

**Федеральное государственное автономное образовательное учреждение
высшего образования
«РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ»
(РУДН)**

Основное учебное подразделение ФФМиЕН

Направление/специальность 02.03.01 «математика и компьютерные науки»

ОТЧЕТ

о прохождении учебной практики

«Улучшение предсказания prompts с помощью ActasGPT»

Яссин Мохамад Аламин

Курс, группа 3 курс, группа НКНбд-01-20

Место прохождения практики «ОТПП департамента технологических и информационных ресурсов РУДН»

Сроки прохождения с «17» апреля 2023 г. по «17» июня 2023 г.

Руководители практики:

от РУДН Фомин М.Б., к.ф.-м.н, доцент, доцент
кафедры информационных технологий

от организации (предприятия)
Салпагаров С.И., к.ф.-м.н., доцент

Оценка _____

Москва 2023 г.

Содержение

Акронимы	3
Введение.....	4
Идея проетка.....	5
Структура проекта	6
🌀 App.py	6
🌀 Utils.py	6
🌀 Папка Model:	6
Technologies used in the project:	8
Результаты:	10
Улучшенные результаты:	11
Заключение	14
Источники	15

Акронимы

AI – Искусственный интеллект

API – Интерфейс прикладного программирования

AWS – Веб-сервис Amazon.

DL – Глубокое обучение

GPT – Генеративный трансформатор с предварительной подготовкой

ML – Машинное обучение

NLP – Обработка естественного языка

Введение

В последние годы в области обработки естественного языка (NLP) и машинного обучения произошел значительный прогресс, что привело к новаторским достижениям и открыло широкий спектр возможностей в области приложений для генерации текста. Эта стремительная эволюция значительно изменила отрасли, внедрив инновационные решения и революционизировав способы нашего взаимодействия с технологиями. Способность генерировать связный и контекстуально подходящий текст стала важнейшим требованием в различных секторах, включая коммуникации, создание контента, обслуживание клиентов и информационные технологии.

Одной из конкретных разработок, которая привлекла внимание и интерес как профессионалов, так и энтузиастов, является ChatGPT. Как продвинутая языковая модель, ChatGPT использует самые современные методы NLP и машинного обучения, чтобы генерировать ответы, подобные человеческим, и участвовать в содержательных разговорах. Его способность понимать и генерировать текст делает его универсальным инструментом с приложениями, охватывающими широкий спектр областей.

Среди разнообразных приложений ChatGPT отлично справляется с такими задачами, как обобщение статей, создание статей, внедрение кода, анализ текста и выполнение домашних заданий. Этот многогранный инструмент стал незаменимым ресурсом для людей, которые ищут эффективные и действенные решения сложных проблем. Доступность таких расширенных возможностей генерации текста произвела революцию в различных отраслях, позволив профессионалам оптимизировать свои рабочие процессы, повысить производительность и получать высококачественные результаты.

Несмотря на огромные преимущества, предлагаемые ChatGPT и аналогичными языковыми моделями, пользователи часто сталкиваются с трудностями при использовании своего полного потенциала. Одним из конкретных препятствий, с которыми сталкиваются пользователи, является выбор оптимальной подсказки или запроса, который дает желаемые результаты и максимизирует полезность этих языковых моделей. Эффективность подсказки играет решающую роль в формировании вывода, генерируемого моделью, поскольку она обеспечивает исходный контекст и руководство для процесса генерации текста.

Признавая необходимость решения этой проблемы и оптимизации взаимодействия с пользователем, ActasGPT становится инновационным проектом, находящимся в авангарде исследований и разработок. ActasGPT стремится предоставить комплексное решение, которое позволяет пользователям использовать все возможности больших языковых моделей, позволяя им последовательно и эффективно получать желаемые результаты.

Идея проекта

ActasGPT фокусируется на совершенствовании и улучшении процесса выбора подсказок за счет использования передовых алгоритмов и методов. Благодаря интуитивно понятному и удобному интерфейсу ActasGPT помогает пользователям формулировать prompts, которые являются контекстуально релевантными, конкретными и согласуются с их целями. Понимая нюансы prompt формулирования и учитывая тонкости языковых моделей, ActasGPT позволяет пользователям получать наиболее точные, проницательные и связные ответы от этих мощных систем генерации текста.

Проект основан на обширных исследованиях и использует самые современные методологии, чтобы предоставить пользователям беспрепятственный опыт взаимодействия с языковыми моделями. ActasGPT создает передовую платформу, которая максимально использует потенциал технологии генерации текста.

С ActasGPT пользователи получают доступ к широкому спектру функций, предназначенных для облегчения prompt формулирования и оптимизации вывода языковых моделей. Платформа включает в себя методы понимания естественного языка, помогающие пользователям генерировать prompts, которые эффективно передают их намерения. Он анализирует структуру приглашения, семантическое значение и желаемые результаты, предлагая предложения и идеи для уточнения и улучшения первоначального запроса пользователя.

Кроме того, ActasGPT использует расширенный контекстный анализ для дальнейшего улучшения генерации prompts. Принимая во внимание контекст и желаемую задачу, ActasGPT помогает пользователям создавать prompts, соответствующие их конкретным целям, и получать ответы, адаптированные к их потребностям. Эта контекстуальная осведомленность гарантирует, что сгенерированный текст будет последовательным, точным и контекстуально подходящим, что позволит пользователям последовательно достигать желаемых результатов.

ActasGPT также предоставляет пользователям всестороннюю аналитику и механизмы обратной связи для оценки эффективности их prompts и оптимизации их взаимодействия с языковыми моделями. Платформа использует GPT-API для создания выходных данных нашей подсказки, предоставляя ценную информацию о качестве, актуальности и согласованности ответов. Этот механизм обратной связи позволяет пользователям повторять и совершенствовать свои prompts, постоянно улучшая их взаимодействие с языковыми моделями и улучшая общее взаимодействие с пользователем.

Структура проекта

Репозиторий ActasGPT тщательно организован для обеспечения эффективного управления файлами и ресурсами. Давайте углубимся в структуру репозитория и изучим содержимое каждой папки и файла:

- 🌀 App.py
- 🌀 Utils.py
- 🌀 Model (папка)
 - 🌀 Prompts.csv
 - 🌀 Chatgpt-prompts (папка)
 - 🌀 Training.ipynb
 - 🌀 My_project.ipynb
 - 🌀 Vectorizer.pkl

- **App.py**

Файл app.py является основой интерфейса веб-сайта ActasGPT. Он служит основным сценарием, который управляет функциональностью веб-сайта, включая взаимодействие с пользователем и генерацию prompts. В этом файле реализованы различные функции и алгоритмы для обеспечения плавного и беспрепятственного взаимодействия с пользователем. Файл app.py действует как мост между пользовательским интерфейсом и базовыми внутренними процессами, позволяя пользователям вводить подсказки и получать точные прогнозы.

- **Utils.py**












Файл «utils.py» содержит важные бэкенд функции для веб-сайта ActasGPT. Он включает в себя различные функции и вызовы API, поддерживающие процесс генерации prompts. Этот файл служит центральным узлом для всех внутренних операций, позволяя интерфейсу веб-сайта беспрепятственно взаимодействовать с необходимыми компонентами и API. Файл «utils.py» имеет решающее значение для обеспечения функциональности веб-сайта, позволяя prompt прогнозировать и обрабатывать связь между пользовательским интерфейсом и базовой моделью.

- **Папка Model:**

Папка «модель» содержит важные компоненты проекта ActasGPT, в том числе модель transformers, набор данных, используемый для обучения, и ipynb файлы, использованные при обучении как для последней, так и для начальной версии программы. Вот разбивка содержимого папки «model»:

- **Prompts.csv**
Набор данных, используемый для обучения модели ActasGPT, является важным компонентом проекта. Он состоит из двух столбцов: входного столбца, обозначенного как «действовать как», и выходного столбца. Этот набор данных служит основой для обучения модели, обеспечивая ее разнообразными prompts и соответствующими текстовыми выводами для обучения.
- **Training.ipynb**
Основной частью ActasGPT является модель transformers, которая играет ключевую роль в точном прогнозировании подсказок. Эта записная книжка подробно документирует процедуру обучения, описывая точные шаги, предпринятые для точной настройки модели transformers с использованием предоставленного набора данных. Тщательно обучая модель трансформатора с помощью набора данных, доступного в репозитории, она становится способной генерировать точные прогнозы в ответ на вводимые пользователем данные. Следовательно, эта книжка служит ценным справочником для понимания эволюции модели и может быть использована для будущих улучшений или корректировок.
- **My_project.ipynb**
В этой записной книжке я провел визуализацию данных, используя различные методы, такие как wordcloud. Кроме того, я применил методы NLP для предварительной обработки данных и использовал косинусное сходство (cosine similarity) от библиотеки scikit-learn, чтобы предсказать оптимальную подсказку вывода, которая согласуется с вводом пользователя. Примечательно, что именно этот подход был использован в начальной версии программы.
- **Vectorizer.pkl**
Файл с именем «Vectorizer.pkl» представляет векторный вывод, полученный путем применения косинусного подобия (cosine similarity) к нашему набору данных. В начальной версии программы этот файл использовался путем импорта его непосредственно в программу Python для немедленного использования в нашем приложении.
- **Chatgpt-prompts folder:**
В этой папке содержатся модель, конфигурация и токенизатор, относящиеся к модели transformers. Эти файлы представляют собой результаты процесса тонкой настройки и впоследствии были импортированы для недавнего использования в нашей программе.

Технологии, используемые в проекте:

-  Pandas
-  Worldcloud
-  NLP
-  Модель
 -  cosine similarity
 -  transformers
-  Prompt-engineering
 -  GPT-API
-  Питон
-  Streamlit
-  Облако (AWS)

Для достижения своих целей ActasGPT использует множество передовых технологий. wordcloud позволяет визуализировать данные. Методы NLP используются для эффективной предварительной обработки данных, обеспечивая правильную обработку входных запросов перед обучением модели. Процесс построения модели включает в себя точную настройку преобразователей и использование сходства косинусов Склеарна(cosine similarity), предлагая оптимизированный подход к prompt прогнозированию. Кроме того, ActasGPT интегрирует GPT-API и использует Streamlit для интуитивно понятного пользовательского интерфейса. Наконец, Amazon EC2 предоставляет надежный и масштабируемый облачный хостинг, соответствующий вычислительным требованиям ActasGPT.

- Pandas
Pandas — мощная библиотека для работы с данными на Python, сыгравшая решающую роль в проекте. В основном он использовался для чтения набора данных в виде фрейма данных, что позволяло эффективно обрабатывать и подготавливать данные. С помощью Pandas набор данных был организован и структурирован в табличном формате, что сделало его легко доступным для последующей обработки и анализа. Возможность чтения и обработки данных с помощью Pandas упростила интеграцию набора данных в pipeline проекта.

- Wordcloud

Wordcloud — это метод визуализации данных, используемый в проекте. Он был использован для визуального представления текстовых данных в краткой и визуально привлекательной форме. Создавая облака слов, проект стремился обеспечить четкое представление наиболее часто встречающихся слов или фраз в наборе данных. Этот метод не только улучшил понимание набора данных, но также помог выявить закономерности и тенденции в тексте, способствуя более точному прогнозированию.

- NLP

НЛП — это область исследования, которая фокусируется на взаимодействии между компьютерами и человеческим языком. В проекте использовались методы НЛП для эффективной обработки и анализа текстовых данных. Это включало различные задачи, такие как предварительная обработка текста, токенизация и извлечение признаков. Используя NLP, проект смог преобразовать необработанные текстовые данные в формат, который легко понять и использовать алгоритмы машинного обучения. НЛП сыграло решающую роль в повышении точности и эффективности быстрого предсказания.

- Cosine similarity

Косинусное сходство — это математический метод, используемый для измерения сходства между двумя векторами или документами. В первой версии программы сходство косинусов использовалось для прогнозирования подсказки ввода на основе предыдущих входных данных из существующего набора данных. Вычисляя косинусное сходство между различными подсказками, программа могла идентифицировать подсказки, наиболее похожие на ввод пользователя, и генерировать оптимальный результат. Хотя этот метод был ограничен подсказками в наборе данных, он обеспечил ценную основу для оперативного прогнозирования на ранних этапах проекта.

- Transformers

Немалую роль во второй версии проекта сыграли трансформеры, именно модели-трансформеры. Эти модели использовались для точной настройки исходной модели в соответствии с конкретными требованиями набора данных. Путем точной настройки модели трансформатора проект был направлен на расширение возможностей оперативного прогнозирования за пределами ограничений существующего набора данных. Преобразователи позволили модели генерировать более точные и контекстуально соответствующие выходные данные, используя возможности механизмов глубокого обучения и методы внимания (attention is all you need).

- **prompt-engineering & GPT-API**
GPT-API, или «Generative Pre-trained Transformer API», служил основным механизмом для создания оперативных прогнозов. Он предоставил интерфейс для взаимодействия с моделью GPT, позволяя проекту генерировать согласованные и контекстуально соответствующие текстовые выходные данные на основе обработанных подсказок. Используя GPT-API, в проекте использовались возможности предварительно обученных моделей трансформаторов для создания высококачественных оперативных прогнозов. Вызов API модели GPT упростил взаимодействие между проектом и системой GPT, что позволило более эффективно и действенно генерировать подсказки.
- **Streamlit**
Streamlit — это библиотека Python, используемая для создания интерактивных и удобных веб-приложений. В проекте Streamlit был использован для создания интерфейса сайта для системы генерации подсказок(prompts). Это позволило пользователям легко вводить свои подсказки и получать прогнозы от модели. Streamlit упростил плавную интеграцию модели в удобный интерфейс, что позволило пользователям эффективно взаимодействовать с системой и легко получать оперативные прогнозы.
- **Cloud (AWS)**
В проекте использовались облачные вычисления, в частности платформа Amazon Web Services (AWS), для размещения и развертывания. Используя облачный хостинг, проект обеспечил масштабируемость, надежность и доступность системы. Программа и связанные с ней ресурсы были загружены в облако, что обеспечило беспрепятственный доступ и использование приложения из любого места. AWS предоставил надежную и гибкую инфраструктуру, позволяющую проекту справляться с различными рабочими нагрузками и обеспечивать бесперебойную работу системы генерации подсказок.

Результаты:

В первой версии проекта использовалось косинусное сходство для прогнозирования выходных подсказок на основе предыдущих входных данных из существующего набора данных. Хотя косинусное сходство является полезным методом измерения сходства между подсказками(prompts), оно имеет свои ограничения.

Что касается результатов, первая версия проекта смогла генерировать оперативные прогнозы, которые были аналогичны входным данным, предоставленным в наборе данных. Однако эффективность прогнозов ограничивалась подсказками, уже присутствующими в наборе данных. Это ограничивало способность системы генерировать разнообразные и контекстуально подходящие подсказки, выходящие за рамки существующего набора данных. Следовательно, первая версия проекта могла быть не столь эффективной для задач генерации подсказок в режиме реального времени или вне набора данных.

Подход на основе косинусного подобия в первой версии служил фундаментальным шагом в понимании быстрого предсказания, но его можно было улучшить.

Основным недостатком было отсутствие возможности адаптации к входным данным за пределами существующего набора данных, что препятствовало способности системы генерировать подсказки, которые действительно соответствовали требованиям пользователя.

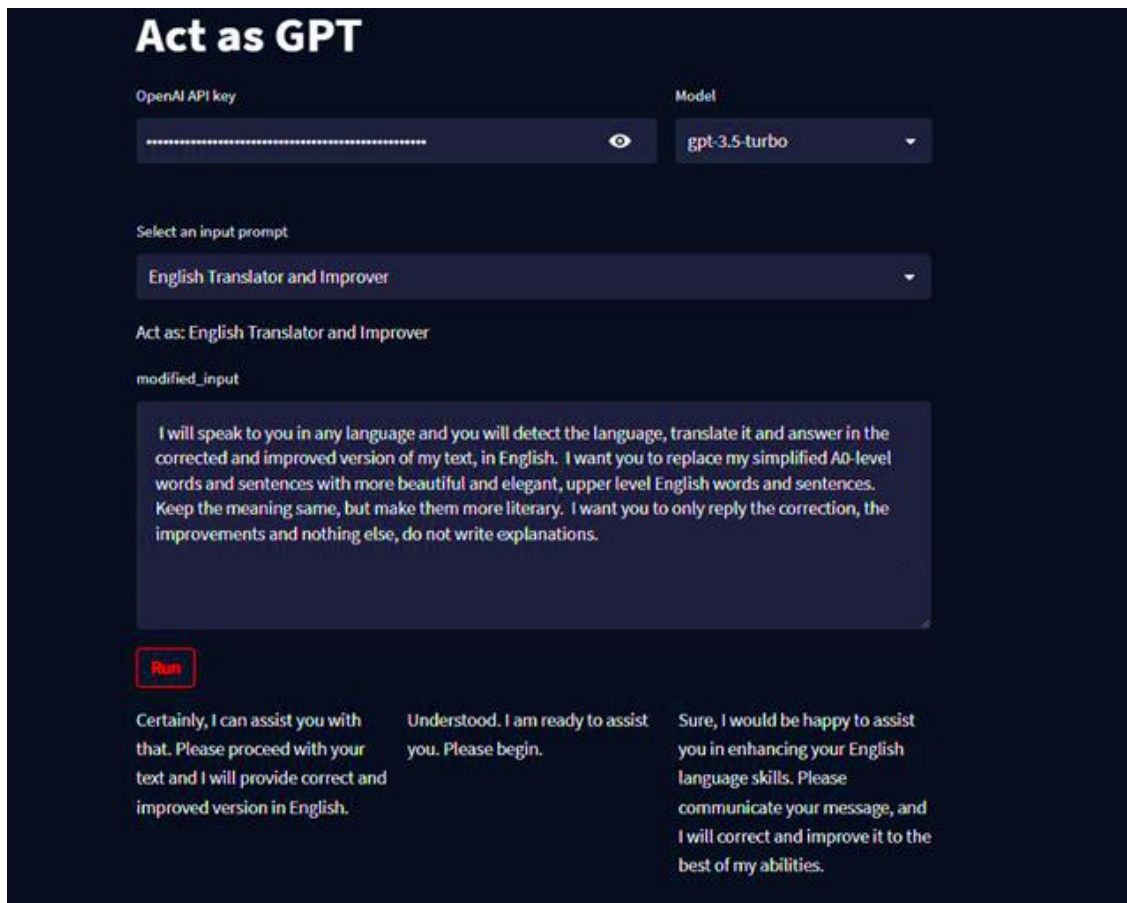


Рис 1. Вывод первой версии на локальное устройство

Улучшенные результаты:

Чтобы преодолеть эти ограничения и добиться улучшенных результатов, проект впоследствии был преобразован во вторую версию, включающую модели трансформаторов, методы prompt engineering и GPT-API. Эти усовершенствования были направлены на расширение возможностей оперативного прогнозирования, создание более точных и соответствующих контексту выходных данных, а также на улучшение пользовательского опыта в целом.

Not secure | 13.53.130.108:8501

Рис 2. IP на сервере AWS

Act as GPT

OpenAI API key Model

..... gpt-3.5-turbo

How many tokens?

50 210 400

Type input here

English teacher

Act as: English teacher

modified_input

I want you to act as an English teacher. I will provide you with a list of topics, and you will need to come up with an effective way to cover them. My first request is, "natural disasters."

Run

Рис 3. Вторая версия со счетчиком жетонов

Run

Absolutely! As an English teacher, there are many ways to approach the topic of natural disasters with your students. Here are a few ideas:

1. **Vocabulary:** Begin by introducing key vocabulary related to natural disasters, such as hurricane, earthquake, tornado, flood, drought, wildfire, and tsunami. You can provide definitions, examples of each term, and images to help students visualize the concepts.
2. **Reading:** Provide students with articles or news stories about natural disasters, and have them read the texts to gain a deeper understanding of the topic. Encourage students to summarize the main points of each text and

Great! Let's talk about natural disasters.

Natural disasters are events produced by natural phenomena that cause widespread destruction or damage to property and human life. Examples of natural disasters include earthquakes, floods, hurricanes, tsunamis, tornadoes, and wildfires.

When discussing natural disasters in English, it's important to understand the appropriate vocabulary to use, as well as the different verb tenses to describe them.

For example, when talking about a past natural disaster, the past simple tense can be used. For instance, "The Hurricane Katrina devastated New Orleans in 2005."

When making predictions about

Great! Let's start!

When we talk about natural disasters, we refer to significant events that occur due to the Earth's natural processes. These events can cause destruction and loss of life, and they include events like earthquakes, tornadoes, hurricanes, tsunamis, volcanic eruptions, and floods.

When teaching about natural disasters, I would recommend starting with an overview of the different types of disasters and their causes. Then, we can dive deeper into each type of disaster and talk about the symptoms or early warning signs and the necessary precautions that can be taken, before, during, and after a disaster striking.

For example, if we talk about an earthquake, we can explain that before the earthquake, there may

Рис 4. Вывод второй версии в облаке

Заключение

В заключение следует отметить, что стремительный прогресс в NLP и машинном обучении вывел область генерации текстов на новый уровень, предоставив возможности преобразования как для отраслей, так и для отдельных лиц. ChatGPT стал мощным инструментом, революционизировавшим различные области и позволяющим пользователям решать сложные проблемы и эффективно выполнять задачи. Однако проблема выбора оптимальных подсказок сохраняется, что мешает пользователям в полной мере использовать потенциал больших языковых моделей. ActasGPT представляет собой новаторский проект, направленный на прямое решение этой проблемы. Предоставляя комплексное решение для быстрой формулировки и оптимизации вывода языковых моделей, ActasGPT улучшает взаимодействие с пользователем и максимально увеличивает полезность этих мощных систем генерации текста. Благодаря передовым алгоритмам, контекстуальному анализу и функциям совместной работы ActasGPT позволяет пользователям получать точные, последовательные и соответствующие контексту ответы, стимулируя инновации и расширяя границы возможностей генерации текста. С ActasGPT пользователи могут раскрыть весь потенциал языковых моделей и вступить в новую эру эффективной и действенной генерации текста.

Источники

- 🌀 Hugging Face: “Документация трансформеров.” [Online]. Available:
🌀 <https://huggingface.co/docs/transformers/training>
- 🌀 Блог TensorFlow: “Introducing TensorFlow Datasets.” [Online]. Available:
🌀 <https://blog.tensorflow.org/2019/02/introducing-tensorflow-datasets.html>
- 🌀 X. Glorot, A. Bordes, and Y. Bengio, "Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach," arXiv:1706.03762 [cs.LG], Jun. 2017. [Online]. Available:
🌀 <https://huggingface.co/docs/transformers/training>
- 🌀 OpenAI: "GPT-4." [Online]. Available:
🌀 <https://platform.openai.com/docs/models/gpt-4>.
- 🌀 H. Khatter, N. Goel, N. Gupta, and M. Gulati, "Movie Recommendation System using Cosine Similarity with Sentiment Analysis," in Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 597-603, doi: 10.1109/ICIRCA51532.2021.9544794. [Online]. Available:
🌀 <https://ieeexplore.ieee.org/document/9544794>.