Architecture and Design - News Article Category Prediction

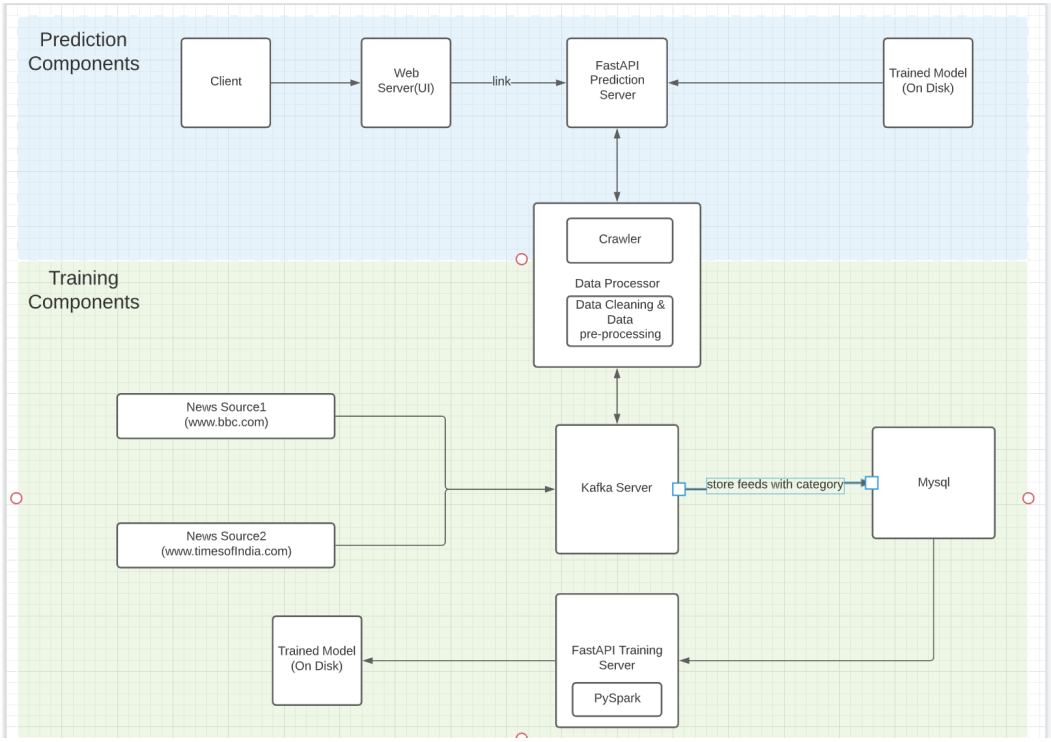
Objective

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



Services

Data Processor

Components

- Crawler
- Data Cleaner
- Data Pre-Processor

Working

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

- It takes the news url as an input, crawls the web page, extracts the article using Beautiful Soup library.
- Then article is cleaned, processed and returned to caller.
- This includes:
  - Normalizing Text
  - Removing Unicode Characters
  - Removing Stopwords
  - Stemming and Lemmatization

Training System

Components

- News Sources
- Kafka Server
- PySpark
- mysql
- Data Processor

Working

- -- We are using google API for fetching top 100 news article links on the basis of categories for every country. --- These links along with its category and country, will be stored in `mysql` database. --- Then every link is sent to `Kafka`. It retrieves the article from `Data Processor` and stores the article in `mysql` database. --- Then `Training server` collects the data from `mysql` databse, and apply some classification algorithm to get the model. --- Then this model is serialized and stored on the disk.

Prediction System

Components

- Client(Browser/API)
- Web Server
- Prediction Server
- Crawler

Working

--- `Client` will send the url(link) through UI interface/API. --- Then that url will be sent to `Fast API Prediction Server`, in turn to `Data Processor` to get the article. --- Then this article is sent to predction model to predict the article category. --- This predicted category is returned to `Client` and displayed on the interface.

Group Details

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