Министерство науки и высшего образования Российской Федерации
Фереральное государственное бюджетное образовательное учреждение высшего образования
Российский государственный гидрометеорологический университет (РГГМУ)
Институт информационных систем и геотехнологий
Направление подготовки: 09.03.03 «Прикладная информатика»
Профиль подготовки: «Прикладные информационные системы и геотехнологии»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
(БАКАЛАВРСКАЯ РАБОТА)

На тему: «Разработка приложения интеллектуального ассистента на базе технологий глубокого обучения.»

Научный руководитель,
к.т.н _____Петров Я.А.

Исполнитель,
студент группы ПИ-Б20-2-2 _____ Попов В.Н.

Санкт-Петербург 2024 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

За последние десятилетия произошел огромный скачок в развитии информационных технологий: от создания первой электронной вычислительной машины, до сложный генеративных нейронных сетей (GAN). Сейчас компании проводят множественные исследования для выявления возможностей таких технологий и необходимости дальнейших инвестиций в данную отрасль. Одним из представителей GAN стали большие языковые модели (LLM). В настоящее время лидерами в данной отрасли стали: OpenAI, которые придумали реализовать интерфейс для взаимодействия с нейройнной сетью в виде чата; ПАО Сбербанк, реализовавшие отечественную LLM в условиях изоляции и ограниченных ресурсов; Meta (признана в РФ экстремистской организацией и запрещена), разработавшие малую языковую модель (SLM) LLaMA, ставшая прорывом для энтузиастов, у которых нет таких ресурсов для реализации LLM, как у больших игроков рынка.

В данной выпускной квалификационной работе будет разработана платформа, позволяющая взаимодействовать с ресурсами предприятия, использующая LLM/SLM для генерации релевантных ответов.

Актуальность данной темы обсуловлена возможностью оптимизации многих процессов, увеличении производительности сотрудников, повышении лояльности клиентов. В контексте высшего учебного заведения (ВУЗ), использование технологий подобного рода повышает конкуретноспособность ВУЗа, что наряду с предыдущими пунктами является положительной метрикой.

Объект исследования — большие языковые модели (LLM).

Предмет исследования — является применение языковых моделей для бизнеса.

Цель работы — проектирование и разработка приложения, которое позволяет взаимодействовать с структурой предприятий посредством интерфейса.

В качестве интерфейса для пользователя была выбрана оболочка в виде чат-бота. Чат-боты давно вошли в жизнь большинства населения. Это подтверждается информацией аналитической компании «eMarketer», согласно которой, чат-ботами пользуются более 1,4 млрд. человек на планете [2].

Для выполнения поставленной цели были поставлены следующие зада-

чи:

- Выполнить анализ предметной области;
- Провести сравнительный анализ информационных систем;
- Рассчитать сроки реализации проекта;
- Смоделировать схему бизнес-процессов;
- Составить описание документов бизнес-процессов;
- Сформировать перечень требований к ИС;
- Исследовать подходы SWOT;
- Описать сценарии вариантов использования;
- Визуализировать описанные сценарии вариантов использования;
- Создать модель диаграммы компонентов;
- Создать модель диаграммы развертывания;
- Реализовать бизнес-логику ассистента и перенести его в интерфейс бота;

В работе будет рассматриваться РГГМУ (далее Университет), но применяться бот сможет не только в конкретном учебном заведении, а для любых предприятий.

Во время разработки ассистента использовалась методология Agile. Она позволила работать в удобном темпе и формировать требования во время разработки.

В ходе выполнения практической части выпускной квалификационной работы были использованы:

Python, LangChain, FAISS, HuggingFace, Transformers, Docker.

1 Предпроектный анализ

1.1 Анализ предметной области

Индустрия информационных технологий является одной из наиболее динамичных и быстроразвивающихся отраслей, где каждый год появляются новые тенденции и совершенствуются технологии, которые позволяют улучшить пользовательский опыт и приносить большую выгоду бизнесу. Одной из ключевых технологий стала технология трансформер, предложенная в статье “Attention is all you need”.

Одной из ключевых особенностей трансформеров является их способность обрабатывать большие объемы текста без потери информации для выполнения задач таких как машинный перевод, обработка естественного языка, когнитивный анализ текста и генерирование текста. Трансформер состоит из блоков кодировщиков и декодеров, которые обрабатывают входные данные и генерируют выходные данные. Большое количество параметров сети позволяет ей улучшить качество работы по сравнению с другими моделями. В последнее время наблюдается тренд на внедрение больших языковых моделей в различные отрасли бизнеса, например системы автоматических ответов на вопросы, чат-боты, умные помощники.

Преимуществом таких решений является быстрый поиск информации и выдача её в удобоваримом виде, когда без использования таких ассистентов на поиск необходимой информации может потребоваться достаточный промежуток времени. В данной дипломной работе предлагается разработать информационную систему-помощника в виде чат-бота который будет представлять собой полезный инструмент как для студентов, так и для сотрудников ВУЗа. Основная идея информационной системы состоит в том, чтобы получить универсальный инструмент для взаимодействия со всей структурой университета. В рамках чат-бота пользователь сможет получить всю необходимую информацию, например информацию о заселении в общежития, списке необходимых документов для поступления т.п.

В общем и целом, интеграция технологии больших языковых моделей является актуальной и перспективной темой для дипломной работы, которая позволит изучить основы построения архитектуры приложения, интеграции технологий в предприятия, основы работы с нейросетями и машинным обучением, а также тестирования решений, где нет очевидных метрик для

измерения результата.

1.2 Сравнительный анализ

На данный момент прямых конкурентов у моего решения нет, но я не отрицаю того, что в настоящий момент может разрабатываться схожее решение. Из схожих решений можно отметить следующие решения:

Боты от университетов. Такие решения не имеют возможности масштабирования, имеют ограниченный пул вопрос/ответ и привязанны к какой-то определенной платформе.

Virtual Spirits. Эта зарубежная компания специализируется на создании на создании чат-ботов для различных предприятий. Из преемуществ имеется возможность настройки внешнего вида бота.

Сравнение моего приложения и приложений конкурентов приведено на таблице 1.1

Таблица 1.1 — Сравнительный анализ

Информационная система	Удобный сбор информации	Возможность неявного поиска	Необходимость аутентификации
Virtual Spirits	-	-	+
ИС от ВУЗ	-	-	+
Моя ИС	+	+	-

Опираясь на проведенный анализ можно подвести некоторый итог: В итоговой системе не будет системы авторизации, так как мне кажется, что вся информация должна быть в открытом доступе для всех возможных пользователей ИС.

Под удобным сбором информации подразумевается интутивно понятный процесс заполнения базы знаний, который может осуществляться как вручную, так и при помощи API, парсинга или других методов получения информации.

Возможность получать информацию не связанную с обучением мне кажется одним из ключевых преимуществ моей информационной системы: для абитуриентов может быть важно получить информацию как о возможном расписании, так и о постулении в ВУЦ, получении БСК или же информации о истории университета.

1.3 Системный анализ

Исследование проектируемой ИС проводилось в виде нескольких типов анализа: SWOT и ISA.

SWOT-анализ — метод стратегического планирования, суть которого заключается в выявлении факторов внутренней и внешней среды организации и разделении их на четыре категории: Strengths, Weaknesses, Opportunities, Threats. Сильные и слабые стороны представляют факторы внутренней среды объекта анализа. В свою очередь возможность и угрозы представляют собой внешнюю среду объекта анализа.

SWOT анализ находится в таблице 1.2

Таблица 1.2 — SWOT анализ

	Положительное влияние	Негативное влияние
Внутренняя среда	— Актуальность — Простота использования	— Нейросетевые галюцинации — Необходимость тщательно прорабатывать интеграцию во избежание проблем с безопасностью
Внешняя среда	— Упрощение навигации по ресурсам ВУЗа — Получение поддержки от государства	— Регуляции со стороны государства — Бюрократия с какой-либо стороны

Из положительных аспектов можно выделить простоту конечного использования и улучшение взаимодействие с предприятием.

Из отрицательных факторов стоит выделить следующие аспекты: нейросетевые галюцинации, проблемы с безопасность и бюрократия.

Нейросетевые галюцинации — аномалия, возникающая во время работы нейронной сети, влияющая на вывод результат непредсказуемым образом. Например, спросив о технической оснащенности предприятия нейросеть может начать галлюцинировать и ответить, что у предприятия есть несколько квантовых суперкомпьютеров, хотя это не так. Частично решить эту проблему может грамотный промт-инженеринг, а так же правильное подмешивание контекста в сам промт.

Проблемы с безопасность можно отнести в ту же категорию. Если занести секретные данные в базу знаний, то велик шанс утечки информации и попадание её в руки злоумышленников. Для обхода этой проблемы нужно

внимательно относиться к информации, которую вы собираетесь хранить в базе знаний и иметь базовые навыки кибербезопасности.

Бюрократия же не позволит внедрить информационную систему, занести все необходимые данные без множества согласований и утверждений со стороны вышестоящего руководства.

1.4 Требования к сервису

Функциональные и нефункциональные требования должны быть определены до начала реализации ИС, чтобы получить представление о конечном продукте.

Для классификации требований использовалась модель $FURPS$. В следующих таблицах были приведены функциональные и нефункциональные требования ИС. Эти требования характеризуют поведение ИС. Функциональные требования приведены в таблице 1.3.

Таблица 1.3 — Функциональные требования

Номер требования	Описание
FUN_1	При начале общения бот должен представиться и рассказать о своих возможностях
FUN_2	По требованию пользователя бот должен предоставить контактную информацию вышестоящего руководства
FUN_3	По требованию пользователя бот должен предоставить информацию о расписании
FUN_4	По требованию пользователя бот должен предоставить информацию о списке направлений, на которые можно поступить с определенными предметами
FUN_5	По требованию пользователя бот должен предоставить информацию о университете
FUN_6	По требованию пользователя бот должен предоставить актуальную информацию о местах проведения практики
FUN_7	По требованию пользователя бот должен предоставить информацию о проводимых мероприятиях внутри университета

В таблице 1.4 приведены нефункциональные требования, затрагивающие удобство использования приложения.

Таблица 1.4 — Удобство использования

Номер требования	Описание
NFR_1	Пользователю предоставляется информация о сервисе
NFR_2	Использование сервиса не требует от пользователя какого-либо обучения
NFR_3	Бот предоставляет как текстовый, так и голосовой интерфейс для общения
NFR_4	По требованию пользователя бот должен предоставить информацию о списке направлений, на которые можно поступить с определенными предметами
NFR_5	Информация, предоставляемая ботом должна быть избыточной

В таблице 1.5 приведены нефункциональные требования, затрагивающие надежность сервиса.

Таблица 1.5 — Надежность сервиса

Номер требования	Описание
REL_1	Сервиса должен работать 24 часа в сутки 7 дней в неделю, за исключением технических перерывов.
REL_2	Точность предоставляемой информации зависит от предоставляемых организацией данных

В таблице 1.6 приведены нефункциональные требования, затрагивающие производительность сервиса.

Таблица 1.6 — Производительность сервиса

Номер требования	Описание
PER_1	На генерацию ответа у сервиса должно уходить не более 10 секунд
PER_2	Для работы боту необходимо: постоянный выход в сеть интернет, 4ГиБ оперативной памяти. Так же, опционально, чтобы бот был полностью автономен необходима видеокарта уровня RTX 3060 и выше для размещения сервиса на SLM В противном случае необходимо пользоваться посредником в виде LLM от Сбербанк, OpenAI, Mistral и т.п.

В таблице 1.7 приведены нефункциональные требования по поддержке сервиса.

Таблица 1.7 — Поддержка сервиса

Номер требования	Описание
SUP_1	Установка сервиса осуществляется при помощи сценариев
SUP_2	Сервис должен поддерживать как работу с различными типами данных, поступающими от предприятия

В таблице 1.8 приведены нефункциональные требования к безопасности сервиса.

Таблица 1.8 — Требования к безопасности

Номер требования	Описание
SAF_1	Сервис должен использовать протокол HTTPS для обмена информацией

Изложенные требования в полной мере описывают работу ИС и позвляют реализовать корректную работу приложения.

1.5 Регламентирующие документы

Для соблюдения законодательства, соответствия требования заказчика, соответствия стандартам качества и урегулирования многих аспектов работы программного обеспечения используются регламентирующие документы: ГОСТ, пользовательское соглашение, рабочие инструкции и прочие.

Список регламентирующих документов приведен в таблице 1.9.

Таблица 1.9 — Регламентирующие документы

№	Наименование документа	Внутренний документ	Внешний документ
1	Политика приватности	+	-
2	ГОСТ 17657-79 Передача данных.	-	+
3	ГОСТ 34.321-96. Информационные технологии. Система стандартов по базам данных. Эталонная модель управления данными	-	+
4	ГОСТ Р 51904-2002. Программное обеспечение встроенных систем	-	+
5	ГОСТ 19.201-78. Техническое задание. требования к содержанию и оформлению	-	+
6	ГОСТ 19.401-78. Текст программы. требования к содержанию и оформлению	-	+

7	ГОСТ 34.201-89. Виды, комплектность и обозначение документов при создании автоматизированных систем	-	+
8	ГОСТ Р 50779.30-95. Приемочный контроль качества	-	+

1.6 Сроки реализации проекта

Для оценки требуемых ресурсов – времени и денег, разработаем график работ. Распишем основные блоки работ – это анализ, проектирование, разработка и документирование. Таким образом получаем, что на разработку системы потребуется 259 дней, что является довольно быстрым сроком. Для оценки стоимости проекта возьмем средние зарплаты специалистов и получим что на оплату труда уйдет 1,38 миллиона рублей. График работ представлен на рисунке 1.1 позволяет наглядно оценить длительность и стоимость работ.

№	Название этапа	Дата начала работ	Длительность	Дата окончания работы	Ответственное лицо	Заработная плата	Итого
1.0	Предпроектный анализ	01.09.2023	37	10.10.2023	Системный аналитик	5000	185000
1.1	Анализ предметной области	01.09.2023	14	15.09.2023	Системный аналитик	5000	70000
1.2	Анализ объекта исследования	16.09.2023	15	01.10.2023	Системный аналитик	5000	75000
1.3	Составление ТЗ	02.10.2023	8	10.10.2023	Системный аналитик	5000	40000
2.0	Проектный анализ	11.10.2023	19	31.10.2023	Системный аналитик	5000	95000
2.1	Анализ функциональных блоков	11.10.2023	10	21.10.2023	Системный аналитик	5000	50000
2.2	Анализ способов реализации	22.10.2023	9	31.10.2023	Системный аналитик	5000	45000
3.0	Проектирование	01.11.2023	40	14.12.2023	Системный аналитик	5000	200000
3.1	UR	01.11.2023	10	11.11.2023	Системный аналитик	5000	50000
3.2	BPMN	12.11.2023	10	22.11.2023	Системный аналитик	5000	50000
3.3	UML	23.11.2023	10	3.12.2023	Системный аналитик	5000	50000
3.4	Концептуальная модель	4.12.2023	10	14.12.2023	Системный аналитик	5000	50000
4.0	Реализация	15.12.2023	91	16.03.2024	Инженер-программист	6000	546000
4.1	Разработка модулей интеграции	15.12.2023	62	15.02.2024	Инженер-программист	6000	372000
4.2	Разработка бота	16.02.2024	29	16.03.2024	Инженер-программист	6000	174000
5.0	Отладка	17.03.2024	10	27.03.2024	Инженер-программист	6000	60000
5.1	Отладка кода модулей	17.03.2024	10	27.03.2024	Инженер-программист	6000	60000
6.0	Тестирование	28.03.2024	32	01.05.2024	Тестировщик	5500	176000
6.1	Альфа-тест	28.03.2024	10	07.04.2024	Тестировщик	5500	55000
6.2	Отчет по итогам альфа-теста	08.04.2024	5	13.04.2024	Тестировщик	5500	27500
6.3	Бета-тест	14.04.2024	11	25.04.2024	Тестировщик	5500	60500
6.4	Отчет по итогам бета-теста	25.04.2024	6	01.05.2024	Тестировщик	5500	33000
7.0	Оформление документации	02.05.2024	30	01.06.2024	Тех.писец	4000	120000
7.1	Заполнение отчетности	02.05.2024	30	01.06.2024	Тех.писец	4000	120000
Примерная стоимость проекта:		01.09.2023	259	01.06.2024			1382000

Рисунок 1.1 — Диаграмма затрат

Так же для наглядности временных затрат была использована «диаграмма Ганта», позволяющая наглядно отследить каждый этап разработки информационной системы.

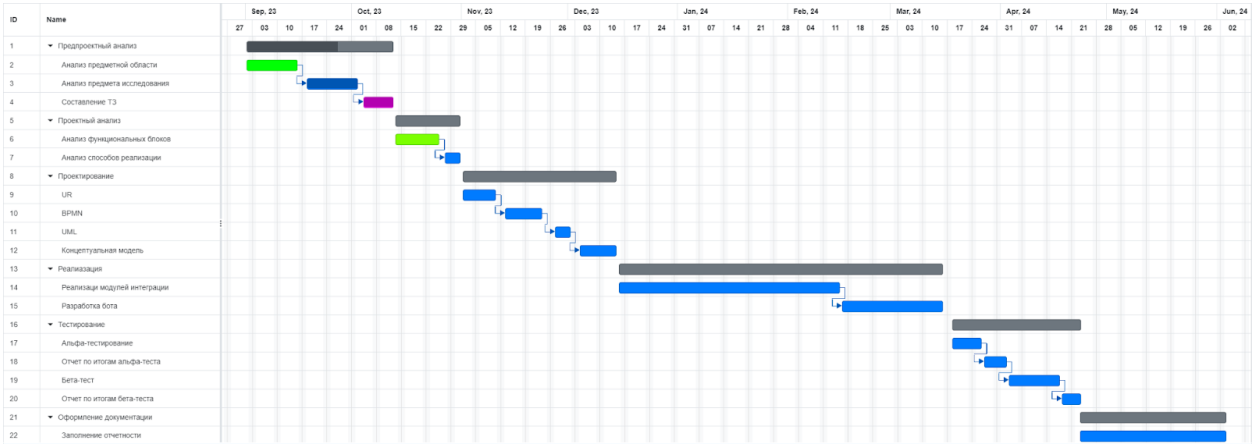

Рисунок 1.2 — Диаграмма Ганта

2 Проектирование информационной системы

2.1 Концептуальное проектирование

Для отображения функциональности системы используется диаграмма вариантов использования, представленная на рисунке 2.1.

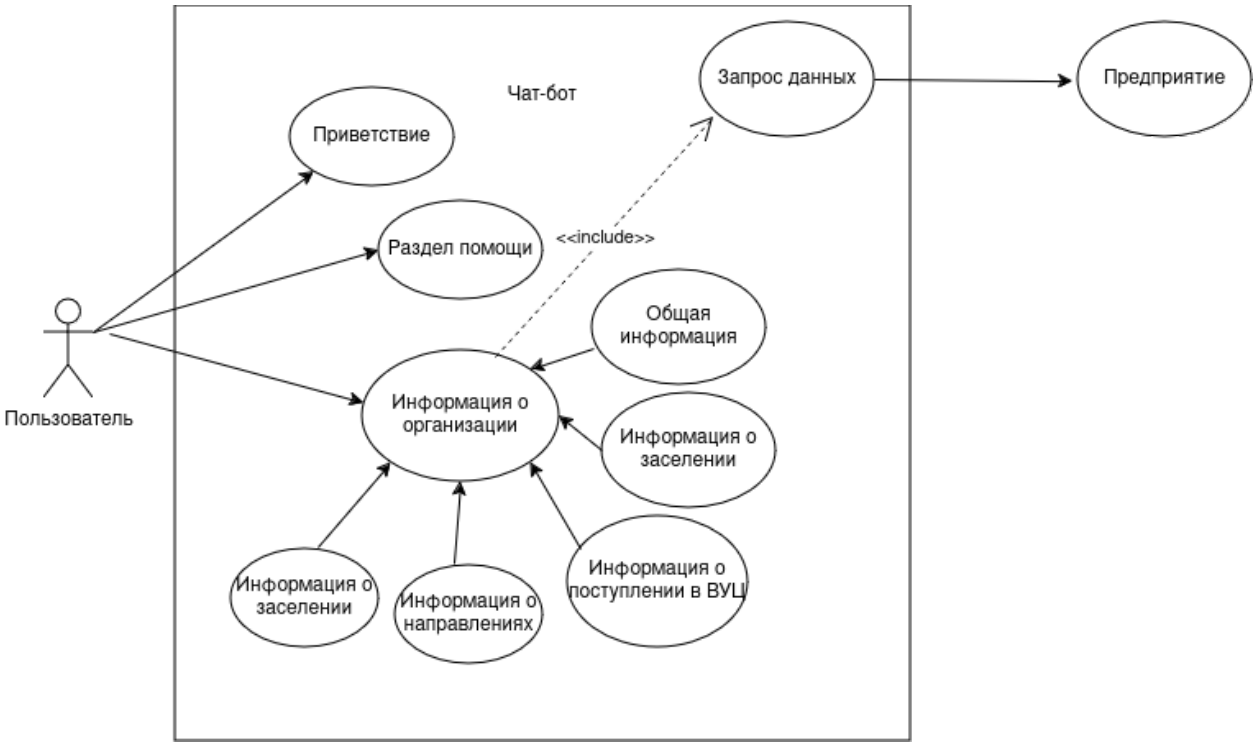

Рисунок 2.1 — Диаграмма вариантов использования

Актор *Пользователь* является обобщением клиентов, которые использует систему. Прецедент Приветствует отражает требование FUN_1. Этот и другие прецеденты описывают субъект Бот, который на диаграмме выделен рамкой. Прецедент Рассказывает о возможностях отражает требование FUN_1. Актор Образовательная организация является источником данных, которые необходимы субъекту. Прецедент *Информация о организации* отражает функциональные требования: FUN_2, FUN_3, FUN_4, FUN_5, FUN_6, FUN_7, и используется для логического объединения других прецедентов, для уменьшения количества связей на диаграмме и облегчить ее восприятие.

Для понимания перемещения данных внутри приложения была разрабработана схема перемещения данных, представленная на рисунке 2.2.

Рисунок 2.2 — Схема перемещения данных

2.2 Диаграмма компонентов

Диаграмма компонентов описывает физическое представление системы и является структурной диаграммой языка унифицированного моделирования. Она определяет архитектуру разрабатываемой системы, устанавливая зависимости между программными компонентами. Кроме того, она предоставляет разработчикам и архитекторам общую картину архитектуры системы, помогает им лучше понимать ее структуру и взаимосвязи, а также

является полезным инструментом для коммуникации и документирования архитектурных решений. Разработанная диаграмма представлена на рисунке 2.3.

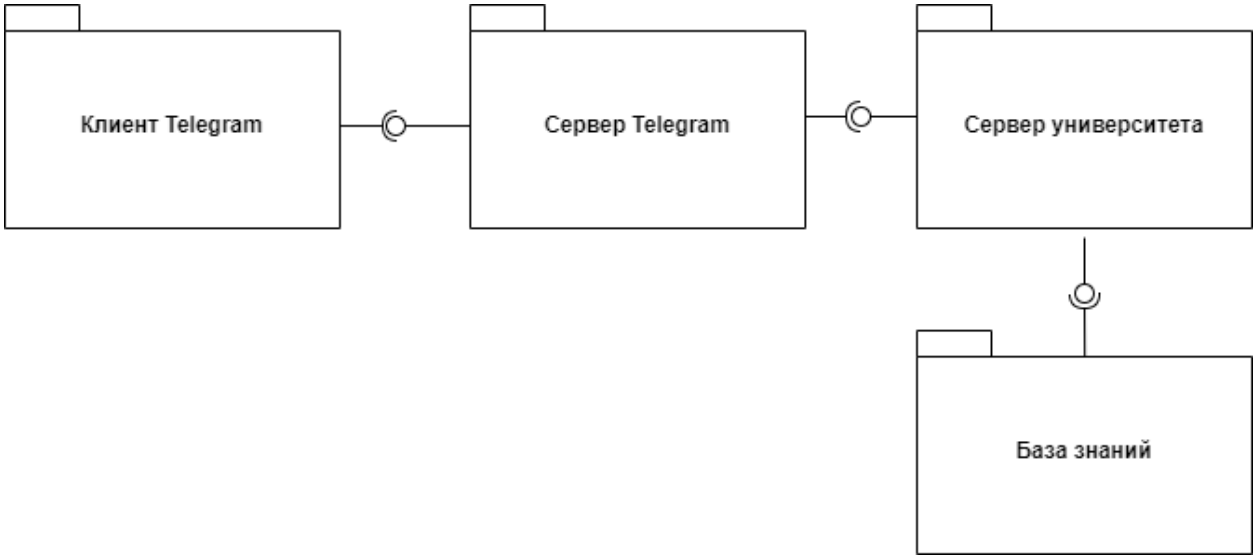

Рисунок 2.3 — Диаграмма компонентов

2.3 Диаграмма развертывания

Диаграмма развертывания – это тип UML-диаграммы, которая показывает архитектуру исполнения системы, включая такие узлы, как аппаратные или программные среды исполнения, а также промежуточное программное обеспечение, соединяющее их. Для каких задач строят диаграммы развертывания:

1) Визуализация полной структуры исходного кода
2) Визуализация узлов системы для определения слабых мест
3) Обеспечение многократного использования отдельных фрагментов программного кода

Диаграмма разрабатываемой информационной системы содержит несколько узлов: сервер базы знаний внутри университета, сервер приложения, сервер мессенджера, на котором базируется бот, конечный аппарат, при помощи которого будет производиться взаимодействие с информационной системой.

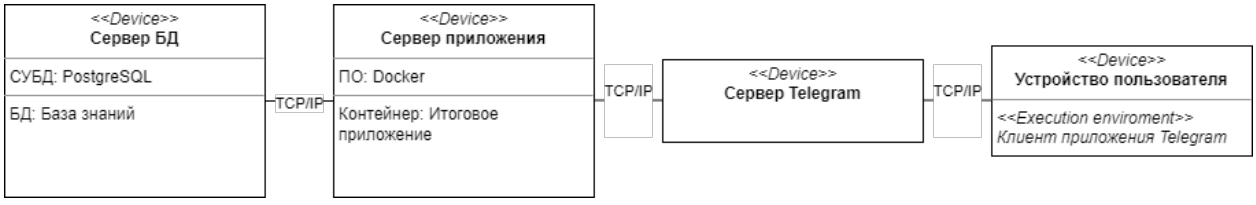

Рисунок 2.4 — Диаграмма развертывания

2.4 Диаграмма последовательности

Пайплайн действий таков: пользователь вводит запрос в текстовом формате или формате аудио, если сообщение передано в аудиоформате, то происходит расшифровка аудиозаписи, после чего полученный текст преобразуется в векторное представление, далее происходит извлечение подходящих документов методом ближайших соседей (KNN) из базы знаний, сформированной из документов в базе данных.

Далее, наиболее релевантные документы подмешиваются к запросу языковой модели, после чего языковая модель генерирует ответ, опираясь на полученный контекст.

Последовательность работы приложения приведена в приложении А.1.

2.5 Схема базы данных

В качестве базы данных для хранения изначальных документов для построения retirivial использовалась система управления базами данных (СУБД) PostgreSQL. Данное решение обладает рядом преимуществ: высокая производительность при работе с большими объемами информации; PostgreSQL является свободным программным обеспечением, что позволяет использовать данное ПО для любых нужд на бесплатной основе; удобство использования и возможность удобного переноса данных из других баз данных. Данные пункты стали ключевыми при выборе СУБД для хранения информации.

Такой выбор СУБД позволяет обеспечить высокую отказоустойчивость. Схема хранения данных внутри базы данных указана на рисунке 2.5

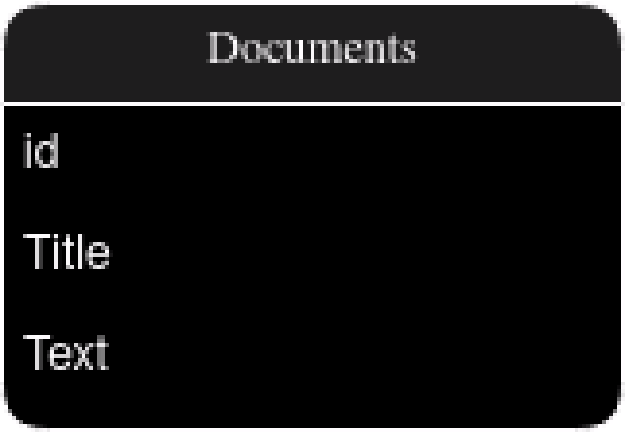

Рисунок 2.5 — Схема хранения документов

В качестве базы данных для хранения представления векторных корпусов из документов была выбрана FAISS, так как: её удобно использовать,

она использует меньше памяти, чем другие базы данных для векторов, можно использовать вычислительные мощности как процессора, так и видеокарты.

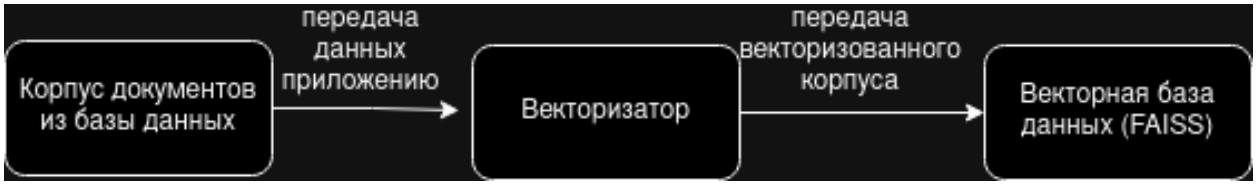

Рисунок 2.6 — Схема векторизации документов

2.6 Развертывание приложения

Планируется, что приложение будет размещаться на мощностях образовательного учреждения. Для упрощения внедрения было принято решение разработать скрипт для контейнеризации приложения.

В качестве базового образа был выбран образ с GNU Linux в качестве операционной системы. GNU Linux является свободным программым обеспечением, что позволяет использовать его для любых нужд. На листинге 2.1 отображена конфигурация Dockerfile.

Листинг 2.1 — Dockerfile

```
1   FROM python:3.12.3-bullseye
2
3   WORKDIR /app
4
5   COPY . .
6
7   RUN pip install -r --no-cashe-dir req.txt
8
9   RUN apt-get update && apt-get install -y nano
10  RUN apt-get update && apt-get install -y wget bzip2
    ↪  libxtst6 libgtk-3-0\
11   libx11-xcb-dev libdbus-glib-1-2 libxt6 libpci-dev
12
13  CMD ["/bin/bash", "-c", "python main.py"]
```

Данный файл позволяет создать контейнер, содержащий все необходимые необходимые для работы зависимости.

3 Разработка системы

3.1 Выбор средств разработки

В качестве языка программирования для разработки системы был выбран Python. Выбор связан с набором ПО, необходимым для функционирования сервиса, который был определен на последнем этапе проектирования. Также Python является наиболее подходящим языком для работы с искусственным интеллектом, для него имеется много библиотек, нацеленных на решение конкретных проблем.

В качестве системы управления версиями был выбран Git. В качестве системы управления репозиториями был выбран GitHub. С недавнего времени GitHub располагает возможностью для создания бесплатных приватных репозиториев — никто, кроме авторизовазанных пользователей не будет иметь доступа к вашему репозиторию.

В качестве системы для написания кода был выбран текстовый редактор NeoVim. NVim ориентирован на разработку без мыши, что позитивно влияет на скорость разработки; можно запустить без графического окружения рабочего стола, что позволяет использовать данный текстовый редактор как на домашнем персональном компьютере, так и на сервере, не имеющим графической оболочки. Конфигурация приведена на листинге Б.1.

Для верстки PDF-документа был использован LaTeX. С использованием пакетов данный инструмент позволяет достичь отличных результатов для написания текстовых документов. Так же данное программное обеспечение является свободным.

3.2 Структура приложения

Созданное приложение имеет модульную структуру. Это позволяет распределить зоны ответственности каждого модуля [4]; к тому же это позволяет повысить отказоустойчивость конечного приложения.

Схема файловой структуры показана на листинге 3.1.

Листинг 3.1 — Файловая структура

```
|-- Dockerfile
|-- documents
|------about.txt
|------checking.txt
|--  common_questions.txt
|--  deductions_recoveries.txt
|-- embeding_store
|------index.faiss
|------index.pkl
|-- LICENSE
|-- main.py
|-- messages.log
|-- README.md
|-- req.txt
|-- retriveval.py
|-- speech2text.py
|-- tex
|-- text_generator.py
|-- videohandler.py
```

Файл *Dockerfile* содержит инструкцию для сборки контейнера и последующего запуска приложения внутри контейнера.

Директория *documents* содержит основные документы в текстовом формате, предназначенные для дальнейшего перевода их в векторное представление. По мере развития проекта данный список может меняться.

Директория *embeding_store* содержит эмбединги документов.

Файл *main.py* запускает основное приложение: запуск retrivial, speech2text, запуск логирования, а так же телеграмм клиента.

Файл *retrivial.py* несет в себе следующий смысл: перевод корпуса текстов в векторное представление, извлечение 3х наиболее релевантных документов для составления контекста, а также обогащение запроса пользователя извлеченной информацией, для составления ответа.

файл *speech2text.py* представляет собой блок расшифровки в аудио в текстовый формат, для использования блоком *retrivial.py*, если это потребуется.

Файл *videohandler.py* представляет собой блок разделения видео- и зву-

ковых дорожек. Извлеченная звуковая дорожка используется блоком *speech2text.py*.

3.3 Расчет надежности программного и аппаратного обеспечения

Для расчета надежности реализуемого продукта необходимо выявить элементы системы, которые могут выйти из строя. Для себя я поделил их на 3 категории, и составил таблицы 3.1, 3.2 и 3.3. Программная часть, аппаратная часть, независящие от обстоятельств. Данные для столбца 'Коэффициент надежности' брались из открытых источников.

Таблица 3.1 — Аппаратная часть

Номер	Коэффициент надежности	Краткое описание	Что повлечет?
1	0.99	Выход из строя жесткого диска (HDD)	Поломка HDD влечет за собой: потерю данных, простой системы.
2	0.99	Выход из строя процессора (CPU)	Поломка CPU влечет за собой: простой системы; если мы используем сервер, то потерю производительности; большие траты на замену и обслуживание
3	0.87	Выход из строя графического процессора (GPU)	Выход из строя GPU грозит: снижением производительности или полным выходом из строя системы, большие затраты на замену и обслуживание
4	0.85	Выход из строя оперативной памяти	Выход из строя модуля оперативной памяти не так критичен для системы, модулей памяти несколько и вышел один из них, но если сломались все плашки оперативной памяти, то система не сможет функционировать.
5	0.95	Выход из блока питания	Выход из строя блока питания сулит полный выход строя системы

Таблица 3.2 — Программная часть

Номер	Коэффициент надежности	Краткое описание	Что повлечет?
1	0.93	Выход из строя модуля, ответственного за перевод звука в текст	Выход из строя данного модуля не полечет за собой больших проблем для обычных пользователей, так как будет доступен обычный текстовый ввод, но для людей с ограниченными возможностями перестанет быть возможным использование моей системы
2	0.84	Выход из строя модуля, ответственного за перевод документов в векторное представление	Поломка данного модуля может делиться на две категории: перестает работать перевод текста в векторное представление и перестает работать извлечение векторов. В первом случае не все так критично, так как часто обновлять список актуальных документов не нужно, ведь большая часть информации не меняется. Если же сломается модуль, ответственный за извлечение векторных представлений, то система в большей своей части перестанет работать, как задумывалось, но все равно будет работать: языковая модель будет выдавать ответы, но не будет иметь контекста о университете, что негативно повлияет на точность ответа.

Продолжение таблицы 3.2

Номер	Коэффициент надежности	Краткое описание	Что повлечет?
3	0.77	Выход из строя модуля клиентского приложения	Выход из строя клиентского приложения сулит собой: репутационные риски, недоступность приложения для пользователей

Таблица 3.3 — Стохастичные переменные

Номер	Коэффициент надежности	Краткое описание	Что повлечет?
1	0.99 – 0.97	Массовое отключение интернета	Без подключения к сети интернет система полностью перестает работать. Это можно будет решить, если поднять локальную сеть
2	0.99 – 0.97	Отключение электричество	Без электричества система полностью перестает работать

Расчитывая стохастические переменные я отталкивался от того, что выключить интернет или электричество на сутки могут от 2 до 10 раз

Так как моя система избыточна: множество модулей позволяют работать другим модулям при выходе из строя, то я буду считать отказоустойчивость системы по следующей формуле: $R_{total} = 1 - (1 - R_1)(1 - R_2)\dots(1 - R_n)$, где R_{total} — общая отказоустойчивости системы, R_n — отказоустойчивость компонента системы.

По расчетам надежность аппаратно-программного комплекса $\approx 1 - 2.5116e-14$, что я считаю довольно хорошим показателем для такой системы.

3.4 Расчет ожидаемого результата экономической эффективности

Основными затратами для реализации информационной системы являются:

– Покупка необходимого аппаратного обеспечения для обеспечения работы языковой модели

– В случае, если предприятие не хочет или не может позволить аппаратное обеспечение для развертывания системы на своих мощностях, то оформление подписки у предприятий и аренда мощностей для получения доступа к языковой модели

– Обслуживание сервера

Предполагается несколько концепций для получения прибыли от конечного продукта: Продажа подписки коммерческим и некомерческим организациям, по предоставлению конечного продукта. Для некоммерческих организаций подписка будет стоить от 5000 рублей в месяц, до 15000 рублей в месяц. В свою очередь для коммерческих организаций планируется разрабатывать уникальный договор по оформлению подписки на конечный сервис, для получения максимальной выгоды от сделки.

К подписочной модели получения прибыли так же не исключается возможность продажи программного обеспечения. В таком случае цена за конечный продукт так же определяется уникальным договором. В таком случае конечный потребитель не платит за подписку и сам обслуживает всю систему, но имеет над ней полный контроль. Со стороны продавца передается документация к программному обеспечению.

Для расчета возврата инвестиций используется следующая формула: $ROI = \frac{profit - cost}{cost} \cdot 100\%$ Если считать минимальный доход, то выходят примерно следующие цифры: $ROI = \frac{60000 - 15000}{15000} \cdot 100\%$, что равно 300%. Но в эти цифры не заложена покупка аппаратного комплекса, так как рассматривается ситуация, что я у предприятия есть мощности для размещения малой языковой модели.

3.5 Word2vec

Для хранения представлений документов в понятном для машины виде используется подход Word2vec.

Преобразование текстовых корпусов в вектора работает следующим образом: после передачи на вход программе текста программа создает для каждого слова вектор, после чего происходит процесс обучения. Во время обучения происходит попытка предсказать контекст слова, основываясь на его окружении. Например, если в тексте явно кореллируют слова “упряжка“

и “конь“, то эти векторные представления данных слов будут иметь сходство. Визуализацию данного свойства можно наблюдать на рисунке 3.1

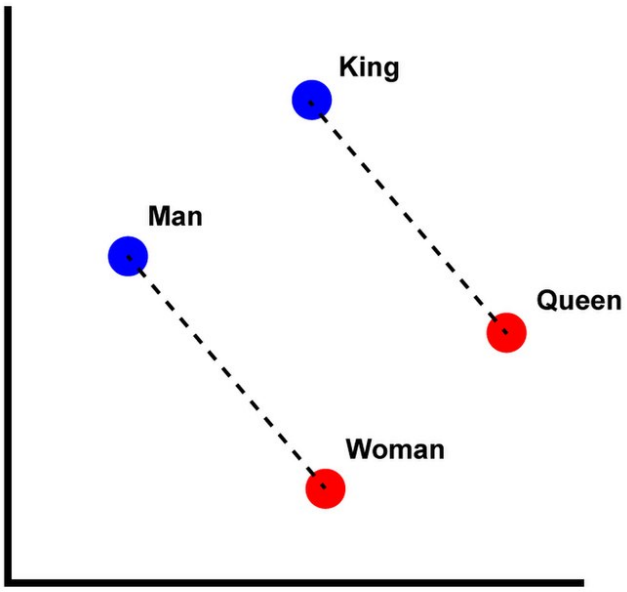

Рисунок 3.1 — Взаимосвязь векторов слов

В конце обучения создается векторное пространство, используемое в дальнейших этапах работы приложения.

Так как большинство решений не оптимизированы для применения на домашних персональных компьютерах (ПК) было принято решение переопределить некоторые методы векторизатора из библиотеки HuggingFace, чтобы улучшить работу на своем ПК. На листинге 3.2 отображены переопределенные методы.

Листинг 3.2 — Векторизатор

```
class HuggingFaceE5Embeddings(HuggingFaceEmbeddings):
    def embed_query(self, text: str) -> List[float]:
        text = f"query: {text}"
        return super().embed_query(text)

    def embed_documents(self, texts: List[str]) ->
    ↪ List[List[float]]:
        texts = [f"passage: {text}" for text in texts]
        return super().embed_documents(texts)

    async def aembed_query(self, text: str) ->
    ↪ Coroutine[Any, Any, List[float]]:
        text = f"query: {text}"
        return await super().aembed_query(text)

    async def aembed_documents(
        self, texts: List[str]
    ) -> Coroutine[Any, Any, List[List[float]]]:
        texts = [f"passage: {text}" for text in texts]
        return await super().aembed_documents(texts)
```

3.6 Поиск подходящих векторов

Созданное векторное пространство делится на множество кластеров методом K-средних. Для каждого кластера выставляется центроида для определения сущности каждого корпуса векторов. Это позволяет не перебирать все вектора, а проходиться по принадлежности к какой-то сущности. При большом наборе данных это позволяет быстрее находить необходимый вектор. На рисунке 3.2 схематично отображен алгоритм кластеризации.

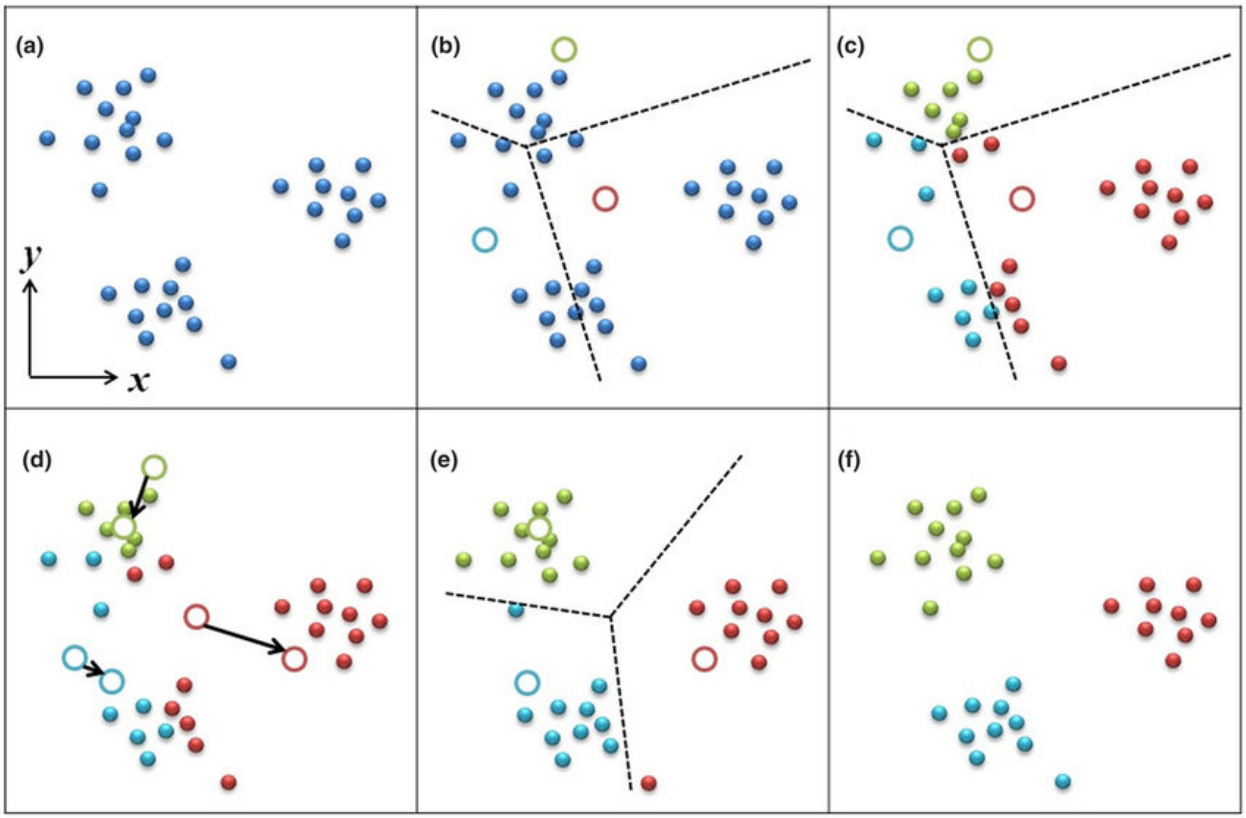

Рисунок 3.2 — Визуализация кластеризации векторов

Наиболее релевантные вектора [3] сопоставляются при помощи метода K-ближайших соседей. Евклидово расстояние между векторами вычисляется по следующей формуле:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Так же для поиска похожих векторов используется косинусное расстояние:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

И хотя данный подход не покрывает все возможные случаи [6], он является наиболее универсальным.

Визуализацию работы поиска релевантных векторов можно на рисунке 3.3

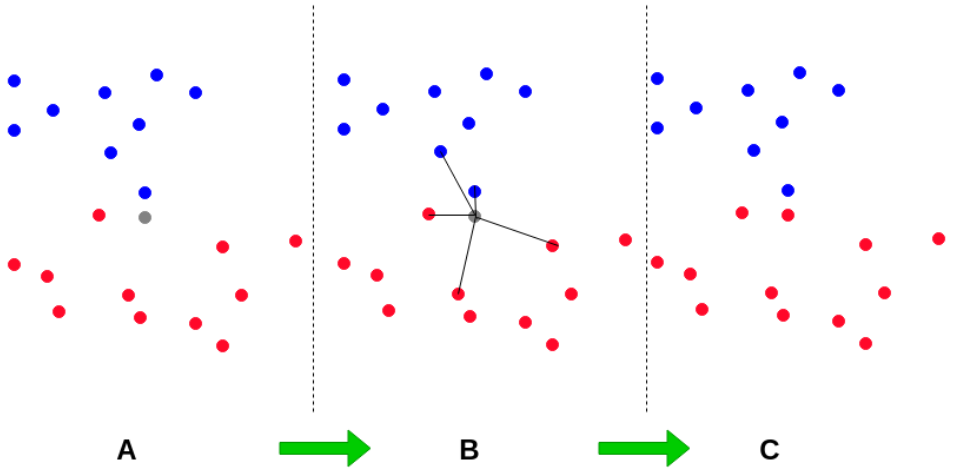

Рисунок 3.3 — Визуализация нахождения релевантных векторов

3.7 Генерация текста при помощи поиска и аугментаций

Для работы системы необходимо подать на вход какой либо текст. В качестве клиента было выбрано клиентское приложение Telegram, а в качестве интерфейса для взаимодействия приложения и пользователя TelegramBotAPI, но в качестве клиента можно использовать и другие платформы: ВК, Яндекс.Алиса, Одноклассники и прочие. За последнее время боты стали очень популярным решением [5] благодаря удобности и доступности, что позволяет бизнесу удобно использовать подобные интерфейсы в своих интересах. Чтобы пользователю было удобно использовать созданное приложение было принято описать идею приложения, а так же дать ссылку на вызов функции-помощника. Данный аспект отражает нефункциональные требования NFR_1 и NFR_2.

Для пользователей предусмотренна возможность ввода как текстового сообщения, так и управления ботом при помощи голоса или видео. Данная возможность отражает нефункциональное требование NFR_3.

На листинге 3.3 отображена реализация приветственного сообщения от бота.

Листинг 3.3 — Стартовое сообщение

```
    dummy_message = "Ищу ответ..."
    stub = hbold("Бот работает в тестовом режиме - не все ответы могут быть достоверными")
    # Start message
    @dp.message(CommandStart())
    async def command_start_handler(message: Message) -> None:
        """
        This is handler for command `/start`
        """
        await message.answer(
            f"Привет, {hbold('студент')}\n\nПеред тобой бот-асситент ИИСИГТ,\
     подготовленный {hlink('мной', 'https://github.com/yaz0p')} к качестве\
     выпускной квалификационной работы\n\nДанный бот позволяет осуществлять\
     удобную навигацию внутри университета и отвечает на не самые очевидные\
     вопросы. \n\nБот работает на основе большой языковой модели, а\
     информацию о внутренних источниках получает благодаря {hbold('RAG')}.\
     Возможно, я прикреплю сюда git репозиторий с реализацией проекта,\
     чтобы студенты могли изучить работу данного ассистента и/или дополнить\
     его своими идеями.\n\n\
    Пример того, что ты можешь спросить:\n- Адрес какого либо-корпуса\n\
    - Номер приемной комиссии\n- Как поступить в ВУЦ\n- Время приема ректората\
     - Информацию о заселении и медкомиссии\n"
        )
        sleep(2)
        await message.answer(
            f"Для получения инструкции о взаимодействии с\
            ассистентом нажмите /help\n\n\
    {hbold('Внимание, все сообщения логгируются!')}"
        )
```

После получения сообщения сервером Telegram и передачей сообщения приложению по протоколу HTTPS, что отражает нефункциональное требование к безопасности SAF_1, происходит перевод полученного текста в вектор и сопоставление полученного вектора с векторами из созданного векторного пространства базы знаний. После поиска подходящих векторов происходит обратная трансформация из векторного пространства в текстовое представление.

На листинге 3.4 отображен процесс создания контекста и промт для корректной работы языковой модели.

Листинг 3.4 — Функция для поиска векторов и создания контекста

```
1     def augment_prompt(self, query: str,
      ↪  _embedings_storage=_embedings_storage):
2         results =
          ↪  _embedings_storage(self).similarity_search(query,
          ↪  k=3)
3         source_knowledge = "\n".join([x.page_content for x
          ↪  in results])
4         augmented_prompt = f"""Используя предоставленный
          ↪  контекст ответь на следующий вопрос.
5
6         Контекст:
7         {source_knowledge}
8
9         Вопрос: {query}"""
10        return augmented_prompt
```

Далее обогащенный текстовый запрос подается на вход языковой модели, после чего происходит генерация текстовых токенов декодером [1]. На листинге 3.5 отображен модуль, ответственный за генерацию текста.

Листинг 3.5 — Функция для генерации ответа на заданный вопрос

```
        def answer(self, message: Message) -> None:
            prompt = PromptTemplate(
                template="Ответь на поставленный вопрос: {subject}",
                input_variables=["subject"],
            )

            output = self.chat(
                [
                    SystemMessage(
                        content="""Ты профессиональный ассистент Российского государственного гидрометеорологического университета, который помогает студентам и преподавателям решать их проблемы: навигация внутри университета, ответ на общие вопросы, касающиеся РГГМУ. Если контекст большой, то используй его по максимуму, но не теряй сути. При ответе на вопросы общайся в деловом стиле и пиши только по существу. """
                    ),
                    HumanMessage(
                        content=prompt.format(
                            subject=self.
                            augmentation.
                            augment_prompt(message)
                        )
                    ),
                ]
            )

            return output.content
```

ЗАКЛЮЧЕНИЕ

В процессе работы были выполнены поставленные задачи, и в результате работы был создан сервис, соответствующий поставленным требованиям. Был проведен необходимый анализ и эстимация. Поставленная цель была полностью выполнена.

Созданный бот можно найти по адресу *https://t.me/IISIGT_bot* или по логину *@IISIGT_bot* внутри клиента Telegram.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Attention Is All You Need / A. Vaswani [и др.]. — 2017. — URL: https://arxiv.org/pdf/1706.03762.

2 *Jaekel B.* H&M, Sephora chatbots gain visibility in Kik's new marketplace. — 2015. — URL: https://www.marketingdive.com/ex/mobilemarketer/cms/news/messaging/22588.html.

3 *Jurafsky D.*, *Martin J. H.* Speech and Language Processing. — 2000.

4 *Mentor O.* SRP: The Single Responsibility Principle [Электронный ресурс]. — Архивировано 2 февраля 2015. — URL: https://web.archive.org/web/20150202200348/http://www.objectmentor.com/resources/articles/srp.pdf.

5 Исследование VK Мессенджера: три четверти россиян активно используют чат-боты. — 2023. — URL: https://vk.com/press/messenger-bots-research.

6 *Левенштейн В. И.* Двоичные коды с исправлением выпадений, вставок и замещений символов. Доклады Академий Наук СССР. — 1965. — URL: https://www.mathnet.ru/links/575633ee4c33f4301f89032ac23cc9e1/dan31411.pdf.

ПРИЛОЖЕНИЕ А
Диаграмма последовательности

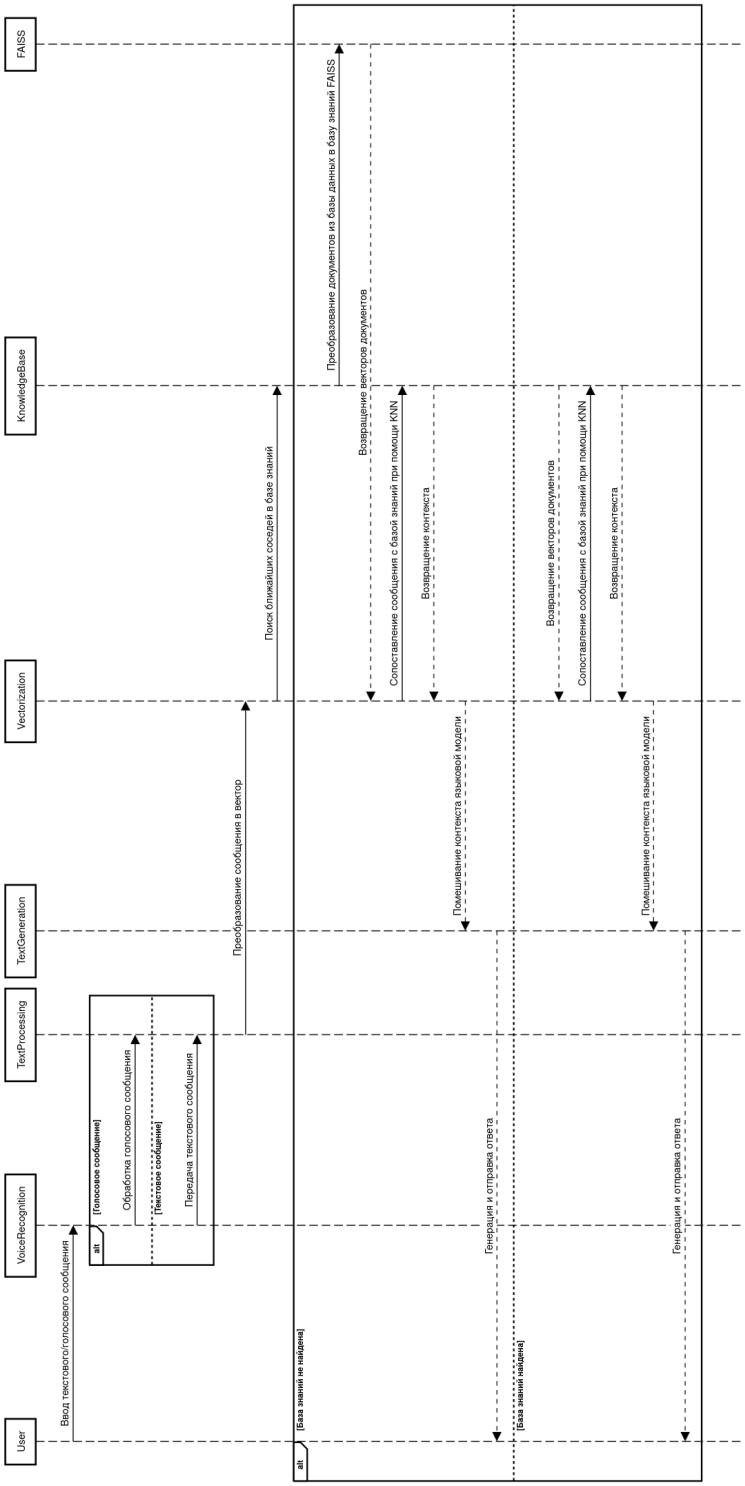

Рисунок А.1 — Диаграмма последовательности

ПРИЛОЖЕНИЕ Б
Конфигурация NeoVim

Листинг Б.1 — Отрывок конфигурации языкового сервера для работы NeoVim

```
local lspconfig = require 'lspconfig'

local capabilities =
 ↪  require('cmp_nvim_lsp').default_capabilities()

lspconfig.pylsp.setup {
  capabilities = capabilities,
  settings = { pylsp = { plugins =
   ↪  require('project.config').pylsp_plugins } },
}
lspconfig.tsserver.setup {
  capabilities = capabilities,
}
lspconfig.ccls.setup {
  capabilities = capabilities,
}
lspconfig.gopls.setup {
  capabilities = capabilities
}

-- `:help vim.diagnostic.*`
vim.keymap.set('n', '<space>e', vim.diagnostic.open_float)
```