Цифровой ассистент “Акела” - фреймворк, содержащий реализацию цифрового ассистента на базе генерации языковыми моделями с улучшенным извлечением информации (RAG - retrieval augmented generation).

Взаимодействие пользователя с системой производится с помощью UI компоненты, реализованной с помощью библиотеки streamlit. После открытия вебсокет соединения с API “Акелы” и поступления запроса пользователя на вебсокет, происходит извлечение контекста с реранжированием и фильтрацией нерелевантных запросов с помощью базы данных Elasticsearch, языковой модели библиотеки SentenceTransformers, обученной на задачу семантической схожести текстов (STS - semantic text similarity), а также Ridge классификатора, реализация которого представлена в библиотеке sklearn. В случае, если запрос пользователя релевантен данным, представленным в базе знаний, контексты и вопрос используются для популяции промпта большой языковой модели. Посредством фреймворка Celery и базы данных Redis, промпт поступает на вход запущенной для распределенного инференса большой языковой модели, которая, в свою очередь, направляет результат генерации в стриминговом режиме с помощью Redis Streams на компоненту API, инкапсулирующую логику генерации ответов. Генерируемые токены, а также извлеченные контексты направляются в виде сообщений по открытому вебсокет соединению. UI компонента обрабатывает приходящие сообщения и отображает пользователю ответ в виде релевантных документов и результата генерации модели.

Пример работы ассистента для ответа на вопросы пользователя:
1. После установки и поднятия всех контейнизированных компонент системы, включая UI-компоненту, необходимо перейти по адресу <web-ui-host>:8501. Отображаемый интерфейс представлен на Рис. 1.



Рисунок 1 - Интерфейс ассистента

2. Затем необходимо выбрать LM модель и индекс ElasticSearch, которые будут использованы для генерации ответа на вопрос. Для этого в выпадающих окнах 1, 2, представленных на Рис. 1, нужно выбрать одну из представленных опций (или оставить значение по умолчанию).

3. Вопрос пользователя вводится в окно, обозначенное числом 3, после введения вопроса необходимо нажать enter.

4. На вкладке “Ответ” внизу страницы будет отображаться полученный результат генерации модели, вкладка “Документы”, будет содержать извлеченные из базы знаний релевантные ответу документы, используемые для генерации ответа.

5. Для перезапуска процесса генерации и ввода нового вопроса, необходимо нажать кнопку “Новый вопрос” (4).