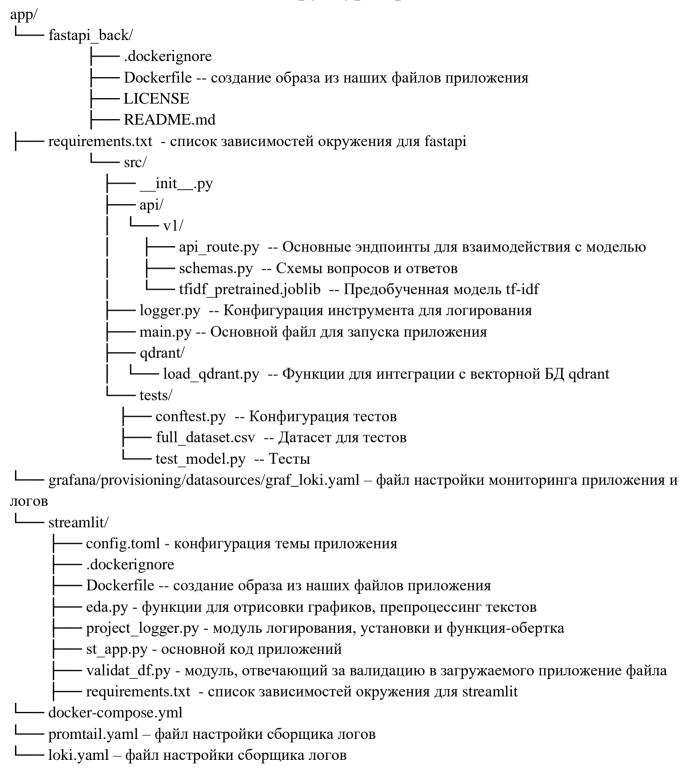
# Построение вопросно-ответной системы с использованием RAG (retrieval-augmented generation)

## 1. Структура проекта:



## 2. Функционал АРІ

- Загрузка датасетов
- Обучение моделей (в настоящее время поддерживается TF-IDF)
- Интеграция с векторной базой данных Qdrant для эффективного поиска по схожести
- Загрузка и выгрузка моделей из оперативной памяти (TODO)
- Поиск контекста для заданных вопросов

- Тестирование качества и скорости модели
- Вывод списка загруженных наборов данных и обученных моделей
- Удаление моделей (по одной или всех сразу)

#### **API Endpoints**

- `POST /api/v1/models/load\_dataset`: Загрузить датасет
- `POST /api/v1/models/fit save`: Обучить модель и сохранить в Qdrant
- `POST /api/v1/models/load model`: Загрузить модель в оперативную память
- `POST /api/v1/models/unload model`: Выгрузить модели из оперативной памяти
- `POST /api/v1/models/find\_context`: Найти контекст для заданного вопроса
- `POST /api/v1/models/quality test`: Оценить точность и производительность модели
- `GET /api/v1/models/get datasets`: Получить список загруженных наборов данных
- `GET /api/v1/models/list models`: Получить список загруженных и обученных моделей
- `DELETE /api/v1/models/remove/{model\_id}`: Удалить конкретную модель
- `DELETE /api/v1/models/remove\_all`: Удалить все модели

## 3. Функционал streamlit-приложения

- Загрузка датасета и анализ данных
- Препроцессинг данных
- Построение графиков
- Конфигурирование и обучение модели
- Сравнение моделей
- Получение инференса

## 4. Запуск приложения

#### А) Приложение развернуто на VPS:

- http://178.130.43.233:8501/
- http://178.130.43.233:3000/ Мониторинг и визуализацияя приложения и логов Сервер очень слабый и полный функционал приложения показать не сможет, максимум

допустимый датасет для использования на сервере должен содержать не более 100 записей.

Б) Запуск приложения на локальной машине:

Для запуска требуется

1. Установить по инструкции docker в зависимости от ОС

https://docs.docker.com/compose/install/

2. В папке арр прописать команду для сборки и запуска приложения:

docker-compose -p app\_rag up

- 3. Перейте по ссылке http://localhost:8501/ для входа в приложение
- 4. Для мониторинга приложения и сбора логов требуется:
- Перейте по ссылке http://localhost:3000/
- Подключить datasources loki, по адресу http://loki:3100/
- В разделе explore/loki ведем мониторинг приложения и логов.

Инструкция использования в файле https://drive.google.com/file/d/1B3RqPt2BVKPCuhqpltTlJFle6sKCiP\_t/vie w?usp=sharing