

## Документация проекта:

### 1. Структура проекта

- `./HSE-Project-RAG` – Директория с самой реализацией RAG и тг ботом. Изначально проект подразумевал оборачивание всего приложения именно в тг бота, но в рамках данного чекпоинта нам пришлось перейти именно к `fastapi` + `streamlit`. Следовательно вряд ли есть смысл сюда заглядывать, но описание же сделать стоило)
- `./data/features` – директория с описанием датасета используемого в рамках данного проекта.
- `./notebooks/eda` – директория с EDA проекта.
- `./pages` – директория со страницами Streamlit-приложения:
  - `Eda_page.py` – файл со страницей Streamlit, связанной с EDA.
  - `Rag_page.py` – файл со страницей Streamlit, связанной с RAG.
- `./utils` – директория со вспомогательными файлами:
  - `Create_functions.py` – файл с функциями для инициализации моделей используемых в RAG.
  - `Config.py` – файл с конфигом.
- `FastApi.py` – основной файл с `fastapi`.
- `RagClass.py` – файл содержащий себе класс в котором реализована большая часть логики проекта связанная с RAG.
- `StreamlitMain.py` – основной файл со Streamlit
- `Bot_icon.png`, `user_icon.png` – иконки для пользователя и бота в Streamlit приложении.
- `Requirements.txt` – файл с необходимыми версиями библиотек.

## Описание функционала:

### FastAPI сервер.

Основной функционал API реализован в FastAPI сервера. Для его запуска :

- 1) Необходимо создать вирт. окружение: `python -m venv venv`
- 2) Установить зависимости: `pip install -r requirement.txt`
- 3) Запустить сервер и приложение :
  - `python FastApi.py` – поднимется сервер FastAPI, по адресу `localhost:8000`
  - `streamlit run StreamlitMain.py` – поднимется сервер Streamlit, по адресу `localhost:8501`

### Более подробно:

`StreamlitMain.py` содержит в себе по сути ссылки на две страницы:

### Реализация RAG:

Тут есть возможность приложить свой датасет (главное чтобы название колонок совпадало) , который далее будет отправлен запросом на back-end, и загружен в векторную базу данных ( qdrant) . После чего будет возможность пообщаться с ботом. Вся логика с RAG отрабатывает на back-end'е.

- В качестве модели для retrieval части используем - sentence-transformers/all-MiniLM-L6-v2  
<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>
- В качестве генеративной модели - meta-llama/Llama-3.2-3B-Instruct.  
<https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>
- В качестве датасета – <https://huggingface.co/datasets/neural-bridge/rag-dataset-12000>.

Данная страница выглядит следующим образом:

## Чат с LLM на основе вашего CSV файла

### Загрузите CSV файл




Drag and drop file here  
Limit 200MB per file • CSV


Browse files



Загрузить новый файл


 Drag and drop file here  
Limit 200MB per file • CSV


Browse files


 rag\_dataset.csv 3.7MB ×

Загрузить новый файл

---

 **Чат**

 What is the Berry Export Summary 2028 and what is its purpose?



Although the provided text contains information about various topics, including berry export industries and BlackBerry devices, there is no direct mention of "Berry Export Summary 2028" or its specific details.

However, according to other available information (not provided in your text snippet), the 'Berry Exports Summary' is an annual report that provides an overview of Australia's berry exports.

Based on this external information, here's a possible answer to your question:

The Berry Import/Export Summary, likely referred to here, is likely a report or document produced by a relevant organization (such as a government agency, industry association, or research institution) that aims to summarize and analyze the current state of berry imports and exports in Australia. Its purpose would be to provide valuable insights and data to help stakeholders (growers, exporters, policymakers, etc.) understand market trends, identify opportunities and challenges, make informed decisions about future trade strategies, and ultimately support the growth and development of these industries.

Please note that this information is not directly extracted from your original text but rather inferred from external sources that provide more context about this specific topic

Задайте вопрос LLM:

What is the Berry Export Summary 2028 and what is its purpose?

Отправить

Итого весь пайплайн схематично:

- 1) Drag and Drop'ом прикладывается датасет.
- 2) Пассажи в этом датасете уже нарезаны, поэтому они сразу просто векторизуются и отправляются в векторную БД.
- 3) Сразу после этого можно задать вопрос LLM. Retrieval часть будет искать релевантный контекст и Llama будет также стараться ответить на вопрос пользователя.

## Реализация EDA:

Тут также есть возможность залить свой датасет, который также отправляется запросом на back-end. После чего можно выбрать любую модель ( у которой есть токенайзер, и посмотреть распределение по количеству токенов для загруженного

пользователем датасета. Количество бинов для визуализации пользователь может настроить, для построения графиков использовалась библиотека plotly.

Данная страница выглядит следующим образом:

## Streamlit EDA

Загрузите CSV файл



Drag and drop file here  
Limit 200MB per file • CSV

Browse files



rag\_dataset.csv 3.7MB

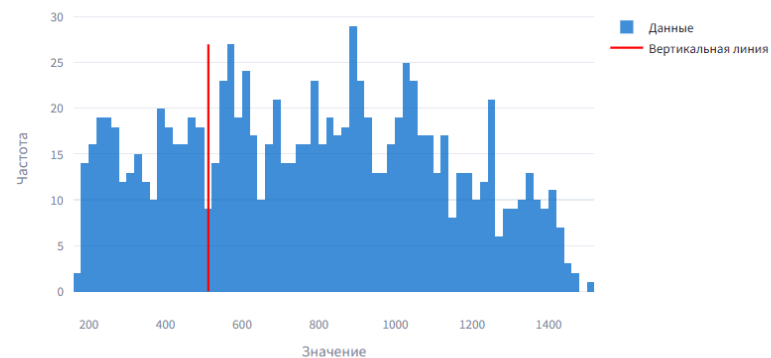


Введите название LLM (например, 'sentence-transformers/all-MiniLM-L6-v2')

sentence-transformers/all-MiniLM-L6-v2

Отправить на обработку

Выберите количество бинов



Итоговое решение было также “упаковано” в docker.