News Classifier Project Report

Objective:

To train machine learning model and classify upcoming news on the fly with good accuracy by building end to end machine learning pipeline.

To make containerized application which is scalable, robust, fault tolerant.

Planning:

We are using agile methodology to build the project. Task level details are mentioned in below gantt chart.

#Sprint: 4

#People: 2

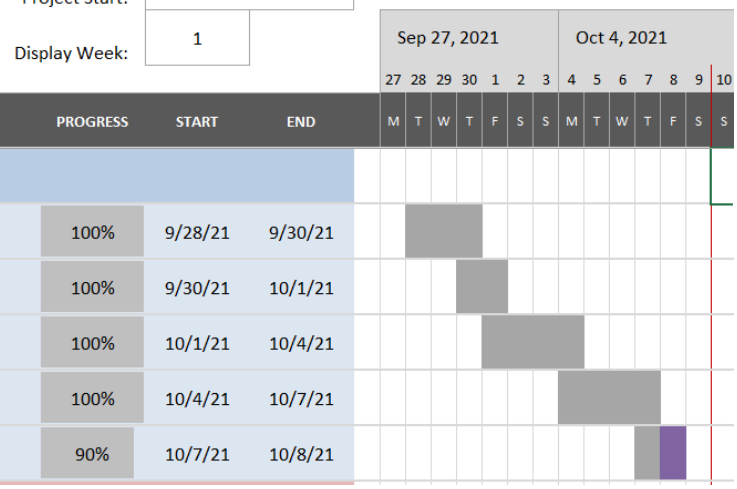
Project members:

Biren Patel, Vaishali khurana

Project Start: Tue, 9/28/2021

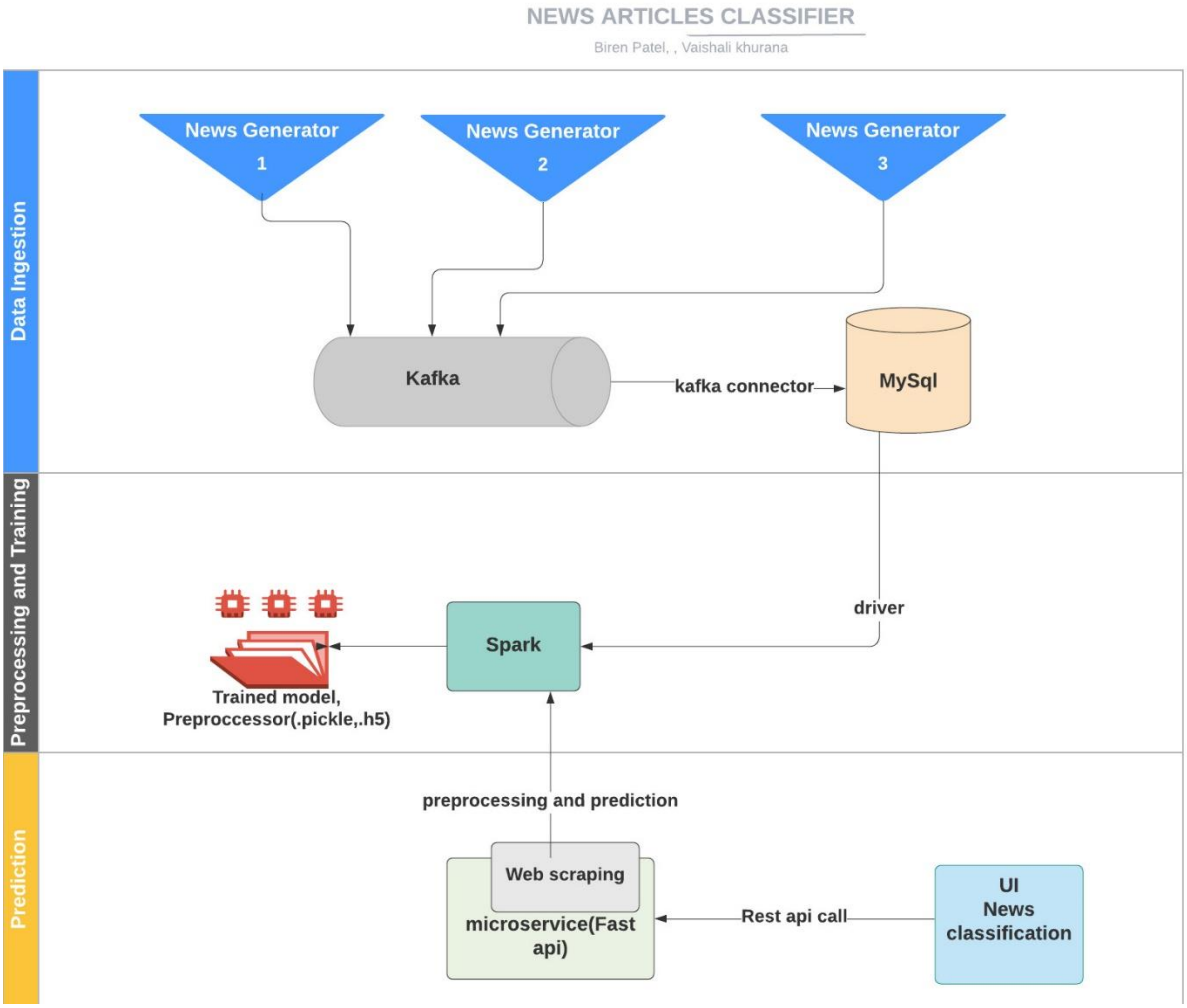
Display Week: 1

TASK	PROGRESS	START	END
Milestone 1: Data ingestion			
Project planning	100%	9/28/21	9/30/21
Building the architecture	100%	9/30/21	10/1/21
Building data ingestion pipeline	100%	10/1/21	10/4/21
Integration	100%	10/4/21	10/7/21
Report writing	90%	10/7/21	10/8/21



News Classifier Project Report

Architecture:



Note: We have used mysql considering the future scope of this application, Please read the future scope at the end of the document

Components and description:

Component	description
Data Ingestion	Multithreaded service that is collecting the data from web using rapid api and custom news generator, passing it to kafka queue. Kafka jdbc connector is used to sync the data between kafka and database. Finally, data is dumped into Mysql database
Preprocessor and trainer	Distributed service to preprocess the data stored in database and train ML model using spark

Prediction	Scraping the weblink provided by user in UI, clean it and predict the news using rest api.

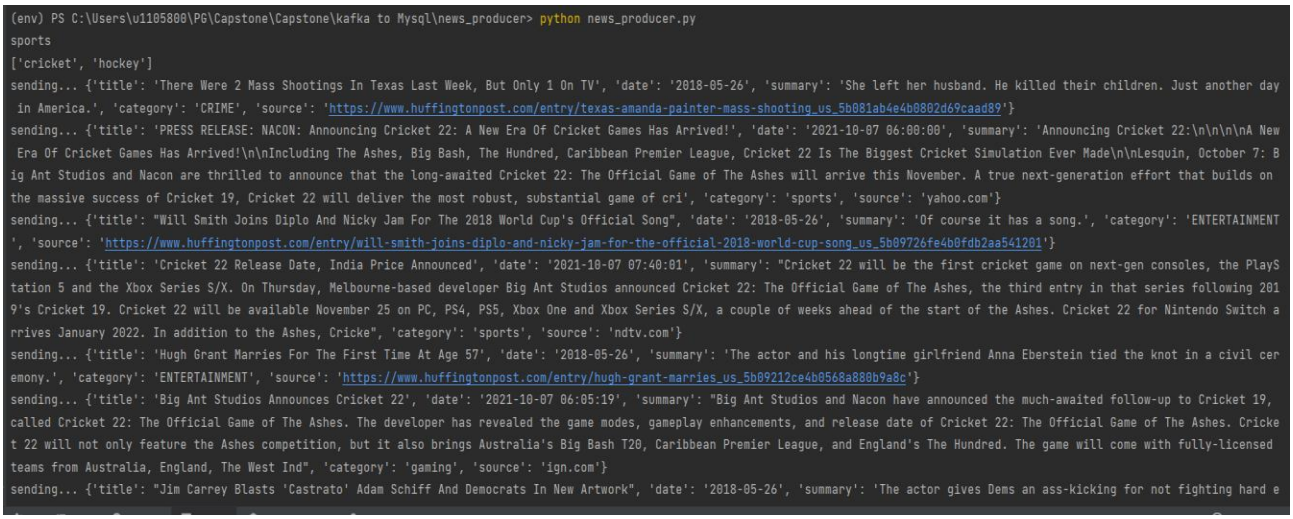
Since data ingestion is the first milestone we have explained it in detail here**.**

Data Ingestion:

- **Environment details**:

  Docker environment with kafka broker running on 9092 port, zookeeper running on 2181, kafka connect running on 8083, mysql running 3306

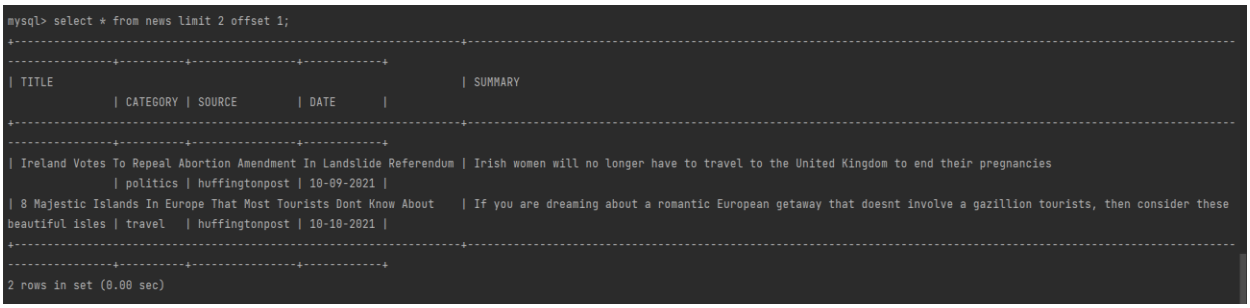
- **What goes in as an input:**
  We have used rapid api and custom news generator as data sources.



News Classifier Project Report

- **How the input is being processed:**
  Multithreaded application that is collecting the data from web using rapid api and custom news generator, passing it to kafka queue. Kafka jdbc connector is used to sync the data between kafka and database. Finally, data is dumped into Mysql database.

- **What comes out as an output:** Data stored in mysql database



- **Tools/libraries used**: Docker, Pycharm, kafka, Mysql, zookeeper, Spark, kafka

Challenges encountered:

We faced below mentioned challenges. However, we have resolved them.

- Finding better legal data sources
- Api rate limiting makes the pipeline slow
- Character encoding
- Data Labeling
- Connecting kafka to data store

- Network configurations for kafka connectors

Future Scope:

- We will be adding more features to build end to end news browsing application like
    - bookmarking the news
    - subscribing to specific news
    - news-recommendations
    - notifications etc.
- Further scale optimizations
- Implementing re-training mechanism using feedback feature
- Once we have large volume of labeled data, we will train our own model in place of transfer learning

Github link:

https://github.com/biren162/Capstone