CS 6350

ASSIGNMENT \_\_\_\_\_\_3\_\_\_\_\_\_

Names of students in your group:

Rui Yang

Number of free late days used: \_\_\_\_\_\_\_\_\_\_2\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

<https://docs.tweepy.org/en/stable/streaming_how_to.html>

<https://docs.tweepy.org/en/stable/extended_tweets.html>

<https://www.digitalocean.com/community/tutorials/how-to-perform-sentiment-analysis-in-python-3-using-the-natural-language-toolkit-nltk>

<https://www.storybench.org/how-to-collect-tweets-from-the-twitter-streaming-api-using-python/>

<https://github.com/robinhood/faust/blob/master/examples/kubernetes/producer/producer.py>

<https://spark.apache.org/docs/2.1.0/streaming-kafka-0-8-integration.html>

<https://www.rittmanmead.com/blog/2017/01/getting-started-with-spark-streaming-with-python-and-kafka/>

<https://stackoverflow.com/questions/44320699/how-to-concatenate-append-multiple-spark-dataframes-column-wise-in-pyspark>

<https://towardsdatascience.com/build-dynamic-visualizations-and-dashboards-with-kibana-for-data-analysis-using-twitter-feeds-6f516bf803a3>

<https://medium.com/@pakanon.pk/tweets-sentimental-analysis-with-spark-streaming-kafka-elasticsearch-and-kibana-bb3d161e1c3b>

<https://towardsdatascience.com/sentiment-analysis-on-streaming-twitter-data-using-spark-structured-streaming-python-fc873684bfe3>

1.code file:

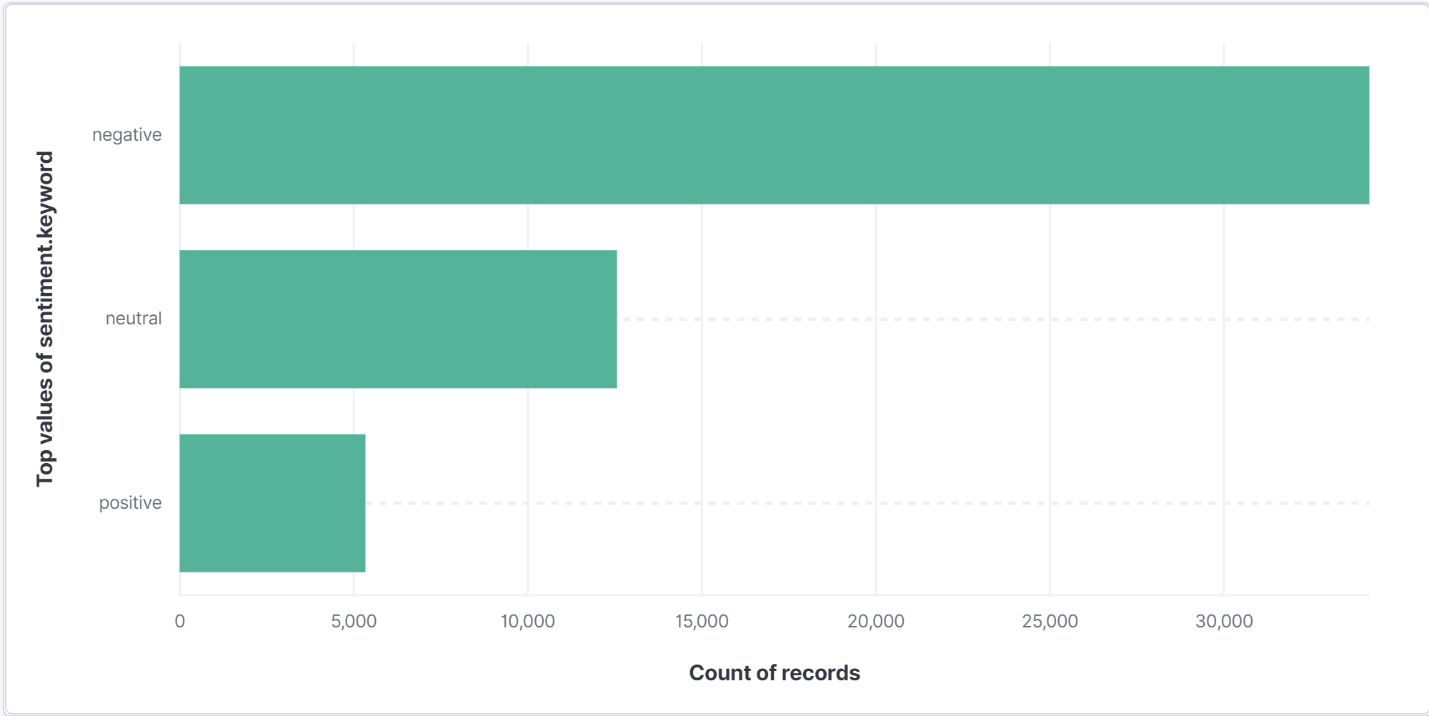
Twitter.py: <https://cs6350a3.s3.amazonaws.com/Q1/twitter.py>

Analysis.py:

<https://cs6350a3.s3.amazonaws.com/Q1/analysis.py>

Readme:

<https://cs6350a3.s3.amazonaws.com/Q1/Q1README.txt>

Graph: 

I didn’t have time to run the program for a long time. And after I got this graph, kibana suddenly doesn’t work, it’s the main reason why I only one graph. But I think my code should work fine. I’m searching “China”. It’s obvious that most common are negative thing. It’s a little sad to know that.

2.

Readme:

<https://cs6350a3.s3.amazonaws.com/Q2/Q2README.txt>

Social network package:

<https://cs6350a3.s3.amazonaws.com/Q2/twitter_combined.txt>

Code:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8750190078783311/160913008196708/5504697096027154/latest.html>

A little mistake, when output the result, it’s Q2 first, then Q1.

According to the result, it's obvious that node 115485051 has the highest PageRank. It means this play an very important role vs other nodes. And node 40981798 has the highest triangle count, it means it connect with many nodes, it's one of the most important node. So when we use different way to measure, the result may be different.