Н И Т У « М И С И С »

Кафедра инженерной кибернетики

Учебная дисциплина
«**Методы искусственного интеллекта**»

ЛАБОРАТОРНАЯ РАБОТА №4

«Разработка асинхронного чат-бота для мессенджера Telegram с использованием языковых моделей архитектур GPT и LLaMA»

бакалавриат по направлению *01.03.04 прикладная математика*
(семестр VII; 2024÷2025 уч. год)

2024 г.

Содержание

Введение

Настоящий документ содержит задание для **лабораторной работы №4** (далее – **л/р №4**) по дисциплине «**Методы искусственного интеллекта**» для учащихся бакалавриата по направлению подготовки 01.03.04 *прикладная математика* кафедры инженерной кибернетики НИТУ «МИСИС». Основным направлением работ учащихся при выполнении **л/р №4** являются процессы и технологии разработки асинхронных чат-ботов, использующих для реализации основных функций общения большие языковые модели архитектур **GPT и LLaMA.**

Чат-бот (англ. *chatbot*; также - виртуальный собеседник, программа-собеседник) – это программа для мобильного устройства и/или настольного ПК, которая выясняет потребности пользователей, а затем помогает удовлетворить их. В чат-ботах реализуется имитация общения с пользователя с «живым» сотрудником некоторой организации. Такое общение ведется с помощью текста или голоса и используются такие программы как альтернатива переписке с живым оператором или звонку сотруднику организации.

Чат-боты упрощают взаимодействие между пользователем некоторого интернет-сервиса или мессенджера (например, Telegram) и организацией. Они помогают экономить средства. Быстрый поиск ответа на вопрос, смена персональных данных, устранение мелких неполадок в приложении, оформление заявки на покупку продукта – с этим чат-бот вполне может справиться самостоятельно. При необходимости он перенаправляет запрос вместе с полученными данными на нужного специалиста.

Студентам нужно сравнить два семейства языковых моделей: **GPT**, **LLaMa**. Провести сравнение: синхронного подхода и асинхронного подхода при проектировании, архитектуры, обучения, цели использования, преимуществ и недостатков каждой модели. В конечном итоге: сформировать вывод, пощупать модели и развернуть у себя.

Режим выполнения учащимися **л/р №4** (индивидуальный, командный) определяет преподаватель. Выбор средств разработки должен быть согласован с преподавателем, ведущим лабораторные работы по учебной дисциплине.

Каждый учащийся при выполнении лабораторной работы обязан соблюдать интеллектуальные, авторские и смежные права и лицензионные соглашения соответствующих правообладателей при использовании в процессах разработки своего программного обеспечения любых выбранных технических средств и инструментов, а также научного инструментария.

1. Задание на лабораторную работу

1.1. Основная цель работы

Основная цель лабораторной работы - выработать у учащихся устойчивые начальные навыки владения процессом создания асинхронных чат-ботов для мессенджера *Telegram*© и использования в структуре чат-бота для реализации основных функций общения больших языковых моделей классов (архитектур) **GPT** и **LLaMA**.

1.2. Общие сведения

Чат-бот — это компьютерная программа, которая относится к категории интеллектуальных систем, и представляет собой электронного (виртуального) «собеседника», созданного для решения типовых задач: поиск информации, ответы на вопросы клиентов-посетителей различных интернет-сервисов, маркетплейсов т.п. электронных площадок, выполнение простых процедур и действий в самых разных сферах, например:

- банковская сфера – консультация, техническая поддержка, а также выполнение стандартных процедур (например, открытие банковского вклада) и простых финансовых транзакций;
- операторы связи – техническая поддержка;
- страхование – помощь в заполнении форм и заявок, консультация по страхованию;
- онлайн-торговля (маркетплейсы) – консультирование по доставке, оплате, адресам точек выдачи;
- здравоохранение – предоставление медицинских материалов, первичная консультация и сбор анамнеза;
- туризм – агрегатор предложений, рассылка горящих туров, бронирование;
- образование – рассылки, приглашение на вебинары;
- государственные услуги – сбор жалоб, быстрый доступ к публичным данным;

– HR-службы – подбор подходящих резюме, автоматизация задач внутри компании.

Задачи, которые могут решать чат-боты, также могут быть самыми разными:

– распознавание и интерпретация запросов,
– консультация по вопросам банковского обслуживания и совершение финансовых транзакций,
– бронирование столиков в ресторанах,
– запись на посещение в разных организациях (медицинские услуги, салоны красоты, фитнес-центры и т.п.);
– решение стандартных задач техподдержки компьютерного и сетевого оборудования (например, при взаимодействии с провайдером)
– и многое другое,

Существуют разнообразные виды чат-ботов, которые можно классифицировать: по платформе внедрения, технологии разработки, способам организации процессов общения с пользователями (синхронные и асинхронные) и по различным аспектам функциональности.

Основное отличие синхронного чат-бота от асинхронного заключается в том, что в **синхронных** – каждый запрос от пользователя блокирует выполнение всего чат-бота до момента выдачи ответа на запрос, а в **асинхронных** такой блокировки нет и чат-бот может обрабатывать несколько запросов параллельно.

Асинхронность является предпочтительным режимом работы, т.к. повышает производительность и улучшает реактивность бота, чтобы он мог эффективнее справляться с различными задачами. Однако реализации асинхронности требует более высокой квалификации от разработчика и больших затрат при создании системы.

В **л/р №4** учащийся должен освоить технологию создания асинхронных чат-ботов с использованием специализированного инструментария (библиотек) – **python-telegram-bot (с модулем asyncio), aiogram, AsyncTeleBot.**

1.3. Языковые модели, рекомендуемые к использованию

В **л/р №4** учащимся рекомендуется использовать языковые модели следующих архитектур (семейств) из списка:

1) Архитектура LLaMa.

а) **Модель LLaMA-2-7B** – это улучшенная версия оригинальной LLaMA-7B, с улучшениями в архитектуре и обучении. Она содержит 7 миллиардов параметров, но имеет более высокую производительность благодаря оптимизациям.

Ссылка: https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

б) **Модель LLaMA-13B** – эта модель несколько больше, чем предыдущие, она всё ещё считается относительно небольшой по сравнению с самыми крупными моделями. Содержит 13 миллиардов параметров и предлагает более высокую точность и способность справляться с более сложными задачами.

Ссылка: https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

в) **LLaMA** -модель с поддержкой русского языка.

Ссылка: https://huggingface.co/IlyaGusev/saiga_llama3_8b

2) **Архитектура GPT**.

а) **Модель GPT2** – это самая маленькая версия GPT-2, обеспечивающая хорошую производительность для многих задач генерации текста с относительно низкими вычислительными требованиями.

Ссылка: https://huggingface.co/openai-community/gpt2

b) **Модель GPT-Neo** – это открытая альтернатива GPT-3 от EleutherAI. Версия с 125 миллионами параметров предназначена для задач, где важны низкие вычислительные затраты и хорошее качество генерации текста.

Ссылка: https://huggingface.co/EleutherAI/gpt-neo-125m

в) GPT-модель с поддержкой русского языка.

Ссылка: https://huggingface.co/ai-forever/rugpt3medium_based_on_gpt2

2. Выполнение лабораторной работы

Основная задача. Учащемуся необходимо создать асинхронный чат-бот для мессенджера Telegram. В функциональную структуру чат-бота должны быть интегрированы две или четыре модели (по одной из каждого класса (архитектуры) или по две из каждого класса (архитектуры) в зависимости от количества человек в команде). После завершения отладки работы чат-бота учащийся должен провести сравнительный анализ выбранных языковых моделей разных классов.

Основные этапы выполнения лабораторной работы учащимся следующие:

1) Выбор библиотеки для создания чат-бота.
2) Выбор предметной области и определение основного назначения чат-бота (Разрешается взять предметную область из **лабораторной работы №3** без согласования).
3) Проектирование чат-бота: назначение, функциональность, используемые NLP-инструментарий.
4) Непосредственная разработка асинхронного чат-бота с глубиной вложенности не менее 5 уровней и проверка его работоспособности.
5) Сравнительный анализ особенностей, преимуществ и недостатков каждой из использованных больших языковых моделей.
6) Подготовка доклада по результатам работы (до 10 минут).
7) Защита результатов работы преподавателю.

Работа выполняется индивидуально.

Внимание! Допускается разработка чат-бота в команде по согласованию с преподавателем. При этом глубина вложенности увеличивается пропорционально числу участников команды.

Дополнительная информация.

Для начального знакомства с особенностями асинхронного программирования на языке Python рекомендуется ознакомится с информацией по ссылке: https://school.kontur.ru/publications/2567

3. Защита лабораторной работы

Выполненная учащимся (группой учащихся) лабораторная работа проходит процедуру защиты. При проведении защиты лабораторной работы учащийся (группа учащихся):

- демонстрирует работоспособность созданного программного обеспечения чат-бота и безошибочность реализации заявленной функциональности (показывает решение задачи с помощью чат-бота) с учетом всех замечаний по реализации ПО с предыдущих лабораторных работ.
- Разработанное ПО должно соответствовать следующим требованиям:
    - отсутствие неявной типизации;
    - наличие модульности в проекте;
    - наличие правильно оформленных комментариев к функциям и классам;
    - наличие аннотаций типов данных;
    - соответствие исходного кода стандарту PEP8;
    - отсутствие явного указания чувствительных данных в исходном коде;
- рассказывает про:
    - решаемую задачу,
    - выбранные и реализованные технологии в чат-боте,
    - проблемы, которые возникли с при работе с языковыми моделями,
    - использованные средства программирования для создания чат-бота (назначения, возможности, ограничения и проч.),
- представляет подготовленный отчет по **л/р №4**.

В процессе и (или) после демонстрации преподаватель задает учащемуся (членам команды) вопросы по различным аспектам выполнения **л/р №4** и полученным в ходе её выполнения результата. В случае наличия существенных замечаний и (или) выявления ошибок в работе созданного программного обеспечения и тексте отчета защита **л/р №4** прекращается до полного исправления выявленных ошибок и устранения сделанных замечаний.

После того как преподавателем будет окончательно принято созданное в **л/р №4** программное обеспечение (чат-бот), результаты его работы и отчет, учащийся обязан сдать[1] преподавателю оформленный отчет по **л/р №4**.

Отчет должен содержать: титульный лист, описание решаемой задачи с помощью телеграм бота, описание архитектуры моделей, демонстрация вложенности, сравнительный анализ моделей, скриншоты работы бота, вывод. Форма и режим сдачи отчетных материалов объявляются преподавателем отдельно.

[1] По решению преподавателя отчетные материалы могут либо сдаваться в распечатанном виде (на твердом носителе) либо загружаются учащимся в электронном виде в LMS

4. Требования к содержанию и оформлению отчета

1) Общие требования.

Язык отчета – русский.

Текст отчета должен быть проверен на наличие и не должен содержать орфографических и синтаксических ошибок.

2) Требования к содержанию отчета.

Отчет по лабораторной работе оформляется в соответствие со следующими требованиями к содержанию.

- **Титульный лист** (оформляется в соответствии с **приложением**).
- **Введение**

Описываются: основная цель и задачи лабораторной работы.

- **Краткое (не более 2 стр.) описание решаемой задачи по интеллектуальной обработки и/или анализу информации**

Описываются: основное содержание решаемой задачи, что является исходными данными (с примерами), что является результатом решения задачи и дополнительные сведения, раскрывающие особенности как самой задачи, так и процесса её решения (при необходимости).

- **Выбранные средства для разработки программного обеспечения**

Описываются: используемые ИИ-сервис/систем и конкретный ИИ-функционал (модули; подсистемы; библиотеки; платформы; API и т.п.), а также использованные для создания программного обеспечения языки программирования, сторонние библиотеки и фреймворки, среды разработки (IDE), СУБД (если используется) и т.п.

- **Описание и примеры исходных данных**

Учащийся должен указать фразы текста на естественном языке, которые были использованы в качестве стартовых фраз для модели.

- **Результаты**

Иллюстрированное описание процесса решения задачи с использованием созданного программного обеспечения. Описание выполняется в форме последовательности скриншотов экранных форм с сопроводительным краткими текстовыми комментариями, иллюстрирующих основной процесс работы созданного приложения при решении задачи для заданного числа примеров исходных данных.

- **Анализ полученных результатов работы с NLP-моделями GPT и LLaMA**

В этой части отчета учащийся (команда учащихся) должен привести развернутый обоснованный анализ, содержащий наиболее значимые выводы и оценку уровня качества генерации русскоязычных текстов выбранными NLP-моделями «семейства» **GPT и LLaMA**, заданной тематической направленности на примере выполненного задания **л/р**;

- **Выводы по лабораторной работе**.
- **Список использованных источников**
- **Приложения (при необходимости)**

2) Требования к оформлению отчета.

Текст отчета, таблицы, графики и диаграммы, формулы, другие специальные обозначения должны быть отредактированы, а также отформатированы единым образом. Сокращения слов, за исключением общепринятых, не допускаются.

Параметры страницы.

−Размер листа – формат «А4»;

−Поля: левое: 25 мм, правое: 10 мм, верх и низ: 15 мм;

−Печать текста на листе – односторонняя.

Форматирование текста.

−Шрифты:

- основной текст – Times New Roman Cyr, 12 пт.;

- названия глав (разделов) отчета – Times New Roman Cyr, 14 пт., набор прописными буквами с полужирным выделением (bold) и выравниванием по центру страницы;
- заголовки пунктов и подпунктов – Times New Roman Cyr, 14 пт. с полужирным выделением (bold);
- подписи рисунков – Times New Roman Cyr, 11 пт.

– Абзацы:

- первая строка – отступ: 1,25 см;
- межстрочный интервал – 1,5;
- выравнивание абзаца – по ширине;
- интервал после абзаца обычного текста – 6 пт.;
- интервал до и после заголовков пунктов и подпунктов – 12 пт.;

– Нумерация страниц: внизу по центру; титульный лист является первой страницей отчета и не нумеруется.

– Нумерация рисунков и таблиц – сквозная целочисленная. Если таблица и/или рисунок в отчете единственные, то тогда они не нумеруются.

– Все таблицы и рисунки должны иметь название.

– Все графики и графические изображения, отражающие какие-либо функциональные зависимости, должны содержать подписи по всем обозначенным осям (абсцисс, ординат, аппликат), включающие в себя: название; единицу измерения соответствующей величины; числовые отметки на оси, характеризующие масштаб отображаемой величины.

– Если рисунок или таблица размещается на странице (страницах), для которой задана «альбомная» ориентация листа, то их размещение должно быть таким, чтобы правильное расположение объекта относительно читателя достигалось поворотом страницы из нормального положения документа на 90° по часовой стрелке.

– Размещаемые в отчете рисунки не должны вызывать затруднений для рассмотрения, а текстовая информация на рисунках должна быть читабельной.

Приложение. Титульный лист отчета по лабораторной работе

Н И Т У « М И С И С »

Кафедра инженерной кибернетики

ОТЧЕТ

по

ЛАБОРАТОРНОЙ РАБОТЕ №4

«Разработка асинхронного чат-бота для мессенджера Telegram с использованием языковых моделей архитектур GPT и LLaMA»

учебная дисциплина
«Методы искусственного интеллекта»

Студент: *ФИО и группа*

Преподаватель: Хонер П.Д.

Оценка:

Дата защиты:

2024 г.