Architecture and Design - News Article Category Prediction (Week 2 - Milestones)

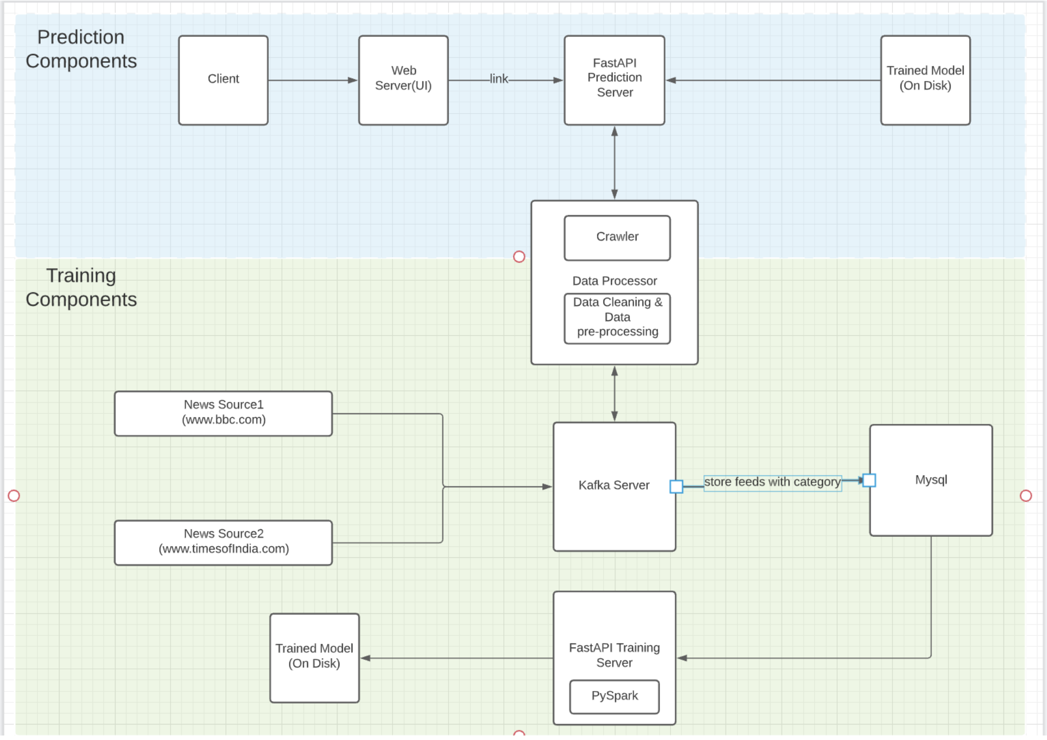
Objective

This document is about classifying News Articles into categories - With information overload today users are inundated with news articles of all topics, even the ones which may not be relevant to users. Design a system which can classify incoming news articles and appropriately tag the corresponding category.

Main Software Components

- html, css, js
- django
- python FastAPI
- Crawler(Beautiful Soup)
- pyspark
- Kafka
- mysql
- NLP libraries

Architecture



Milestones

Understand cleaning + preprocessing steps necessary to transform the raw data and complete the data preparation step

Components

- Crawler
- Data Cleaner
- Data Pre-Processor

Working

This component is used by training as well as predicting framework.

Capture news URLs for every category using www.google.com

- Crawler takes the source(www.bbc.com) as the input through API call, searches the google news for every category and get the top news urls.
  - It stores the links for every category in mysql DB.

Extracting article from the URLs

  - It takes the news url as an input, crawls the web page, extracts the article using Beautiful Soup library.

Article cleaning and pre-processing

  - Then article is cleaned, processed and returned to caller.
  - Cleaning and processing includes:
    - Normalizing Text
    - Removing Unicode Characters
    - Removing Stopwords
    - Stemming and Lemmatization

Setup the model-training-service project

Services

- Kafka Server
- PySpark
- mysql
- Data Processor

Working

Saving Articles

- `Kafka Producer` fetches the news URLs and category from mysql DB. Passes the URLs one by one to `Crawler` to fetch the corresponding cleaned article.
- This article and its category are sent to the kafka topic `news_feed`.
- `Kafka Consumer` consumes the article, category from the topic `news_feed` and stores them into the mysql DB.

Loading articles from mysql DB

- Articles are loaded into pysparkRDD from mysql.

```
select source,url,category,article from article_details;
```

Training Service

- We have defined the following categories:

  ```
  {'business':1,'computers':2,'covid':3,'entertainment':4,'health':5,'lifestyle':6,
  ```

- Extracted the features(article, category) from the data.
- Then splitted the data into training and testing dataset.
- Training and testing dataset is transformed into count vectors using `CountVectorizer`.
- `MultinomialNB` is used as a classification training algorithm.

Group Details

- Group-Name: VK Learners
- - Member1: Vinod Adwani
- - Member2: KSRC Murthy