CS 6350

Names of students in your group:

Meghana Sai Bijivemula - MXB210011

Number of free late days used: 0

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.

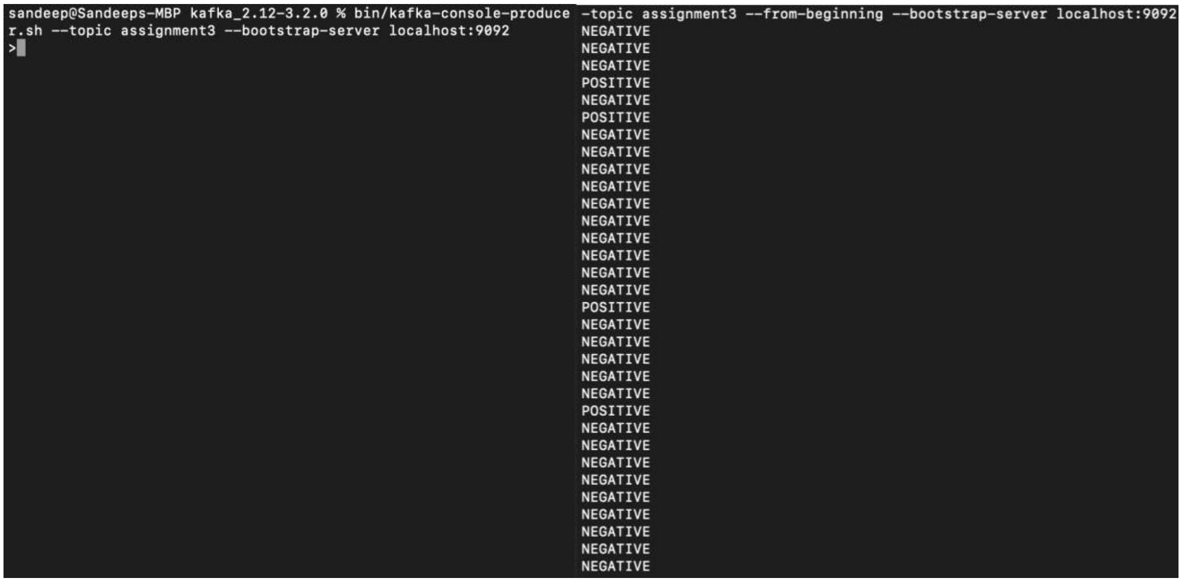
References:

https://graphframes.github.io/graphframes/docs/_site/user-guide.html#strongly-connected-components
https://graphframes.github.io/graphframes/docs/_site/user-guide.html#pagerank
https://www.elastic.co/guide/index.html
https://docs.tweepy.org/en/stable/
https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-the-twitter-api

1. Spark Streaming with Twitter and Kafka

The motive of this assignment is to perform sentiment analysis and then classify the tweets from Twitter into Positive and Negative
#covid19 is the filter that has been used for this assignment.

Output from the terminal based on sentiment analysis:

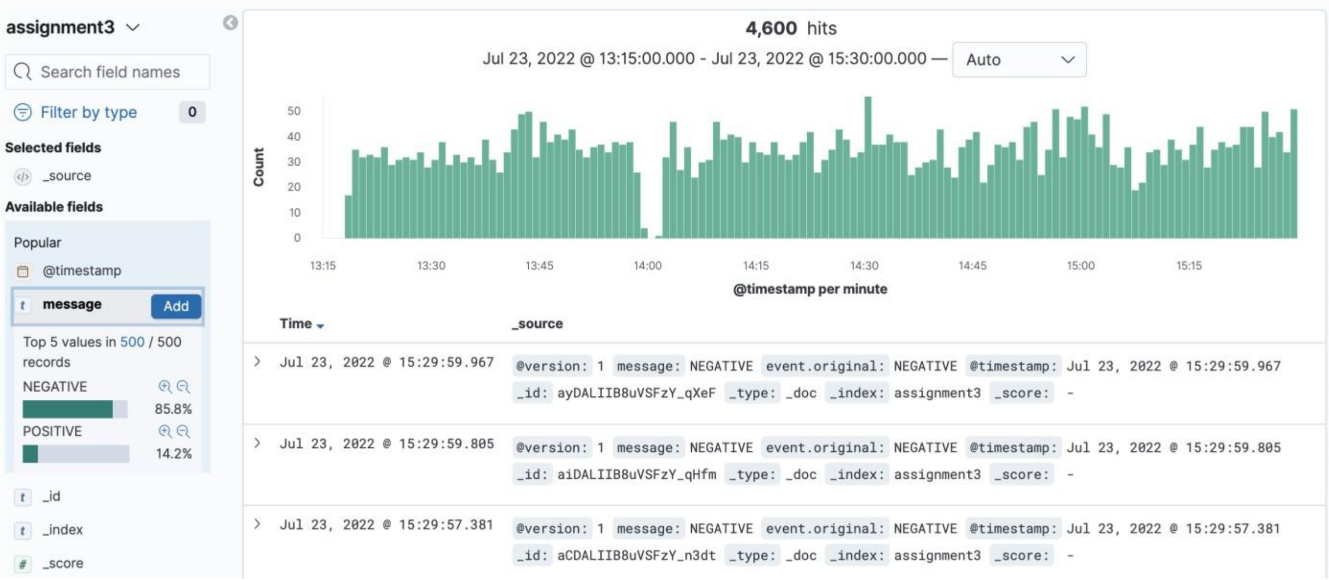


```
sandeep@Sandeeps-MBP kafka_2.12-3.2.0 % bin/kafka-console-produce
r.sh --topic assignment3 --bootstrap-server localhost:9092
>
```

```
-topic assignment3 --from-beginning --bootstrap-server localhost:9092
NEGATIVE
NEGATIVE
NEGATIVE
POSITIVE
NEGATIVE
POSITIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
POSITIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
POSITIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
NEGATIVE
```

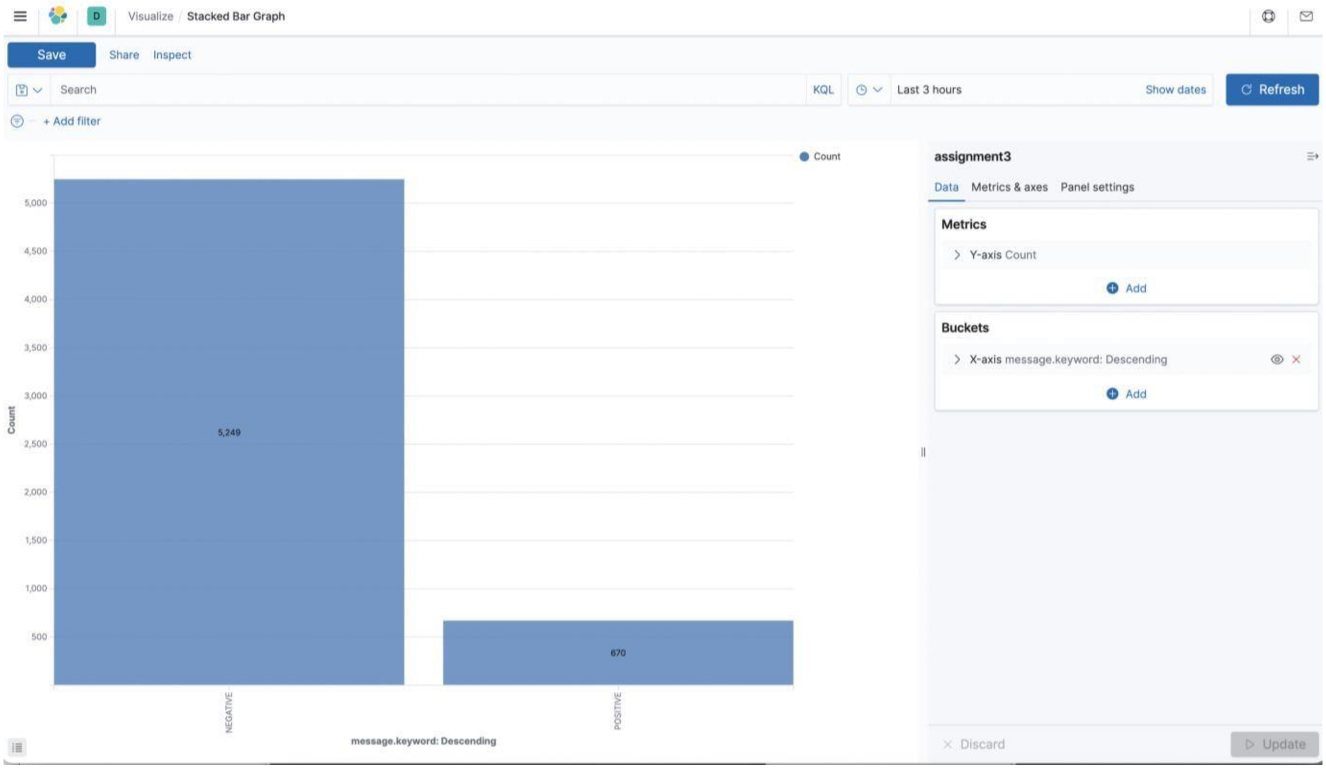
After performing the sentiment analysis, visualized the processed tweets using ELK Stack.

Below is the output of the visualization of the processed data from the ELK stack:



```
Time          _source
> Jul 23, 2022 @ 15:29:59.967   @version: 1 message: NEGATIVE event.original: NEGATIVE @timestamp: Jul 23, 2022 @ 15:29:59.967
                                _id: ayDALIIB8uVSFzY_qXeF _type: _doc _index: assignment3 _score: -
> Jul 23, 2022 @ 15:29:59.805   @version: 1 message: NEGATIVE event.original: NEGATIVE @timestamp: Jul 23, 2022 @ 15:29:59.805
                                _id: aiDALIIB8uVSFzY_qHfm _type: _doc _index: assignment3 _score: -
> Jul 23, 2022 @ 15:29:57.381   @version: 1 message: NEGATIVE event.original: NEGATIVE @timestamp: Jul 23, 2022 @ 15:29:57.381
                                _id: aCDALIIB8uVSFzY_n3dt _type: _doc _index: assignment3 _score: -
```

The below graphs represent the categorization of the overall tweets obtained from Twitter for a 2-3 hour time interval:



assignment3

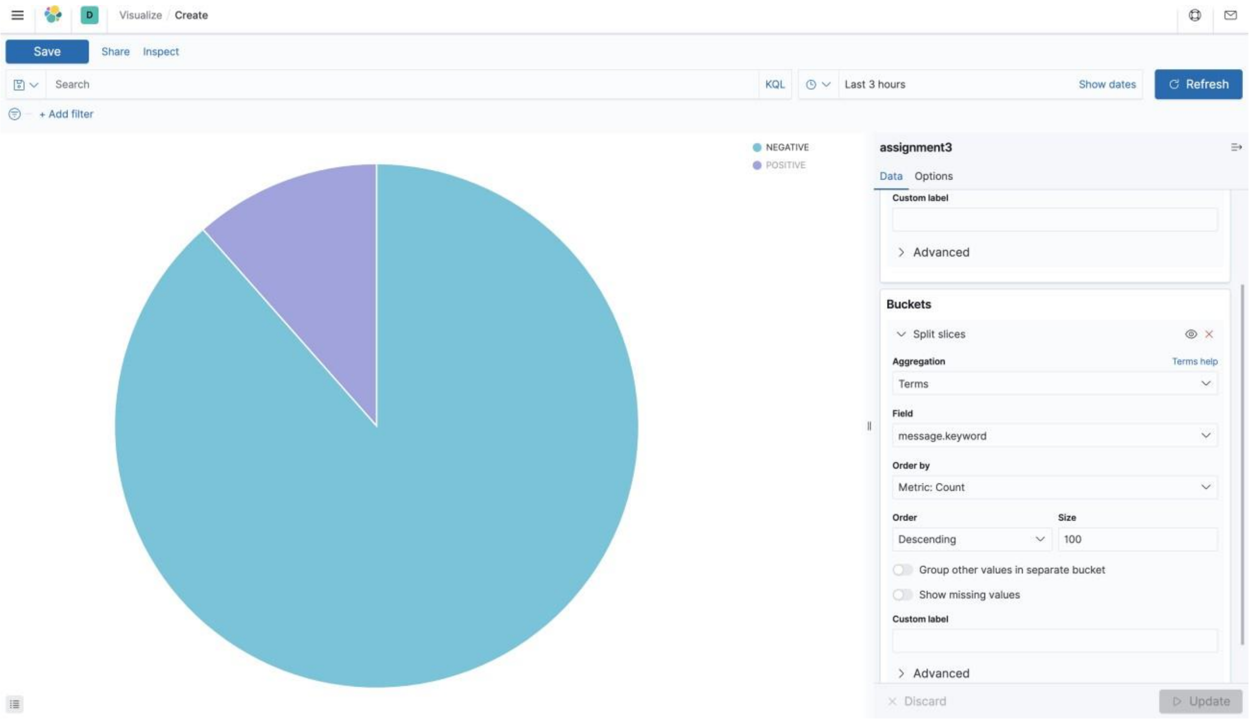
Metrics

Y-axis Count

Buckets

X-axis message.keyword: Descending

The variation of the positive and negative tweets over a period of time in the form of pie chart:



assignment3

Custom label

Buckets

Split slices

Aggregation: Terms

Field: message.keyword

Order by: Metric: Count

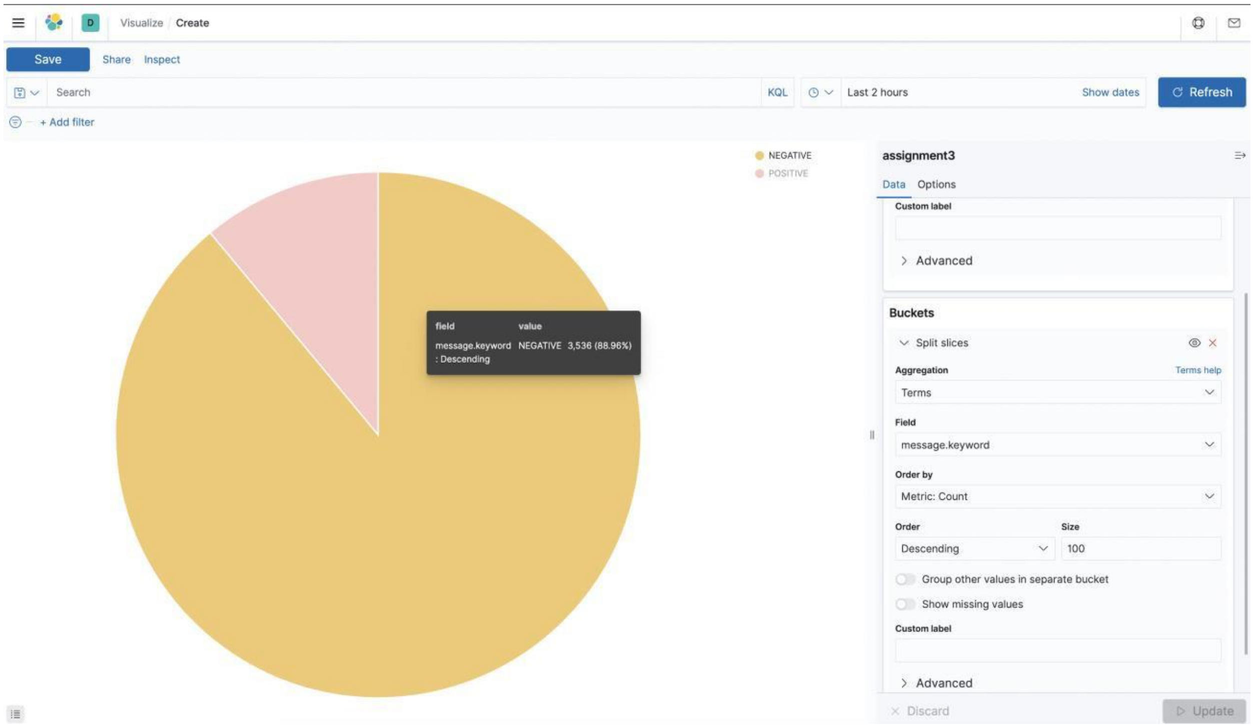
Order: Descending

Size: 100

Group other values in separate bucket

Show missing values

Custom label



field	value
message.keyword : Descending	NEGATIVE 3,536 (88.96%)

Custom label

Aggregation: Terms

Field: message.keyword

Order by: Metric: Count

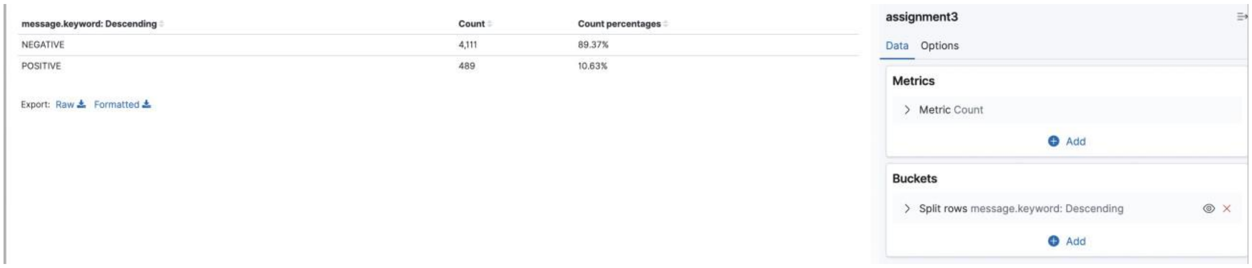
Order: Descending

Size: 100

○ Group other values in separate bucket

○ Show missing values

Custom label



message.keyword: Descending	Count	Count percentages
NEGATIVE	4,111	89.37%
POSITIVE	489	10.63%

Export: Raw Formatted

As per the analysis done, there are 89.370% of negative tweets and 10.630% of positive tweets. The ratio of negative to positive tweets is around 9:1.

From the above observations, we can see that most of the tweets related to covid19 have a negative sentiment associated. Number of positive tweets stayed unchanged with only a few dips during the period of analysis. There exists huge dissatisfaction with the keyword covid19, and very few people have expressed positive thoughts about it.

2. Analyzing Social Networks using GraphFrame

DATASET:

Dataset Link: https://snap.stanford.edu/data/soc-Epinions1.html

Input Path Link: /FileStore/tables/soc_Epinions2.txt

Output Path Link: /FileStore/tables/res-graph.txt

Code Public Link:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3592547200979724/142679872762799/2120617183887627/latest.html

QUERY ANALYSIS:

a. Find the top 5 nodes with the highest outdegree and find the count of the number of outgoing edges in each

Result:

Top 5 high out degree nodes with their degrees

Node-645 with 1801 outgoing edges

Node-763 with 1669 outgoing edges

Node-634 with 1621 outgoing edges

Node-71399 with 1128 outgoing edges

Node-3924 with 976 outgoing edges

b. Find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each

Result:

Top 5 high in degree nodes with their degrees

Node-18 with 3035 incoming edges

Node-143 with 1521 incoming edges

Node-737 with 1317 incoming edges

Node-790 with 1284 incoming edges

Node-136 with 1180 incoming edges

c. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define any suitable parameters.

Result:

Top 5 page rank nodes with their values

Node-18 with 345.13733694777454 page rank

Node-737 with 240.25396712965414 page rank

Node-118 with 162.0208569978614 page rank

Node-1719 with 158.853456075026 page rank

Node-136 with 151.6712231987729 page rank

d. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

Result:

Top 5 strongly connected nodes

Node-0 with 32223 nodes

Node-137438953803 with 15 nodes

Node-60129542153 with 9 nodes

Node-360777253180 with 9 nodes

Node-317827580180 with 8 nodes

e. Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, you can randomly select the top 5 vertices.

Result:

Top 5 highest triangle count nodes

Node-645 with 48674 triangles

Node-18 with 47203 triangles

Node-27 with 25817 triangles

Node-634 with 25230 triangles

Node-44 with 24752 triangles

INSIGHTS:

- Node 18 has the max PageRank which is linked to a large number of nodes indicating the number of triangles that pass through it. There is a considerable difference in page rank between the Node eighteen and the second node 737.

- The page ranks of the top three to five pages differ slightly. This indicates that the dataset contains two frequently visited nodes.