**CS 6350.002 : Big Data Management & Analytics**

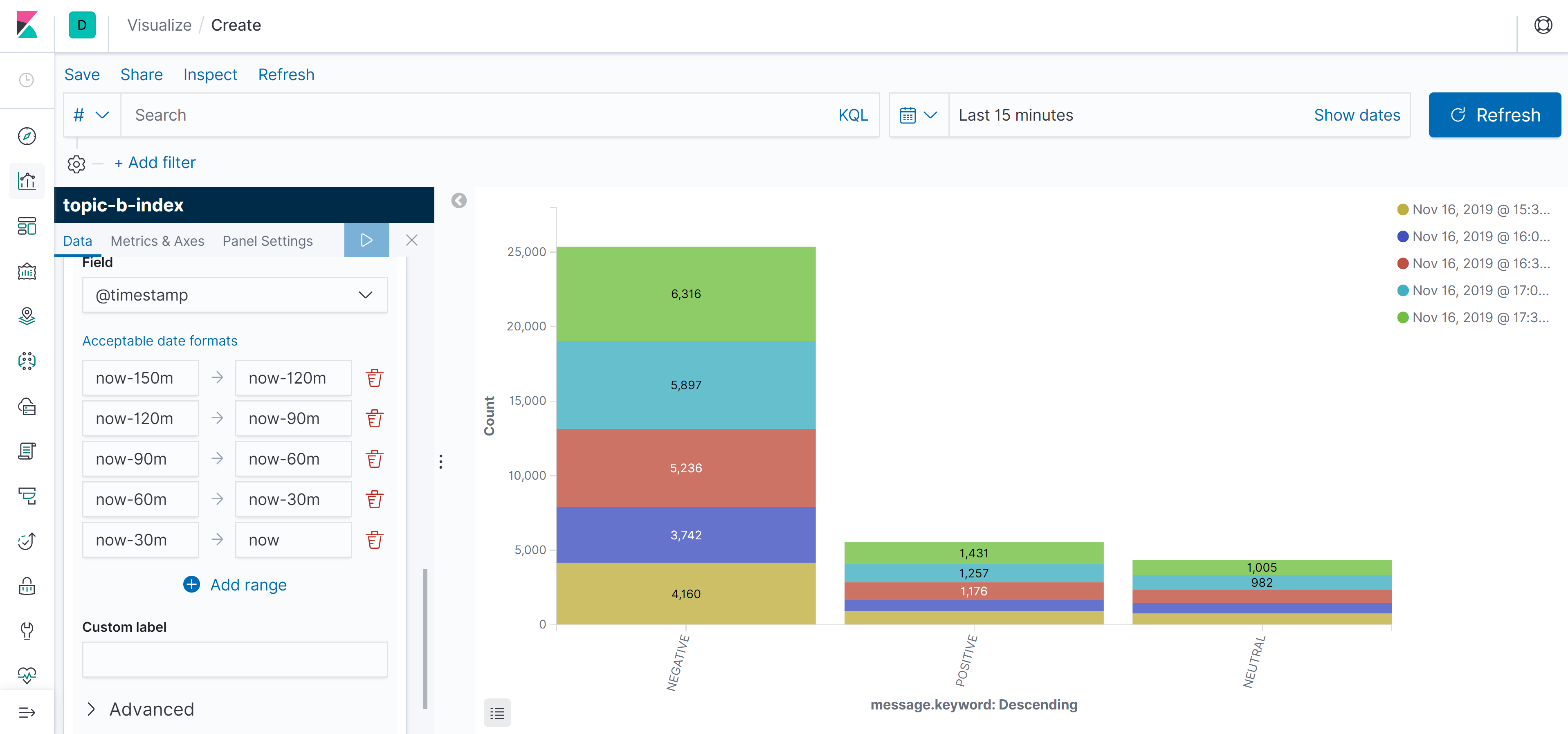
Assignment 3

Team members: Kavya Jampani(Netid : kxj170004)

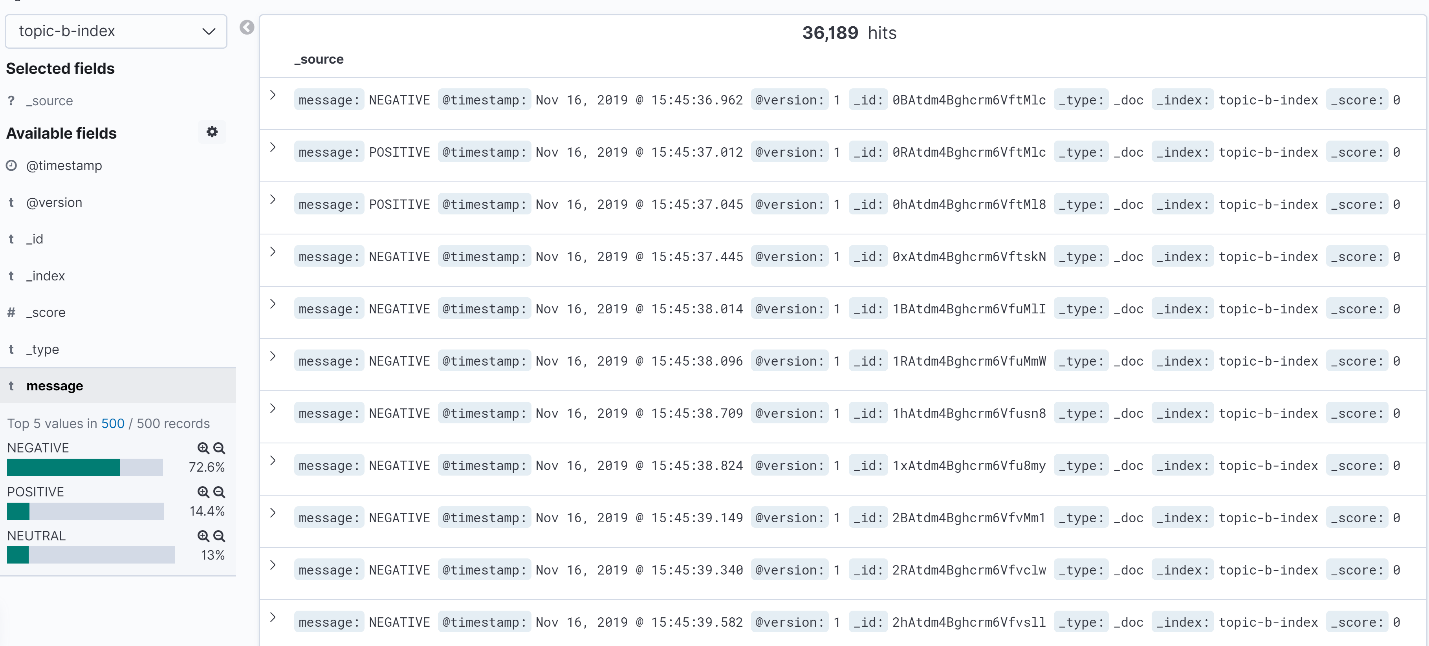
Hari Chandana Epuri (Netid: hxe17000)

**Q1. Spark Streaming with Twitter and Kafka**

The following is the visualization generated in kibana for question 1. For streaming of tweets, we have used the language “en” and “movie” filters. The tweets are monitored for a time span of 150 minutes (2 ½ hours) and the number of positive, negative and neutral sentiments are plotted over 30 min intervals. From the plot, we can observe that number of negative sentiments is more than the number of positive and neutral sentiments.

****

From the below screenshot from kibana, we can see that number of tweets collected = 36189 and out of which 72.6% are negative sentiments, 14.4% are positive and 13% are neutral sentiments.

****

**Question 2:**

**Dataset used:** <https://snap.stanford.edu/data/wiki-Vote.html>

Dataset Information:

Number of nodes: 7115

Number of edges: 103689

The dataset represents the administrator elections of Wikipedia vote history data. Nodes in the network represent Wikipedia users and a directed edge from node i to node j represents that user i voted on user j.

Insights from results:

Highest number of votes a single user casted is 893 whose id is 2565.

Highest number of votes a user received is 457 whose id is 4037.

Pagerank algorithm here is used to get the importance of the user i.e, who is most likely to get highest number of votes. The user with highest pagerank is 4037 with probability 32.761392590353616 We can see that the user with highest indegree has highest pagerank as well.

There is one strongly connected component with highest number of users 3504 and other components have same number of users.

The highest number of triangle counts is 30940 for user 2565

1. Top 5 nodes with highest outdegree

Id Count

2565 893

766 773

11 743

457 732

2688 618

b) Top 5 nodes with highest indegree

Id Count

4037 457

15 361

2398 340

2625 331

1297 309

1. Top 5 nodes with highest pagerank

Id Pagerank

4037 32.761392590353616

15 26.25300495762173

6634 26.16452443488874

2625 23.511515933028402

2398 18.728389390671293

1. Top 5 components with largest number of nodes

Id Nodes

0 3504

919123001344 3

42949672971 3

850403524621 3

489626271762 2

1. Top 5 vertices with largest triangle count

Count id

30940 2565

22003 1549

18204 766

17361 1166

14220 2688