


# Assignment 3 Report

Data Source: <https://newsapi.org/>

1. All articles about Tesla from the last month
  - a. API:  
<https://newsapi.org/v2/everything?q=tesla&from=2024-03-14&sortBy=publishedAt&apiKey=2cb949194e494e57a58e34929fb89df0>
  - b. In the query parameters, date in “from” changes everyday. Please make sure to change the value in the code and apiKey is your personal API KEY.
2. About the code:
  - a. newsTopicOne takes the response from news API, performs named entity extraction by removing stopwords and alphanumeric words from description of the news articles and sends them to topic 1.
  - b. writeTopicTwo reads from kafka topic 1 and performs the count on named entities and writes to topic 2.
  - c. We configure logstash to take input from kafka topic 2 and output to elasticsearch and make use of kibana to visualize the count of named entities at 15, 30, 45 and 60 mins.
3. About the result:
  - a. We get the top 10 most frequent words to be: Tesla, Musk, Elon, price, month, subscription, full, software, mile, electric
  - b. At every 15 mins of intervals, the word “Tesla” is most frequent as it goes from 300 -> 1200 -> 2000 -> 3000 which takes up almost 27% of the total count in every interval.
  - c. “Musk”, “Elon” are the next most frequent words.
  - d. “Software”, “mile”, “electric” are the least frequent words in the top 10 named entities and they keep switching positions at the 45 min and 60 min mark.
4. References:

- a. <https://spark.apache.org/docs/2.3.1/structured-streaming-kafka-integration.html#deploying>
- b. <https://kafka-python.readthedocs.io/en/master/apidoc/KafkaProducer.html>
- c. <https://stackoverflow.com/questions/62149261/how-to-produce-kafka-messages-with-json-format-in-python>
- d. [https://medium.com/@mukeshkumar\\_46704/consume-json-messages-from-kafka-using-kafka-pythons-deserializer-859f5d39e02c](https://medium.com/@mukeshkumar_46704/consume-json-messages-from-kafka-using-kafka-pythons-deserializer-859f5d39e02c)
- e. <https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.Session.builder.config.html#pyspark.sql.Session.builder.config>
- f.  How to Use Logstash to import CSV Files Into ElasticSearch
- g. <https://stackoverflow.com/questions/64922560/pyspark-and-kafka-set-are-gone-some-data-may-have-been-missed>
- h. <https://stackoverflow.com/questions/58035890/index-pattern-is-not-showing-up-in-kibana-management>
- i. <https://discuss.elastic.co/t/elasticsearch-unreachable-http-localhost-9200-manticore-clientprotocolexception-localhost-9200-failed-to-respond/325897/3>
- j.