# Spark Streaming Report

## Data Source

The data source used in this project is NewsAPI. In order to get trending news, the /v2/top-headlines endpoint was called. Unsure of how often this endpoint is updated with new articles, the script calls this endpoint every hour to retrieve more news. The script locally stores all articles it has already sent to the Kafka topic, so no duplicate news data is sent.

## Bar Plot Results

4 bar plots were created for 4 different batches of data received at 4 different intervals, over the course of 5 hours. Each plot shows the top 10 named entities by count for that batch of data. As expected, the counts for each batch increase over time, as more named entities are gathered from articles. The most commonly named entities are days in a week, which makes sense, as articles usually have dates and events associated with them. If an event is currently trending, there will be more count of named entities associated with that event, for example, "Black Friday" and "Cyber Monday", which will make it on the top 10 bar plots.

Even though 83 articles were parsed for this data, it might be weird as to why the counts are lower. This is due to the fact there are a lot of special and escape characters in the content of the API request, such as **\n** or **\r**, which are next to actual named entities, but without any additional preprocessing, they will be counted as a unique named entity, which is not entirely the case.