Nihal Bhat
nsb170002
CS 6350.001

Assignment 3 Report

Section 1: Spark Streaming with Real Time Data and Kafka

For my streaming source, I used the PRAW Python library to stream comments from a subreddit. I decided to use the r/news subreddit since it has 27 million members and thousands of posts and active members, giving a high chance of comments on new posts. After I made a variable that gets the stream of comments from the News subreddit, I used the skip_existing=True option to ensure that all comments that the stream provided would be posted after my program had started. I then sent these comments to a Kafka topic, created another Python file to read the Kafka topic and write the named entities and their counts in JSON format to another Kafka topic. I then used the ELK stack to read from the second Kafka topic and visualize the counts using Elasticsearch and Kibana. The following graphs show the named entities and their counts at 15, 30, 45, and 60 minute intervals.

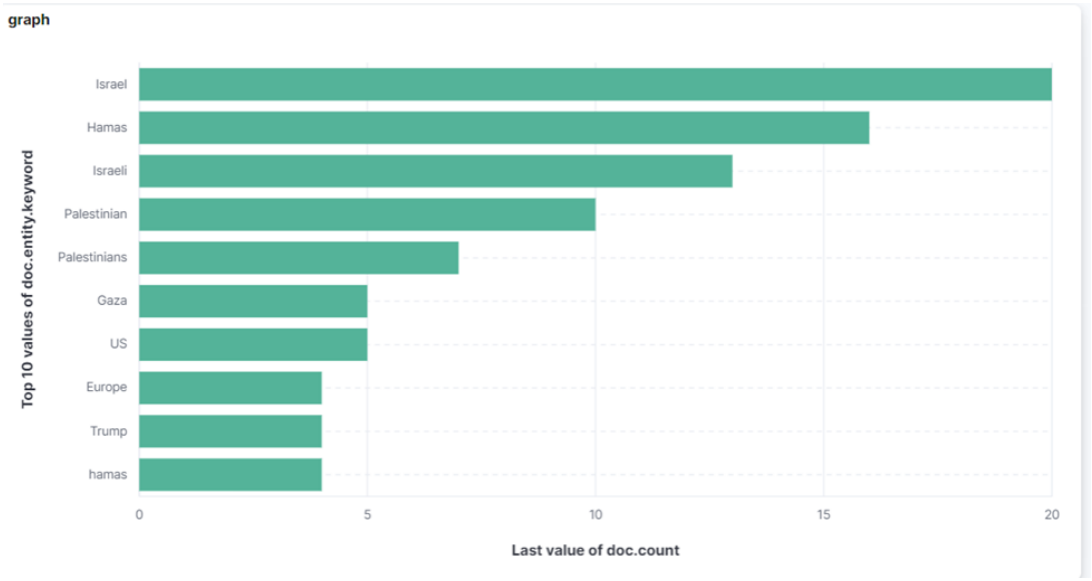

Figure 1: Named entities and counts after 15 minutes

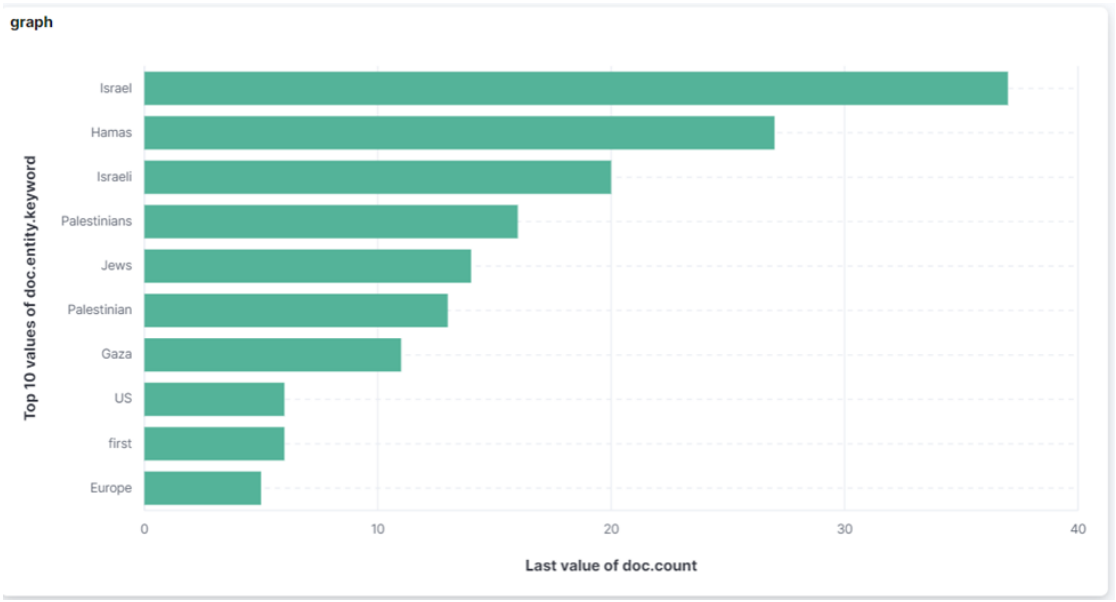

Figure 2: Named entities and counts after 30 minutes

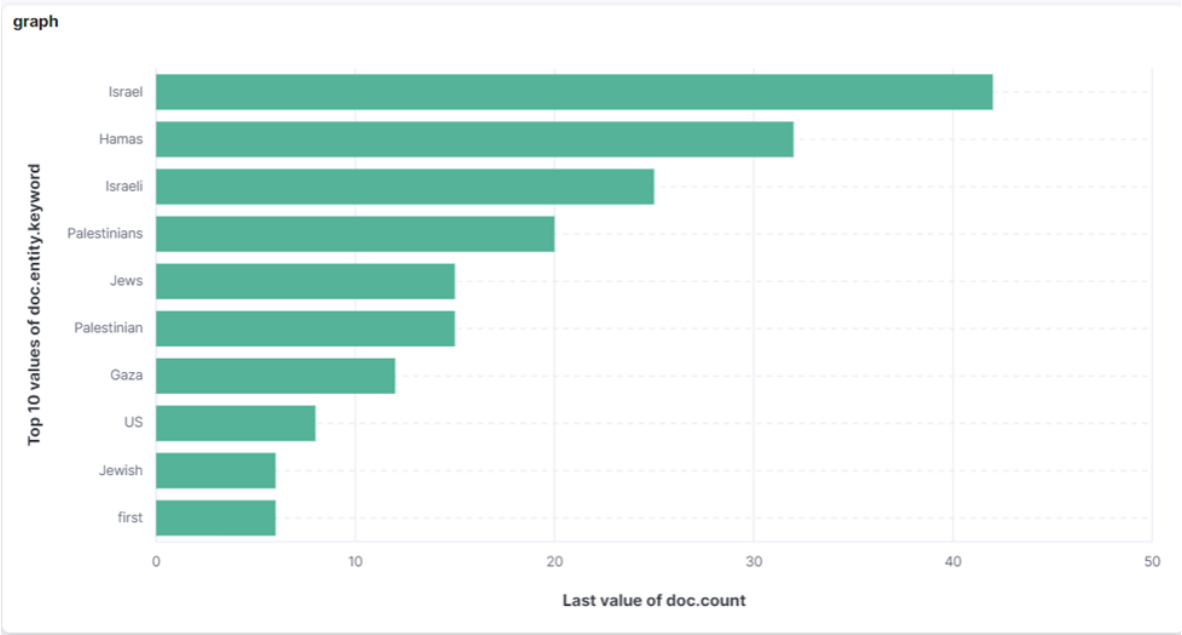

Figure 3: Named entities and counts after 45 minutes

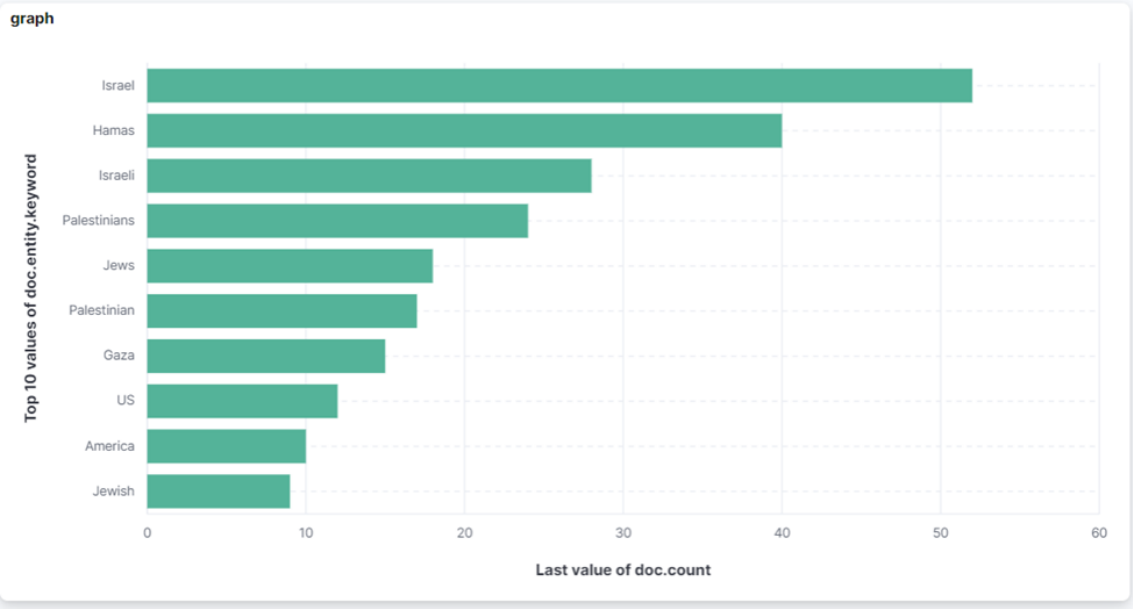

Figure 4: Named entities and counts after 60 minutes

The above graphs show the top 10 named entities and their counts over a period of 60 minutes, with intervals of 15 minutes. As can be seen from the graphs, the most occurring named entity was Israel, and followed by Hamas, Israelis, and Palestine. This is no surprise due to the Israel, Hamas, and Palestine conflict raging in the Middle East. We can see that Israel gets mentioned on average 13 times every 15 minutes, and Hamas gets mentioned on average 15 times every 15 minutes. Despite the number of active members and volume of posts and comments, the named entity counts are as high as one might think. Since Reddit is a social media platform, commenters typically use slang and colloquial language more than utilizing the full name of proper nouns and other named entities.

Section 2: Analyzing Social Networks using GraphX/GraphFrame

For this section on analyzing a social network dataset using GraphFrames, I used a dataset representing the Twitter network between the senators and representatives of the United States Congress. Here were the results of running the following operations on the graph:

1. The top 5 nodes with the highest outdegree and their counts

a.

id	name	outDegree
367	SpeakerPelosi	210
322	GOPLeader	157
393	RepBobbyRush	111
71	SenSchumer	97
399	SteveScalise	89

   b. The outDegrees explain how many people each person is following. SpeakerPelosi having the highest out-degree makes sense, because as the Speaker of the House it is her job to follow other members of her party and know what they are posting.
2. The top 5 nodes with the highest indegree and their counts
   a.

id	name	inDegree
322	GOPLeader	127
208	RepFranklin	121
190	RepJeffDuncan	120
111	RepDonBeyer	109
254	LeaderHoyer	108

   b. The inDegrees explain how many people are following you. The GOPLeader (Minority Leader of the House) having the highest in-degree makes sense because although he was the Minority Leader of the House, the Senate had a Republican majority at the time. The Republicans from the House and Senate would follow the GOPLeader, giving him the highest inDegree.
3. The top 5 nodes with the highest PageRank
   a.

id	name	pagerank
322	GOPLeader	4.418292220404877
208	RepFranklin	4.379837373041833
190	RepJeffDuncan	4.192224541660599
385	RepJohnRose	4.003239432976625
192	RepTomEmmer	3.983446282048152

b. The PageRank of a vertex explains the popularity of the vertex, utilizing features of the graph involving the vertex. Since the GOPLeader had the highest inDegree, it makes sense that he would also have the highest PageRank.

4. The top 5 connected components with the largest number of nodes
   a.

id	name	component
99	RepAuchincloss	0
0	SenatorBaldwin	0
236	RepRaulGrijalva	0
4	SenBlumenthal	0
300	RepTeresaLF	0

   b. In this dataset there are no connected components, because the graph might be disconnected meaning there could be members of Congress who do not have any followers and are not following other people.
5. The top 5 vertices with the largest triangle count
   a.

id	name	count
367	SpeakerPelosi	3281
322	GOPLeader	2777
190	RepJeffDuncan	1900
208	RepFranklin	1894
254	LeaderHoyer	1893

   b. The triangleCount explains how many triangles each vertex is involved, either through outDegrees or inDegrees. Since SpeakerPelosi and the GOPLeader have the highest outDegrees and inDegrees respectively, it makes sense that they have the highest triangleCounts.

Overall, the results of the algorithms ran makes sense and was useful in gaining insights into how the congress Twitter network is structured. It helped to see the popularity of the top members of Congress and how they affect the entire network.