Data Source Explanation

For this assignment, I used NewsAPI, a real-time web service that provides JSON-formatted news articles from a wide variety of sources and publishers. I queried the API with the keyword "trump" and set the from and to parameters to retrieve articles published between yesterday and today. Additional filters included language set to English, results sorted by publication date (sortBy: publishedAt), and a limit of 5 articles (pageSize: 5). The API returned structured data including metadata such as the title, description, publication time, and source of each article.

This data was then processed in a streaming pipeline where it was ingested through Kafka, analyzed with PySpark to extract named entities and word frequencies, and finally visualized using the ELK stack.

Result Summary

The Named Entity Recognition (NER) analysis conducted on news articles about "Trump" revealed consistent patterns over multiple intervals. The top entity types detected include GPE (Geopolitical Entities), ORGANIZATION, and PERSON, each appearing frequently across batches. This indicates that news coverage involving Trump heavily references geopolitical regions (like countries or cities), people (e.g., Trump himself or other political figures), and organizations (such as government bodies or political parties).

Entities labeled as FACILITY and GSP were much less frequent, suggesting that places and some specialized named categories were not as relevant in the context of this news cycle. The consistent ranking across time intervals suggests stable news topics within that period, offering insights into trending themes and the prominence of various named entities in real-time news streams.