SYSTECH

Middle Python Developer

Тестовое задание

Описание задачи

Представим, что Компания разрабатывает бота, который должен распознавать запросы пользователей на русском языке и находить нужную информацию в базе данных.

Для этого нужно:

- Реализовать базовый NLP-пайплайн для обработки текстовых данных.
- Настроить систему поиска по тексту (по принципу RAG или упрощенной версии).

Часть 1: NLP-пайплайн

Задание

Напишите на Python пайплайн для обработки текстовых запросов.
Пайплайн должен включать следующие этапы:

- Токенизация текста.
- Очистка текста от стоп-слов.
- Приведение к нижнему регистру и лемматизация.

Результат

REST API, которая на вход принимает строку и возвращает очищенный и нормализованный список слов (приложить скрипт обращения к апи)

Инструменты

Можно использовать библиотеки nltk или spaCy.

Часть 2: Поиск по тексту (упрощенный RAG)

Задание

Реализуйте упрощенный поиск по тексту, чтобы по запросу пользователя находить наиболее релевантные результаты из заранее подготовленного текстового файла (например, база отзывов на товары).

- Создайте индекс из текстов в базе данных, используя модель TF-IDF.
- Реализуйте функцию, которая принимает на вход текст запроса и возвращает топ-3 наиболее релевантных текста из базы.
- Применить функцию в апи.

Результат

REST API, которая на вход принимает запрос и возвращает список из 3 наиболее релевантных результатов (приложить скрипт обращения к апи).

Инструменты

Можно использовать библиотеки scikit-learn, nltk или spaCy.