CS 6350
ASSIGNMENT 3

Names of students in your group:

Tarun Teja Obbina (txo220011)

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.

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

https://snap.stanford.edu/data/wiki-Vote.html
https://spark.apache.org/docs/latest/structured-streaming-kafka-integration.html
https://subhamkharwal.medium.com/pyspark-structured-streaming-read-from-kafka-64c40767155f
https://logz.io/learn/complete-guide-elk-stack/

Report

Part 1:

Data Source:

I've implemented a robust data pipeline using the News API's "everything in US '' endpoint, allowing me to retrieve the latest headlines every 5 minutes. These headlines are then stored in a designated kafka topic, topic1. Subsequently, I utilize a structured read stream to access the content of all news articles within topic1. Employing natural language processing techniques, I extract all named entities from the articles and counted their occurrences. These named entities and their respective counts are structured into a JSON object, where the entities serve as keys and the counts as values. This structured data is stored in topic2. Next, I leverage Logstash to ingest the named entities from topic2 and seamlessly index them into Elasticsearch, under the named_entities index. Finally, I utilize Kibana to visualize the top 10 named entities with the highest word counts, presenting them in an insightful vertical bar format for easy interpretation and analysis.

Results:

The analysis of news data from the News API using the ELK stack, with visualization in Kibana, revealed intriguing patterns over various time intervals. Notably, there was a consistent prominence of the words "Wednesday" and "Thursday" across all observed intervals, suggesting a heightened frequency of news events or topics related to mid-week activities or events occurring on these days. Additionally, a notable shift was observed in the position and frequency of the word "Israel," which surged from its previous placement at the tail end to the sixth position. This upward trajectory in occurrence indicates a potential increase in news coverage or attention towards Israel-related events or developments. Conversely, the word "third" exhibited a contrasting trend, diminishing from its seventh position to the lower ranks, implying a decrease in its relative importance or occurrence within the news corpus during the observed intervals. These fluctuations highlight the dynamic nature of news content and the insights gleaned from analyzing named entities, offering valuable context for understanding evolving trends and topics within the news landscape.

Part 2

About Dataset:

The Wikivote dataset, available in the SNAP (Stanford Network Analysis Project) library, provides a valuable resource for studying social networks and online communities within the context of Wikipedia. This dataset comprises directed edges representing votes between Wikipedia users during the 2008 US presidential election. Specifically, users are nodes in the network, and directed edges denote votes from one user to another. The dataset offers insights into user interactions, preferences, and influence dynamics within the Wikipedia community during a politically significant period. Analyzing the Wikivote dataset enables researchers to explore patterns of influence, identify key contributors, and understand the underlying structure of collaborative decision-making processes within online communities.

Summary:

Based on the provided results from the Wikivote dataset analyzed using GraphFrames, several insights can be gleaned:

1. Out-Degree Analysis: The top nodes with the highest out-degrees are identified, indicating users who have made the most votes in the network. For instance, node 2565 has the highest out-degree of 893, suggesting significant engagement or activity in the voting process.

2. In-Degree Analysis: Conversely, the top nodes with the highest in-degrees represent users who have received the most votes from others in the network. Node 4037, with an in-degree of 457, stands out as a prominent recipient of votes, indicating perceived influence or relevance within the community.

3. PageRank Analysis: Notably, node 4037 has the highest PageRank score of 32.84, reaffirming its significance within the network.

4. Connected Components: The network is partitioned into connected components, with each component representing a subset of nodes where each node is reachable from any other node within the same component. The largest component, labeled as component 0, contains 7066 nodes, indicating a cohesive and interconnected portion of the network.

5. Node Counts: Finally, the counts of nodes are provided, revealing the distribution of nodes based on their frequency in the network. Node 2565 appears most frequently, with a count of 30940, followed by nodes 1549, 766, 1166, and 2688.

Notebook link
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3664271699826134/3415248045185465/373535317718030/latest.html

Dataset link
https://drive.google.com/file/d/1Dx_wKGvcdlKvfEaM2Z_6C0gJkTn181sK/view?usp=drive_link