CS 6350 – ASSIGNMENT 3 – REPORT

Group Members

- Krittika Paul (KXP230001)
- Harsha Priya Daggubati (HXD220007)

Q1. Spark Streaming with Real Time Data and Kafka

Data Source:

- Downloaded real time news using the News API available at: https://newsapi.org/
- Generated a new API KEY: "b2e62f4d3bb24b2fb5bf6d4cceecf71b".
- Searching through all the articles that have been published by more than 150,000 news sites and blogs over the past five years is the primary usage of News API. It can be considered as a programmed equivalent of Google News.

Results and Bar Plot:

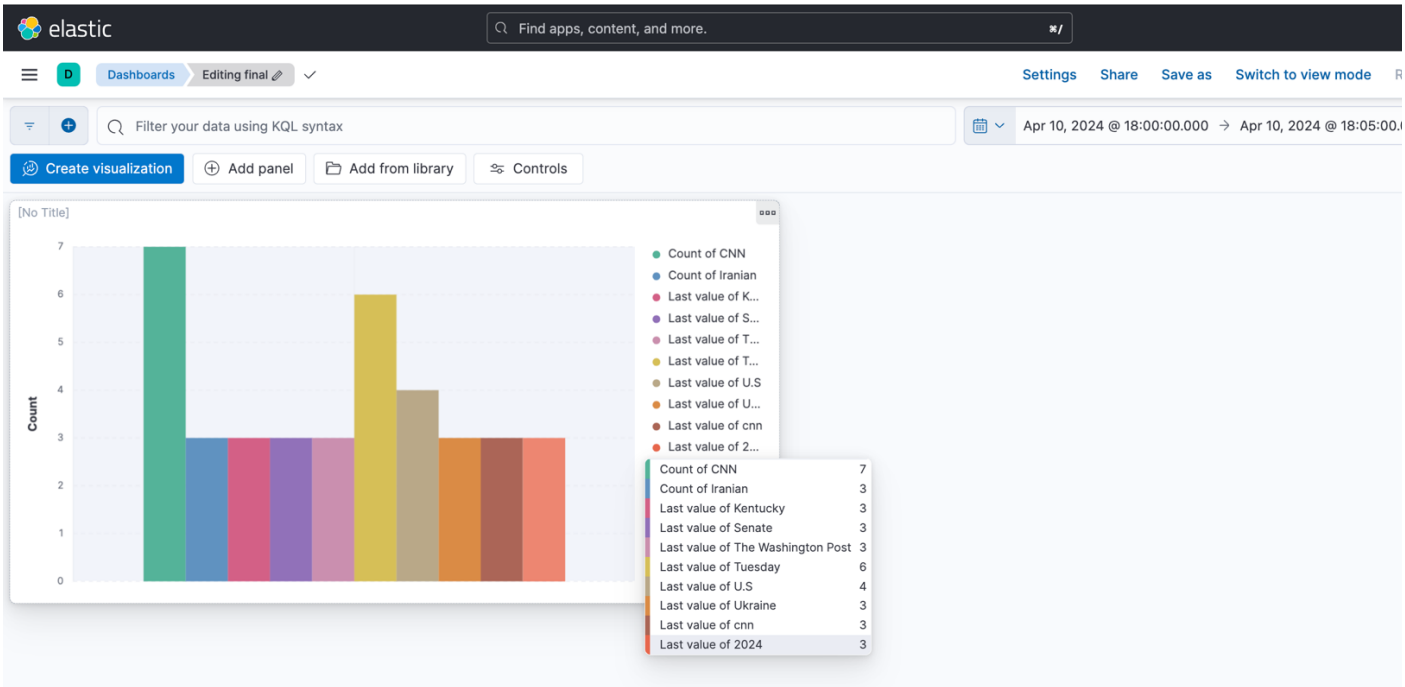


Fig 1: A screenshot of wordcount bar plot taken after **5 minutes**, shows the count of top 10 named entities appearing in the news article.

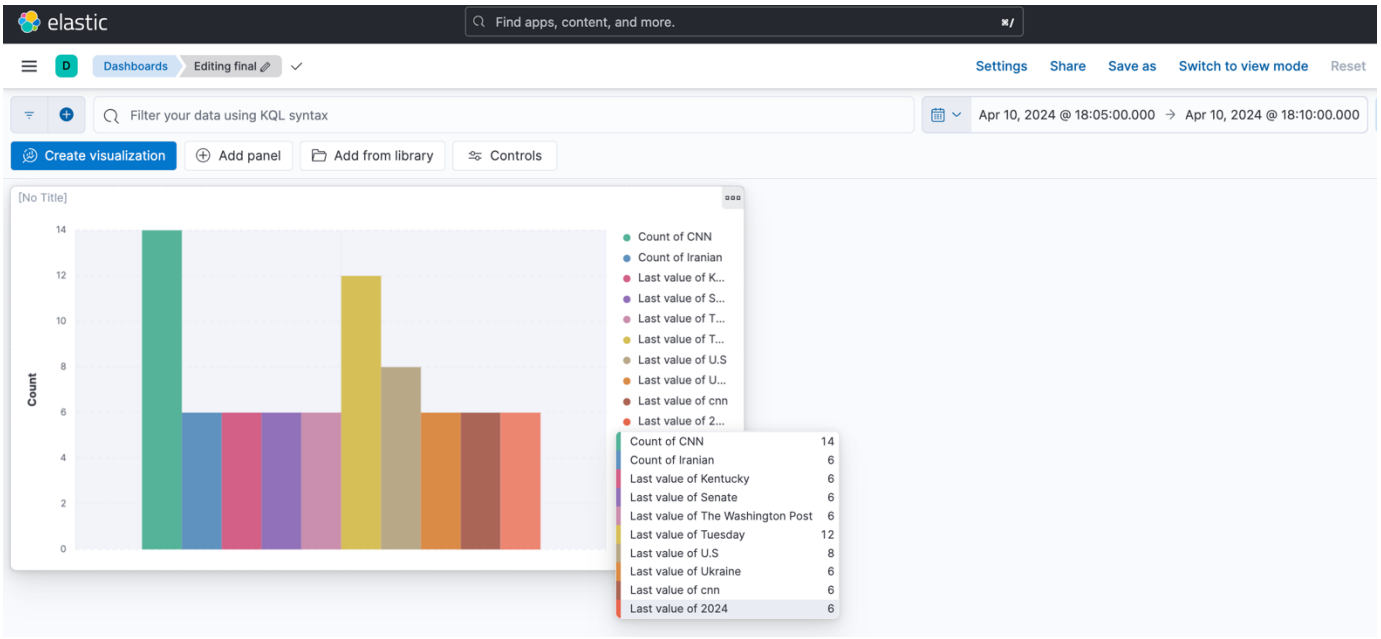


Fig 2: A screenshot of wordcount bar plot taken after **10 minutes**, shows the count of top 10 named entities appearing in the news article.

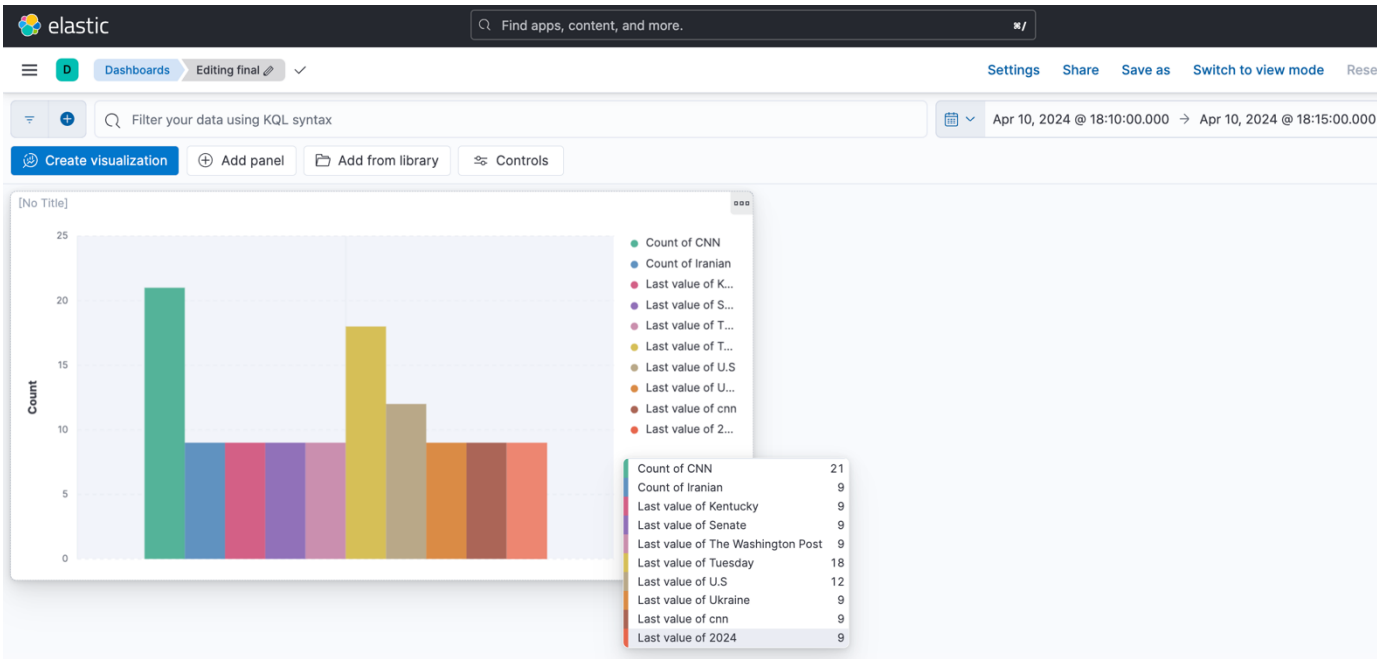


Fig 3: A screenshot of wordcount bar plot taken after **15 minutes**, shows the count of top 10 named entities appearing in the news article.

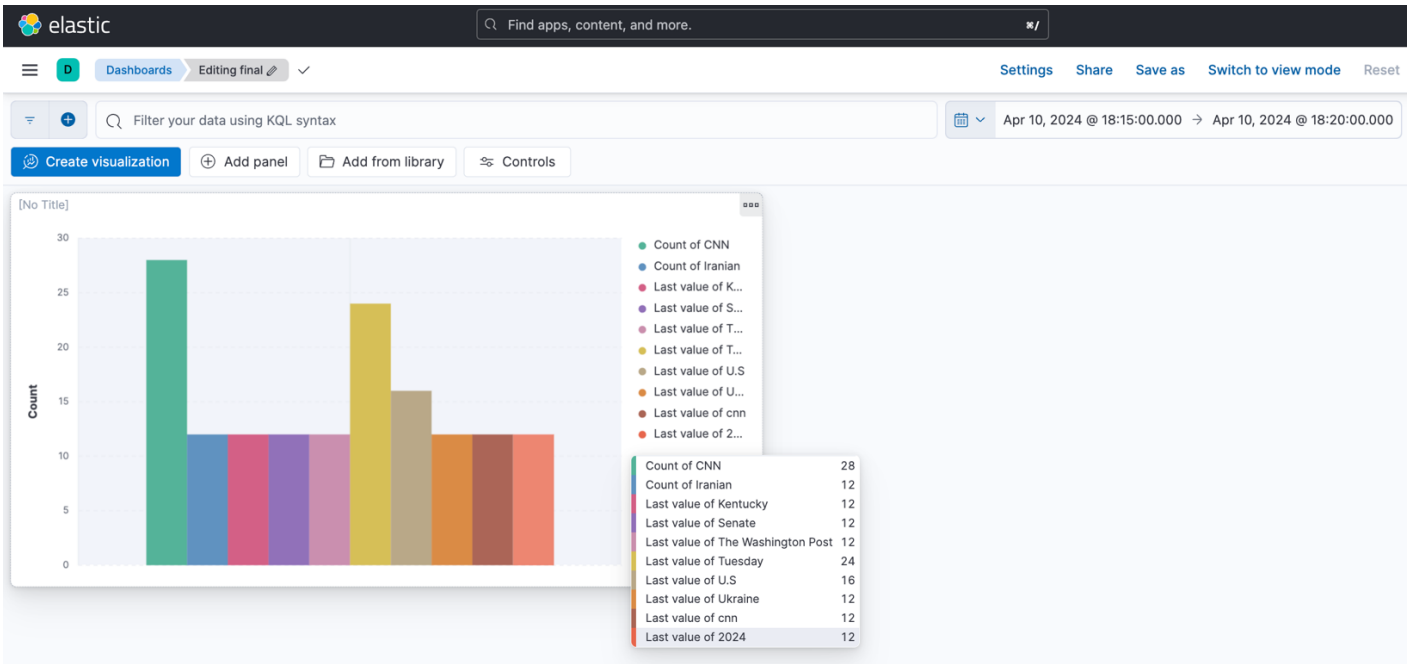


Fig 4: A screenshot of wordcount bar plot taken after **20 minutes**, shows the count of top 10 named entities appearing in the news article.

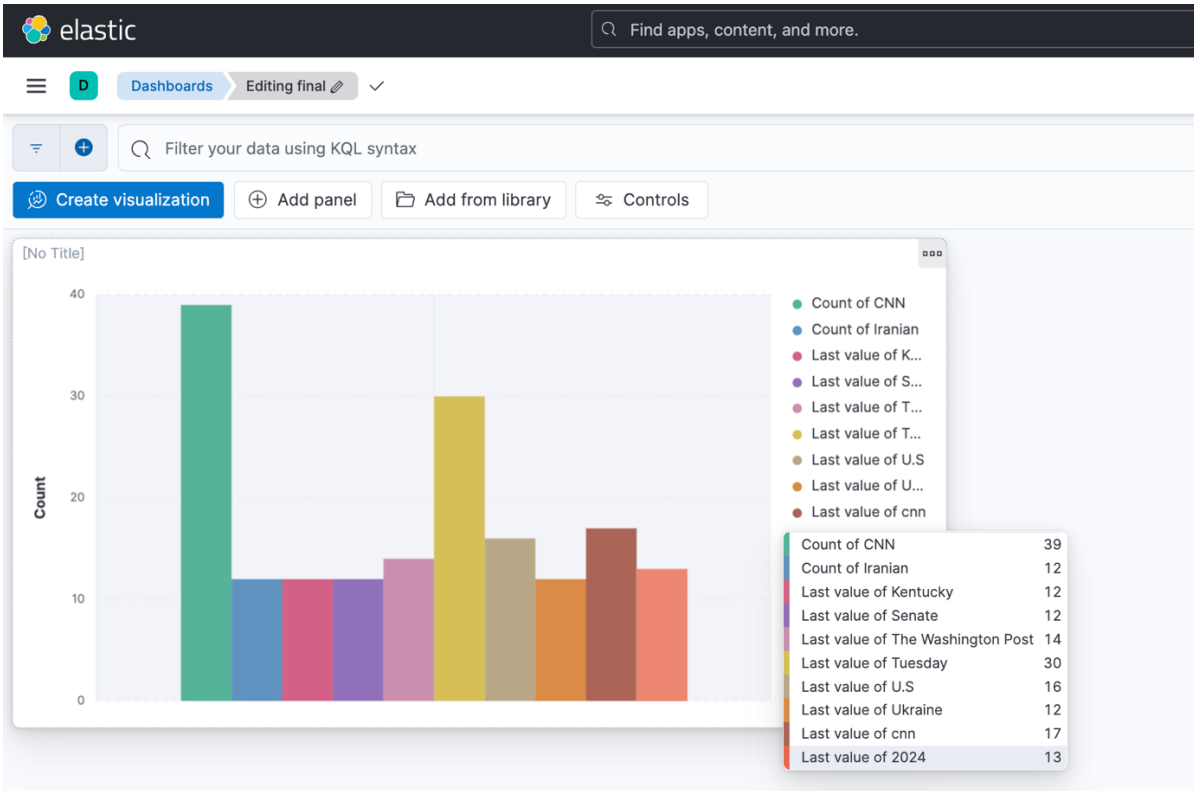


Fig 5: A screenshot of wordcount bar plot taken after **25 minutes**, shows the count of top 10 named entities appearing in the news article.

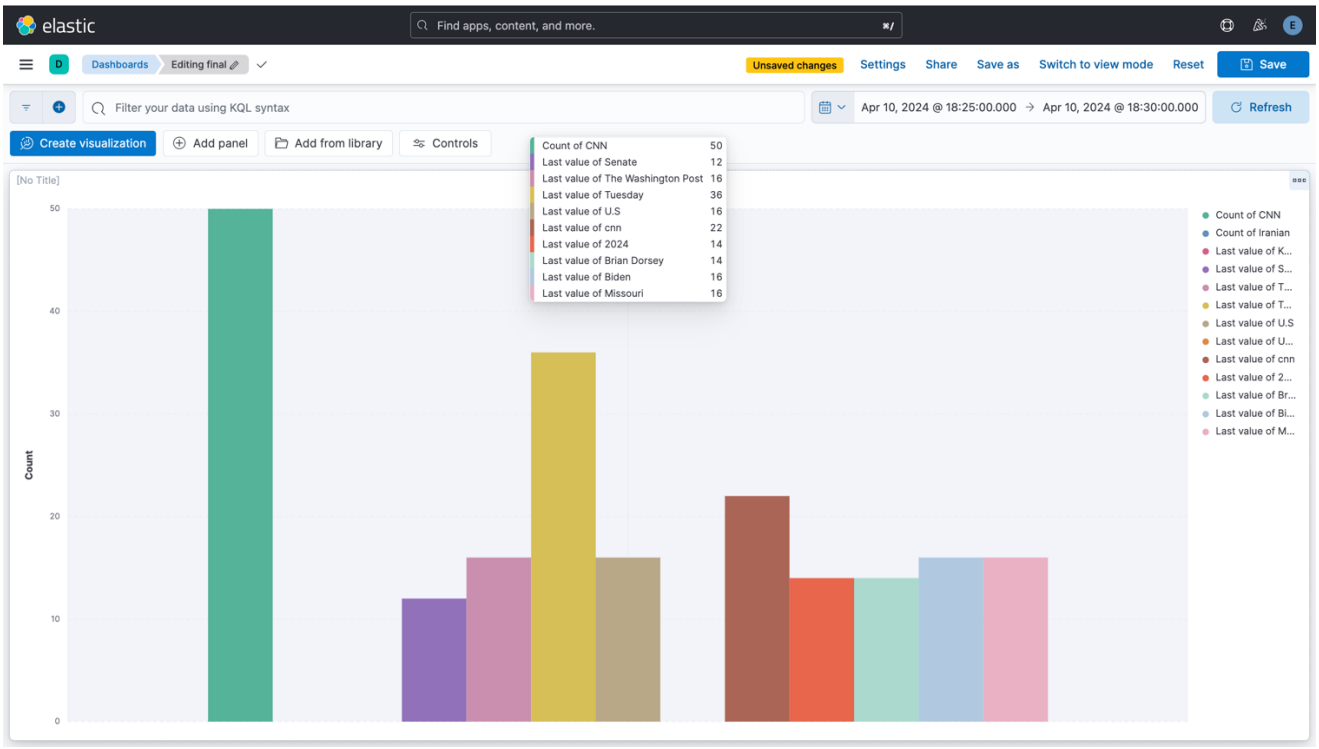


Fig 6: A screenshot of wordcount bar plot taken after **30 minutes**, shows the count of top 10 named entities appearing in the news article.

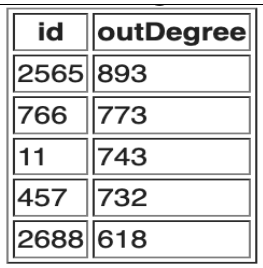
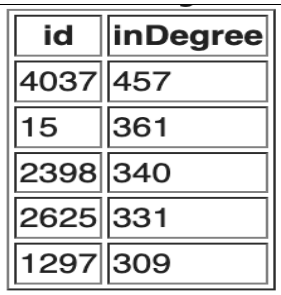
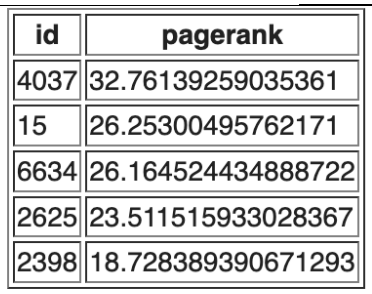
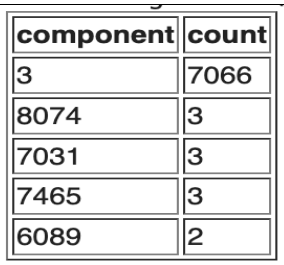
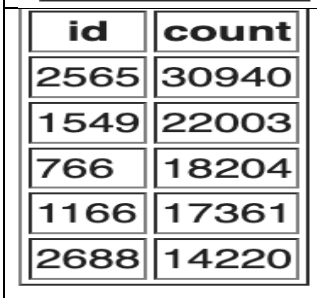
- In the figures, the vertical axis shows the last value/count of top 10 named entities at the interval of 5 minutes (300 secs).
- With every passing minute the count of every named entity increases as shown.
- The top 10 named entity list might not contain same list of words all along. As we can see in the results, till Figure 5 (25th minute), the top 10 named entities were CNN, Iranian, Kentucky, Senate, The Washington Post, Tuesday, U.S, Ukraine, cnn, 2024. But after 30 minutes, Iranian, Kentucky and Ukraine went out of the list and new words Brian Dorsey, Biden and Missouri got added to the list.

README (STEPS TO RUN): Follow the Q1.readme file.

Q2. Analyzing Social Networks using GraphX/GraphFrame

Google Collab Notebook link to Q2:
https://colab.research.google.com/drive/1ehCbUqoqalXQ9mbbhNFkXA5ET82d0POA?usp=sharing

RESULTS

 id | outDegree 2565 | 893 766 | 773 11 | 743 457 | 732 2688 | 618	*a. Find the top 5 nodes with the highest outdegree and find the count of the number of outgoing edges in each.* The output shows a list of top 5 nodes with their outdegrees arranged in descending order. Node with id 2565 has maximum outdegree of 893, that means it has 893 outgoing edges.
 id | inDegree 4037 | 457 15 | 361 2398 | 340 2625 | 331 1297 | 309	*b. Find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each.* The output shows a list of top 5 nodes with their indegrees arranged in descending order. Node with id 4037 has maximum indegree of 457, that means it has 457 incoming edges.
 id | pagerank 4037 | 32.76139259035361 15 | 26.25300495762171 6634 | 26.164524434888722 2625 | 23.511515933028367 2398 | 18.728389390671293	*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.* The output shows a list of top 5 nodes with their page ranks arranged in descending order. Node with id 4037 has maximum PageRank of 32.76, that means it has highest incoming edges and has high importance.
 component | count 3 | 7066 8074 | 3 7031 | 3 7465 | 3 6089 | 2	*d. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.* The output shows the top 5 components and their corresponding node count. Component 3 has the highest node count of 7066 followed by the rest of the components.
 id | count 2565 | 30940 1549 | 22003 766 | 18204 1166 | 17361 2688 | 14220	*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.* The output shows a list of top 5 nodes with their corresponding triangle counts arranged in descending order. Node with id 2565 has maximum count of 30940, that means it has maximum connected components when triangle algorithm is used.

README: Follow Q2.readme file for running the code.