ПРОЕКТ ПО НЕЙРОИНФОРМАТИКЕ

Авторы проекта:
Артемий Знай (frontend)
Василий Амурский (backend, devops),
Арсен Яруллин (backend, ml)

ОПИСАНИЕ ПРОЕКТА

Данный проект посвящён созданию системы для автоматического определения языка текста и его перевода, используя микросервисную архитектуру и машинное обучение. Архитектура проекта включает несколько взаимосвязанных компонентов, каждый из которых выполняет отдельную функцию в рамках общей системы.

БИЗНЕС-ЦЕЛЬ

Целью проекта является разработка инструмента, который упростит пользователям взаимодействие с текстами на различных языках. Это решение может быть полезно для организаций, стремящихся улучшить доступность своих продуктов на международных рынках.

ЦЕЛЬ В ОБЛАСТИ МАШИННОГО ОБУЧЕНИЯ

Основная задача в рамках машинного обучения – внедрение модели, способной точно определять язык текста. Это позволит автоматически направлять текст к соответствующему сервису перевода, гарантируя удобство и скорость обработки.

АРХИТЕКТУРА

Система строится на основе микросервисной архитектуры, что обеспечивает модульность и масштабируемость. Структура включает следующие основные компоненты:

АРХИТЕКТУРА

1. **Frontend (Фронтенд):** Веб-интерфейс, с которым взаимодействуют пользователи.

2. **BFF (Backend for Frontend):** Прокси-сервер, координирующий поток данных между фронтендом и внутренними сервисами.

3. **Сервис для определения языка текста:** Использует модель машинного обучения для идентификации языка.

4. **Сервис перевода:** Подключён к API для выполнения перевода текста на целевой язык.

5. **ML модель:** Является ключевой частью сервиса определения языка.

МОДЕЛЬ

PAPLUCA/XLM-ROBERTA-BASE-LANGUAGE-DETECTION

Базовая архитектура
В основе модели лежит современная архитектура XLM-RoBERTa, которая позволяет ей эффективно работать с многоязычными данными

Цель модели
Основная задача модели — точно определять язык текста. Это задача классификации, где целью является правильно обозначить язык из заданного набора языков.

ОСОБЕННОСТИ

Предобученная модель:
Основа модели — XLM-RoBERTa base, которая эффективно обрабатывает данные на различных языках.

Файнтюнинг:
Модель прошла этап дообучения (fine-tuning) на задаче идентификации языка с использованием текстовых примеров.

Количество языков:
Поддерживает большое количество языков (потенциально более 50).

Практическое применение:
Подходит для задач мультиязычных платформ, требующих автоматического определения языка для персонализации контента или маршрутизации запросов.

ОБОСНОВАНИЕ АРХИТЕКТУРЫ

Выбранная микросервисная архитектура позволяет отдельно разрабатывать и развёртывать каждый модуль, что сокращает время на исправление ошибок и внедрение новых функций. При необходимости, системы могут масштабироваться в зависимости от нагрузки, что повышает общую устойчивость.

ОБОСНОВАНИЕ ВЫБОРА ТЕХНОЛОГИЙ

Используемые технологии отвечают поставленным задачам и требованиям:

Frontend: TypeScript(ReactJS) выбран за его гибкость и распространённость.

Backend: Python за его простоту.

СЕТЕВЫЕ АСПЕКТЫ

Система разделена на несколько зон безопасности:

DMZ (Demilitarized Zone): Для размещения компонентов, имеющих доступ к интернету, таких как фронтенд.

Secure Zone: Ограниченный доступ к BFF и внутренним сервисам для усиленной защиты данных.