##### **Team Members:**

1. Karan Patel - kxp173930
2. Pandit Joshi - pxj171830
3. Prashanth - sxp178231

# Large Scale Data Collection and preprocessing in Spark

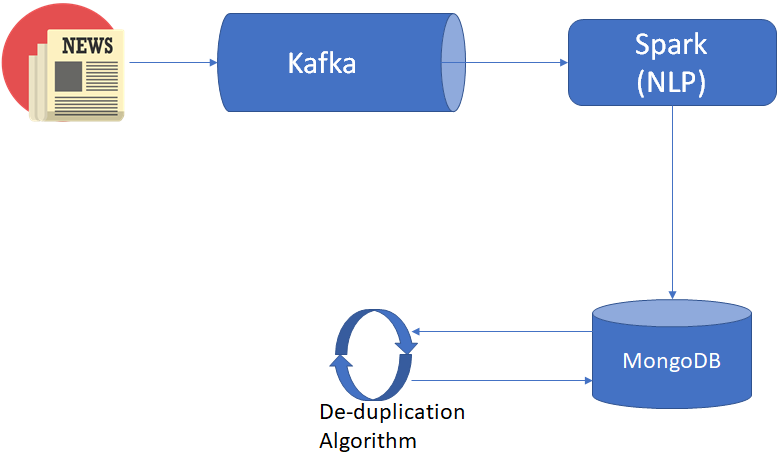
## Description

In this project, we are planning to develop a novel distributed framework using

Apache Spark, MongoDB, Stanford CoreNLP, PETRARCH, and Apache Kafka to gather the Spanish news articles from the URLs provided as input and then run Spark jobs to figure out the duplicate articles.

We are planning to use the PETRARCH as part of the content level deduplication algorithm using the encoded events from the metadata.

## Tentative Method



## Tentative Schedule

1. Setup the Kafka and web crawler to extract the data from the URLs and store it in the XML file in HDFS - approx 1 week
2. Run the Spark Job to extract the metadata from the XML and parse it using the UD parser and then store it in the MongoDB. - approx 2 week
3. Run the Spark Job using the PETRARCH software to find out the who-did-what-to-whom. - approx 1 week
4. Trying out several Implementation of good content level deduplication algorithm to find the duplicate news-article. - approx 1 week.

## Programming environment

1. Python
2. Apache Spark
3. MongoDB
4. Stanford CoreNLP
5. PETRARCH
6. Scrapy
7. newsplease
8. Apache Kafka
9. Ubuntu

## References:

1. <https://github.com/fhamborg/news-please>
2. <https://scrapy.org/>
3. <https://kafka.apache.org/>
4. <https://ieeexplore.ieee.org/document/7474330>
5. <https://universaldependencies.org/>