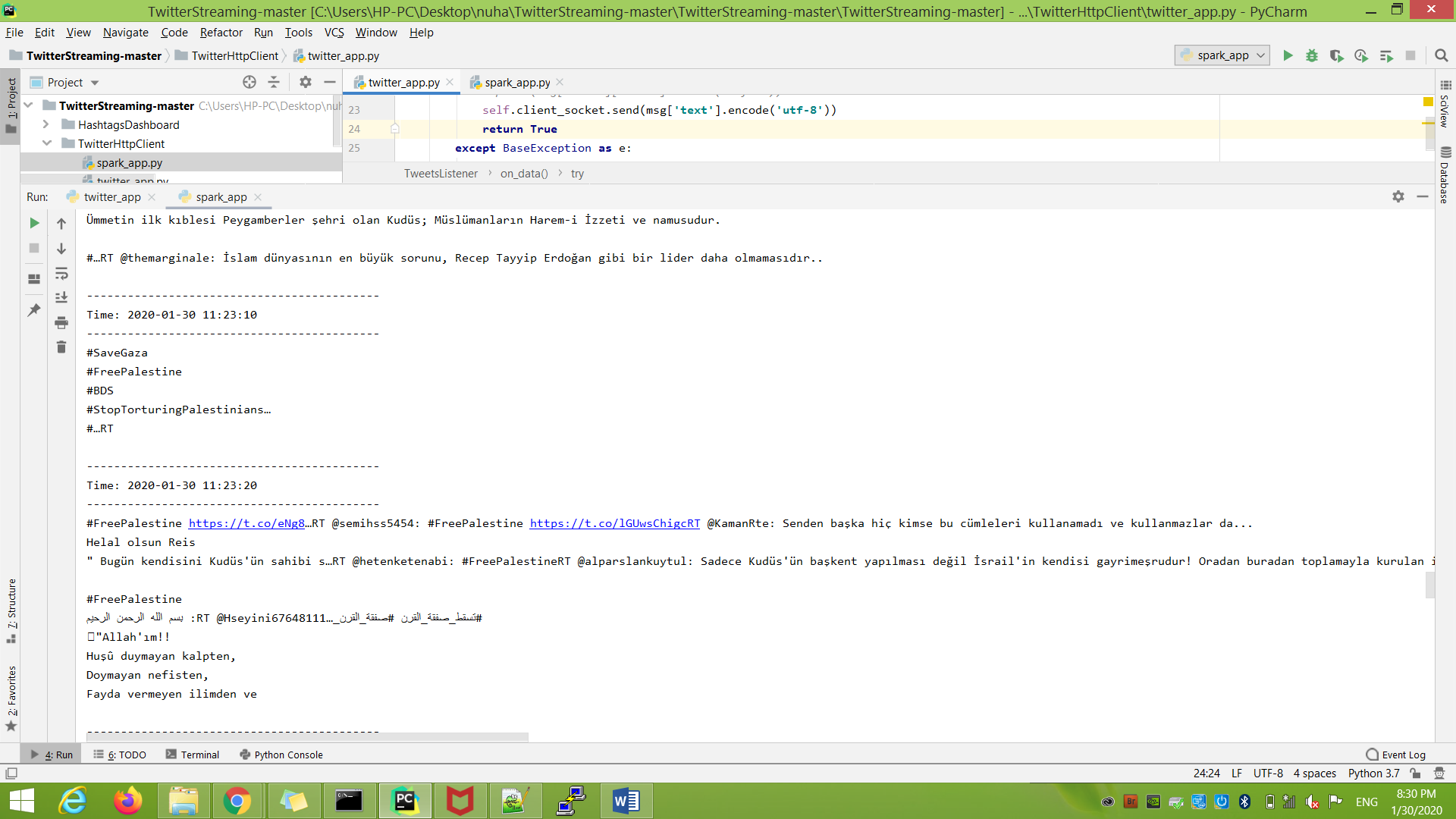
A. In Twitter\_App.py:

* I import tweepy which is library in python handling twitter’s API.
* Then I imported OAuthHandler from tweepy which is used for authentication protocol of twitter API.
* Import stream from tweepy to import stream & stream listener, listening to stream data get from twitter.
* Import socket package to create socket connection with localhost or port in local machine.
* Import json because data came in Jason format from all the world different languages.
* I used my keys to connect to API.
* Class tweetsListener.
* Class on\_Data.
* Class set\_Data.
* Create object from socket.
* Call send\_Data function.

B. In Spark\_App.py:

Import PySpark, Spark context.

* Create spark context and streaming context.
* Set socket\_stream to create stream.
* Make word count for hashtags.
* Sort hashtags (popular one).



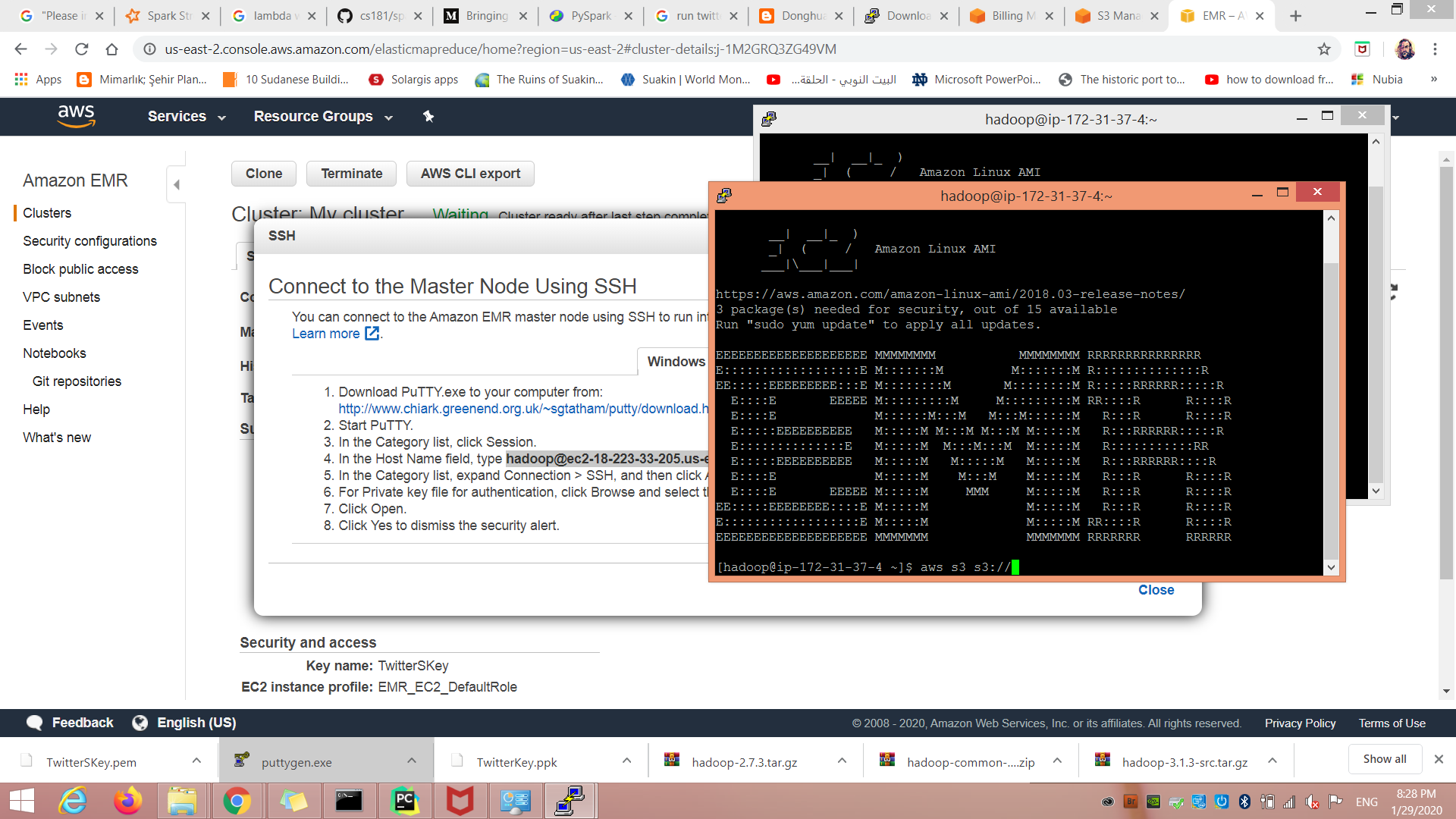
Running the project in pycharm

**Running the project in EMR cluster:**

* Setting up aws EMR to use in spark.
* Upload the project’s files twitter\_app.py, spark\_app.p.
* Upload bash file to install tweepy library in EC2 the file contains

Pip install tweepy.

* Open putty and upload the project’s pair keys and open session.
* One session for opening a port for the twitter streaming.
* The other session for splitting the tweets and specifying the hashtags.



* Upload files from s3 using aws s3 cp s3://your bucket name/file name .
* Running project using spark-submit file name.py
* I faced this error while running my files

UnicodeEncodeError: 'ascii' codec can't encode character

* It’s solved by adding PYTHONIOENCODING=utf8 before spark-submit file name
* And it’s running …