Описание функциональных характеристик

TextAI.NTR

НТР

Содержание

Терминология.....3
1 Функциональные характеристики.....4
1.1 Назначение и область применения..... 4
1.2 Компоненты системы TextAI.NTR.....4
2 Информация, необходимая для установки и эксплуатации системы..... 5
2.1 Эксплуатация системы..... 5
2.2 Минимальные требования к оборудованию, на котором предполагается эксплуатация системы.....5
2.2.1 Веб-интерфейс системы.....5
2.2.2 Backend.....6
3 Уровень подготовки пользователей.....6

Терминология

LLM (large language model) - большая языковая модель, состоящая из нейронной сети со множеством параметров, обученной на большом количестве неразмеченного текста с использованием обучения без учителя.

Автоматизированное рабочее место – обозначение программного приложения, доступного в сети интернет.

1 Функциональные характеристики

1.1 Назначение и область применения

TextAI.NTR – программное обеспечение на основе LLM, предназначенное для решения задач в области поиска информации и ответа на вопросы. Пользователь задаёт в систему интересующие его вопросы по заранее загруженным документам, система ищет подходящие для ответа источники, а затем формирует ответ.

Основные решаемые задачи:

- поиск по документации;

- формирование ответа на вопросы в свободной форме;

- суммаризация информации из нескольких источников с помощью LLM.

Область применения: характеристики ПО позволяют внедрять его на всей цепочке производства и реализации продукта.

1.2 Компоненты системы TextAI.NTR

Описание компонентов TextAI.NTR:

- Пользовательский интерфейс. Первая часть предназначена для получения результатов поиска и обработки запроса. Вторая часть предназначена для изменения набора данных.
    - Функции первой части:
        - Задать запрос в систему и получить ответ;
        - Скачивание документа, по которому сформирован вопрос, в формате pdf;
        - Просмотр документа, по которому сформулирован вопрос, в системе;
        - Оценить ответ;
    - Функции второй части:
        - Удаление документов из базы;
        - Загрузка документов в базу;

    - Обновление документов в базе;

- Backend-сервис. Обеспечивает работу векторного поиска, загрузку системы аббревиатур, взаимодействие с LLM-бэкендом, с elasticsearch и постобработку ответа от LLM.
- LLM-backend. Предназначен для обеспечения корректной работы модели обработки естественного языка (LLM).
- Elasticsearch. Elasticsearch используется для реализации полнотекстового поиска, что позволяет повысить качество результатов поисковой обработки за счет комбинации с векторным поиском.

Компоненты системы поставляются в виде Docker-образов и предназначены для развёртывания в среде контейнеризации.

2 Информация, необходимая для установки и эксплуатации системы

2.1 Эксплуатация системы.

Эксплуатация системы происходит посредством развертывания Docker-контейнеров и автоматизированного рабочего места. Для адаптации и настройки TextAI.NTR под решение конкретных задач Заказчика система развертывается в локальной среде. Далее по требованиям заказчика вносятся необходимые изменения.

Адрес для доступа к системе создается новый при решении каждой новой задачи.

2.2 Минимальные требования к оборудованию, на котором предполагается эксплуатация системы.

2.2.1 Веб-интерфейс системы

В веб-интерфейсе реализована возможность журналирования запросов-ответов LLM на диск. В случае включения этой функции, нужно подключить том для записи.

Минимум:

- vCPU 0.5
- RAM 1 Гб.

2.2.2 Backend

Для данных и модели векторного поиска требуется том, смонтированный в /data.

Минимум:

- vCPU 1
- RAM 2 Гб
- VRAM 16 Гб

2.2.4 Сервер LLM API

Нагрузка на CPU происходит только при загрузке модели в видеопамять, после чего нагрузка на процессор минимальна. Оперативная память не задействуется интенсивно во время работы.

Минимум:

- vCPU 1
- RAM 2 Гб
- VRAM 80 Гб (Nvidia A100 80G)

3 Уровень подготовки пользователей

Пользователь «TextAI.NTR» должен иметь навык работы с любым из поддерживаемых интернет-браузеров (Google Chrome, Mozilla Firefox, Apple Safari, Microsoft Edge, Microsoft Internet Explorer).

Для работы с базой документов пользователь должен знать соответствующую предметную область.