

Министерство науки и высшего образования Российской Федерации
Федеральное 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	3
1 Предпроектный анализ	5
1.1 Анализ предметной области	5
1.2 Сравнительный анализ	6
1.3 Системный анализ	6
1.4 Требования к сервису	8
1.5 Сроки реализации проекта	9
2 Проектирование информационной системы	11
2.1 Концептуальное проектирование	11
2.2 Диаграмма компонентов	13
2.3 Диаграмма развертывания	13
2.4 Диаграмма последовательности	14
2.5 Схема базы данных	15
2.6 Развертывание приложения	15
3 Разработка системы	17
3.1 Выбор средств разработки	17
3.2 Расчет фактических затрат на реализацию	17
3.3 Расчет ожидаемого результата экономической эффективности	17
3.4 Расчет надежности программного и аппаратного обеспечения	17
3.5 Что-то про структуру приложения	17
3.6 Что-то про извлечение сущности из фраз пользователя и их метчинг с сущностями из базы знаний	17
3.7 Что-то про метод ближайших соседей и обогащение ответа . .	17
3.8 Что-то про генерацию ответа	17
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	18

ВВЕДЕНИЕ

За последние десятилетия произошел огромный скачок в развитии информационных технологий: от создания первой электронной вычислительной машины, до сложных 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 нейронной 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 обусловлена возможностью оптимизации многих процессов, увеличении производительности сотрудников, повышении лояльности клиентов. В контексте высшего учебного заведения (ВУЗ), использование технологий подобного рода повышает конкурентоспособность 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 приложения для коммуникации.

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.

Для выполнения поставленной цели были поставлены следующие зада-

чи:

- Выполнить анализ предметной области;
- Провести сравнительный анализ информационных систем;
- Изучить 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 3.13, LangChain, FAISS, HuggingFace.

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 преимуществ имеется возможность настройки внешнего вида бота.

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 интуитивно 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 поступлении 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.

Таблица 1.2 — SWOT анализ

	Положительное влияние	Негативное влияние
Внутренняя среда	<ul style="list-style-type: none"> — Актуальность — Простота использования 	<ul style="list-style-type: none"> — Нейросетевые галлюцинации — Необходимость тщательно прорабатывать интеграцию во избежание проблем с безопасностью
Внешняя среда	<ul style="list-style-type: none"> — Упрощение навигации по ресурсам ВУЗа — Получение поддержки от государства 	<ul style="list-style-type: none"> — Регуляции со стороны государства — Бюрократия с какой-либо стороны

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

Из отрицательных факторов стоит выделить следующие аспекты: нейросетевые галлюцинации, проблемы с безопасностью и бюрократия.

Нейросетевые галлюцинации — аномалия, возникающая во время работы нейронной сети, влияющая на вывод результат непредсказуемым образом. Например, спросив о технической оснащённости 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 безопасностью 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.

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 следующих таблицах описаны нефункциональные требования.

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 следующей таблице описаны требования к надежности 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 следующей таблице описаны требования к производительности 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.

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

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 позволяют реализовать корректную работу приложения.

1.5 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	ВРМН	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.

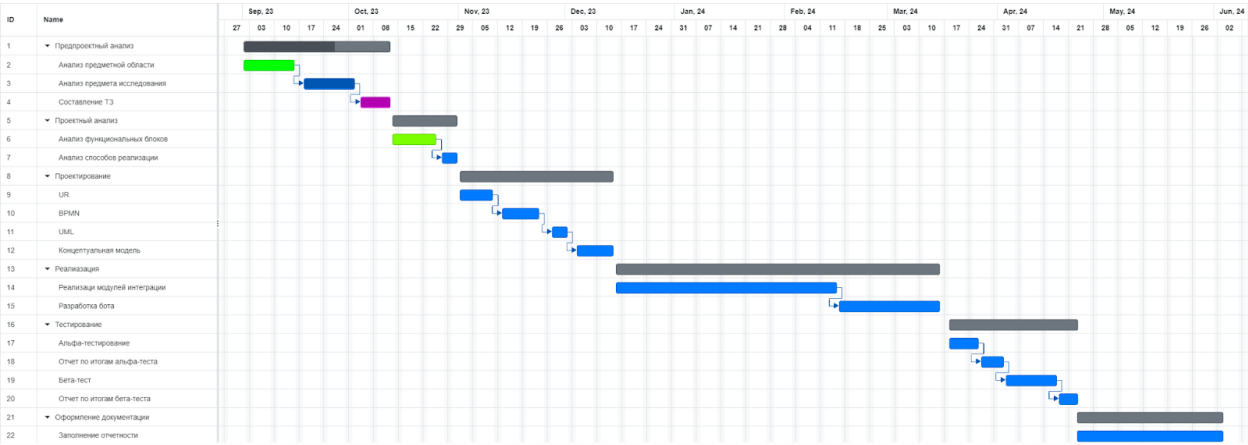


Рисунок 1.2 — Диаграмма Ганта

2 Проектирование информационной системы

2.1 Концептуальное проектирование

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

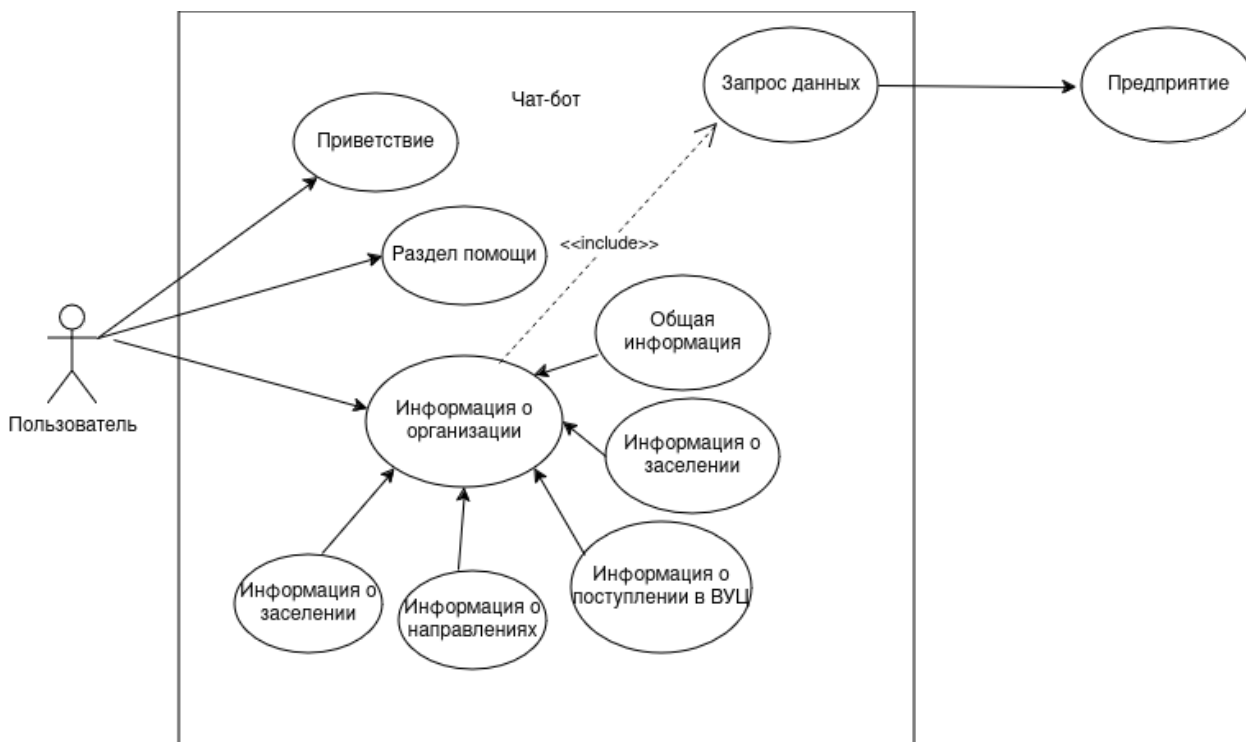


Рисунок 2.1 — Диаграмма вариантов использования

Актор *Пользователь* является обобщением клиентов, которые используют 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* используется для логического объединения других прецедентов, в цели уменьшения количества связей на диаграмме и облегчить ее восприятие.



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.

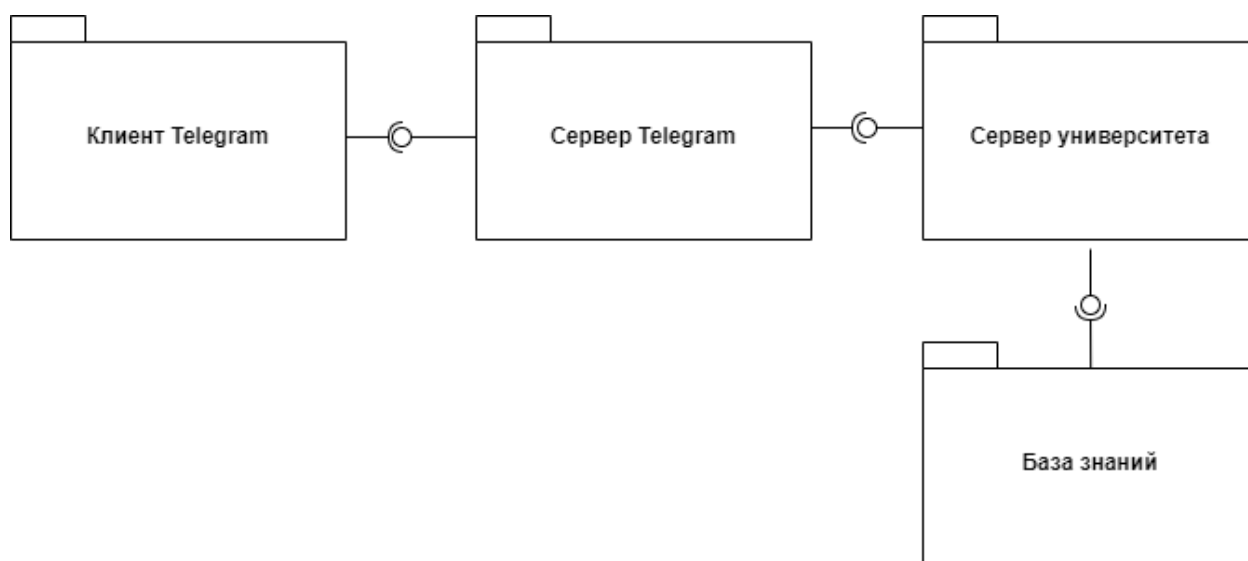


Рисунок 2.3 — Диаграмма компонентов

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.

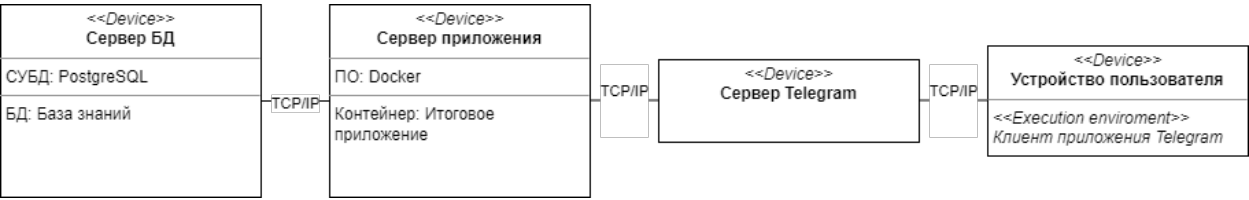


Рисунок 2.4 — Диаграмма развертывания

2.4 Диаграмма последовательности

Последовательность работы приложения показано на рисунке 2.5

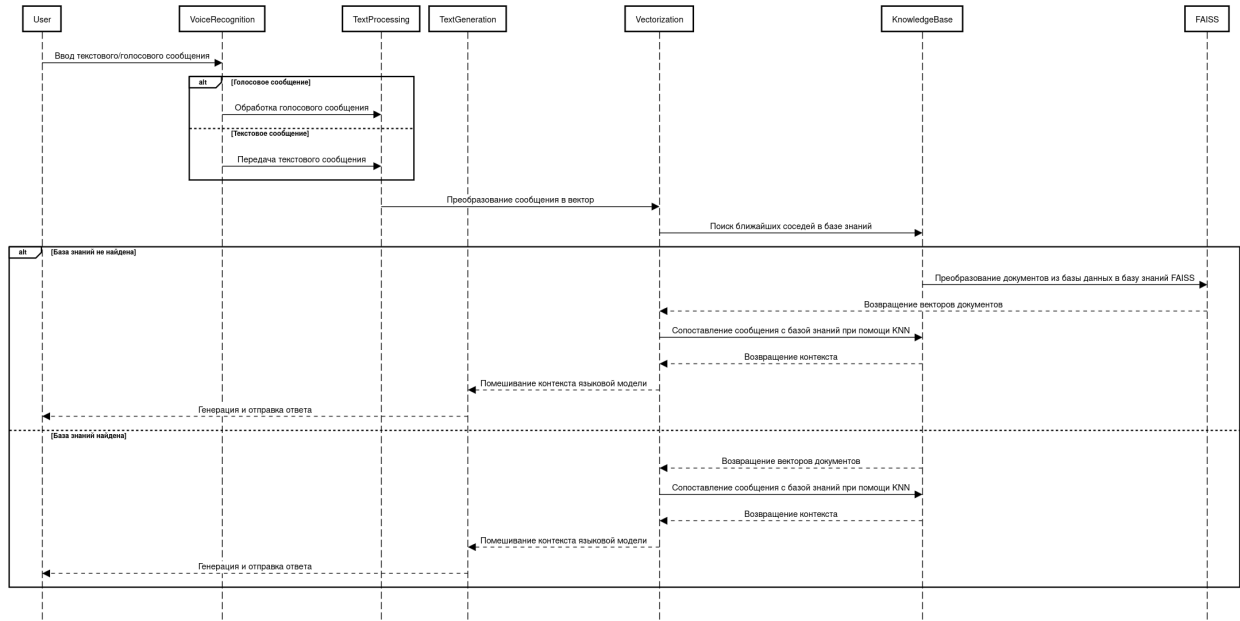


Рисунок 2.5 — Диаграмма последовательности

Пользователь вводит запрос в текстовом формате или формате аудио. Если сообщение передано в аудиоформате, то происходит расшифровка аудиозаписи. Далее полученный текст преобразуется в векторное представление, после чего 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.

2.5 Схема базы данных

В качестве базы данных для хранения изначальных документов для построения retrieval 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 2.6

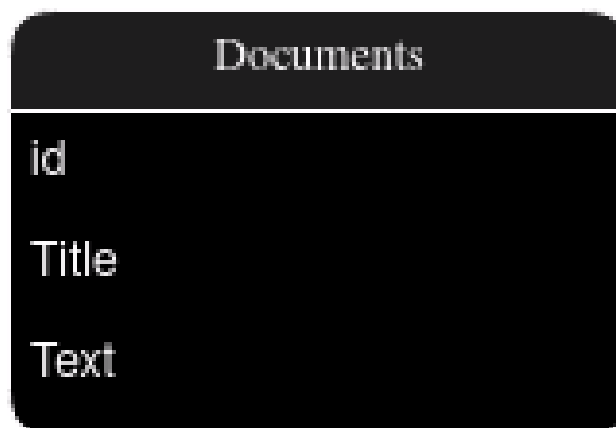


Рисунок 2.6 — Схема хранения документов

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.

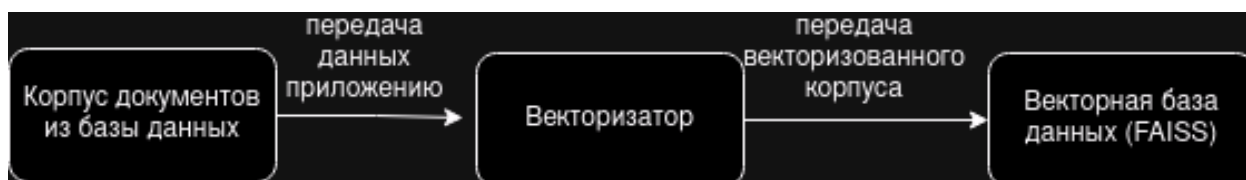
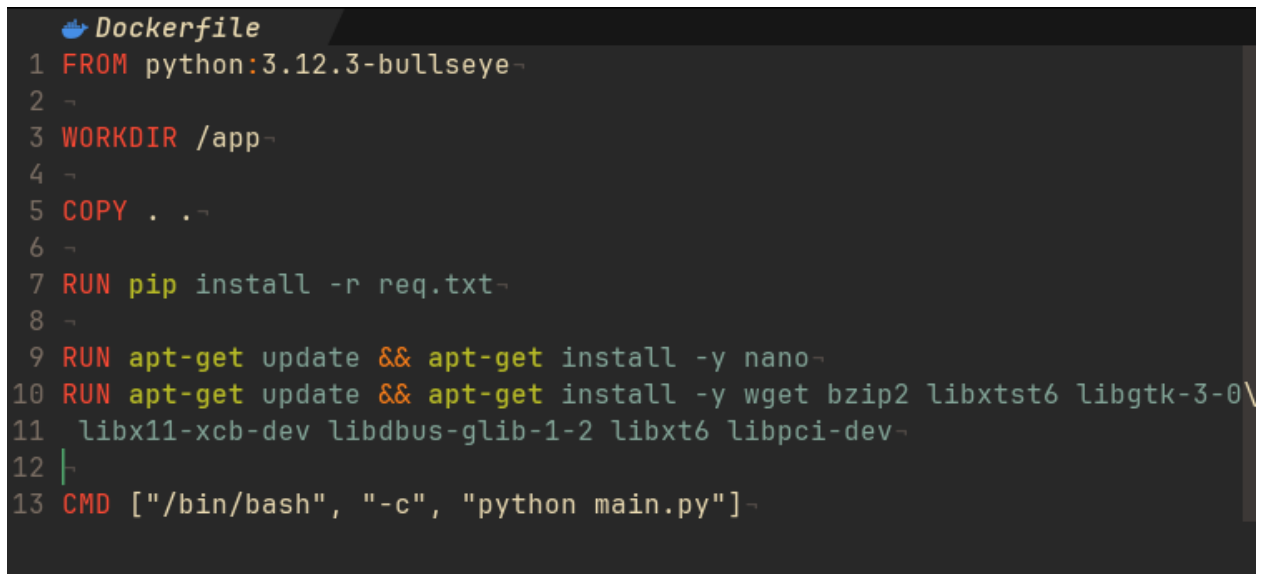


Рисунок 2.7 — Схема векторизации документов

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 программным обеспечением, что позволяет использовать его для любых нужд.



```
Dockerfile
1 FROM python:3.12.3-bullseye
2
3 WORKDIR /app
4
5 COPY . .
6
7 RUN pip install -r 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"]
```

Рисунок 2.8 — Файл контейнеризации

Данный файл позволяет создать контейнер, содержащий все 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 частных репозиториях — никто, кроме авторизованных 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.

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

3.2 Расчет фактических затрат на реализацию

3.3 Расчет ожидаемого результата экономической эффективности

3.4 Расчет надежности программного и аппаратного обеспечения

3.5 Что-то про структуру приложения

3.6 Что-то про извлечение сущности из фраз пользователя и их метчинг с сущностями из базы знаний

3.7 Что-то про метод ближайших соседей и обогащение ответа

3.8 Что-то про генерацию ответа

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ