A MACHINE LEARNING PIPELINE FOR CLASSIFYING FINANCIAL LAW DOCUMENTS

Task

- **Premise:**

  In the first phase of your work as a data consultant for the law firm, you implemented a rule-based system to predict the relevance of legal documents to financial law. While rule-based systems are interpretable and straightforward, they lack the flexibility and adaptability of machine learning models. In this phase, your task is to build a machine learning pipeline that can automatically learn from the data and improve predictions over time.

- **Objective:**

  Your goal is to build an end-to-end machine learning pipeline that automates the process of identifying documents relevant to financial law. The pipeline should cover everything from data preprocessing to model deployment.

- **Data:**

  You will continue to work with the same datasets provided in Phase 1:

  - **Regulations Dataset**: Contains information about the documents themselves.
  - **Relevance Data**: Contains labels indicating whether each document is relevant to financial law or not.

Project Structure

Overview

This project implements an end-to-end machine learning pipeline for predicting the relevance of legal documents to financial law. The pipeline includes data processing, model training, and deployment of a web application for document classification.

Table of Contents

Project Description

The goal of this project is to build a robust machine learning pipeline that can classify documents based on their relevance to financial law. The pipeline involves:

- **Data Processing**: Cleaning and preparing data for modeling.
- **Model Training**: Experimenting with different machine learning algorithms.

- **Model Deployment**: Serving the model using FastAPI and a frontend interface with Streamlit.
- **Cloud Deployment**: Deploying the application on Render for live use.

Features

- **Data Ingestion**: Upload and process CSV files containing legal documents.
- **Text Processing**: Vectorize text using TF-IDF and concatenate relevant features.
- **Model Prediction**: Classify documents as relevant or not relevant to financial law.
- **Frontend Interface**: A user-friendly interface for uploading files and viewing results.
- **Deployment**: Hosted on Render with automatic scaling and monitoring.

Data

The dataset used includes the following columns:

- **Title**: The title of the document.
- **Content**: The content of the document.
- **SourceLanguage**: The language of the document content.
- **DocumentID**: Unique identifier for the document.

Data Preprocessing

- **Feature Engineering**: Combined **Content** and **SourceLanguage** for vectorization.
- **Encoding**: Applied TF-IDF vectorization to text data.

Installation

Prerequisites

- Python 3.10 or higher
- Docker
- Docker Compose

Setup

1. Clone the repository
2. Install dependencies
3. Build Docker images
4. Run Docker Compose

Usage

a. **Run the Streamlit App**
b. **Run the FastAPI App**
c. **Upload a CSV File**: Use the Streamlit interface to upload a CSV file and get predictions.

Deployment

The application is deployed on Render

Steps for Deployment

1. **Build Docker Images**: Ensure the Docker images are built with the latest code.
2. **Push Docker Images**: Push the Docker images to a container registry (e.g., Docker Hub).
3. **Configure Render**: Set up Render services for FastAPI and Streamlit applications, linking the Docker images and configuring environment variables.

Testing

Testing is performed using unit tests for both the FastAPI and Streamlit applications. To run tests:

a. **Run FastAPI Tests**:
b. **Run Streamlit Tests**:

Monitoring

Monitoring is set up using Render's metrics and logging services. For more detailed monitoring and alerting, integrate tools like Deepchecks and Evidently AI.

Challenge(s): Difficulty in using SHAP for model explainability, due to the virtual environment (Kaggle Notebook), running out of space, even with very small data samples.

Links to Required Submissions:

- **Link to Weight and Biases Report:** https://wandb.ai/oluwafemiolasupo123-student/financial-law-docs?nw=nwuseroluwafemiolasupo123
- **Link to Render Dashboard:** **A-Machine-Learning-Pipeline-for-Classifying-Financial-Law-Documents · Web Service · Render Dashboard**
- **Link to fastAPI: localhost:8080**
- **Link to Streamlit Frontend: Streamlit**

Acknowledgements

- **Streamlit**: https://streamlit.io
- **FastAPI**: https://fastapi.tiangolo.com
- **Docker**: https://www.docker.com
- **Render**: https://render.com

License