**НИТУ «МИСИС»**

**Кафедра инженерной кибернетики**

**ОТЧЕТ**

по

**ЛАБОРАТОРНОЙ РАБОТЕ**

*«Исследование возможностей и оценка качества генерации русскоязычных текстов NLP-моделями «семейства» ruGPT-3.х»*

учебная дисциплина «Методы искусственного интеллекта»

**Студент:** Прокуденко Кирилл Игоревич

**Группа**: \_БПМ-21-3\_\_\_\_\_\_\_

**Преподаватель: Хонер П. Д.**

**Оценка: \_\_\_\_\_\_\_\_5\_\_\_\_\_\_\_\_\_**

**Дата защиты: \_\_30.10.24\_\_\_\_**

**2024 г.**

Содержание

[1. Введение 3](#_Toc183885251)

[2. Описание задачи 3](#_Toc183885252)

[3. Выбранные средства для разработки ПО 3](#_Toc183885253)

[4. Исходные данные 3](#_Toc183885254)

[5. Результат 4](#_Toc183885255)

[6. Анализ полученных результатов 7](#_Toc183885256)

[7. Выводы по лабораторной работе 7](#_Toc183885257)

[8. Список использованных источников 8](#_Toc183885258)

# Введение

В данной лабораторной работе целью является получить навыки по организации взаимодействия на уровне программного кода между собственным программным обеспечением и программным обеспечением, реализующим любой из вариантов «семейства» NLP-моделей ruGPT-3.x, при решении задачи генерации связного русскоязычного текста на заданную тему.

Также необходимо выполнить оценку уровня качества генерации русскоязычного текста «семейства» NLP-моделей ruGPT-3.x, используя полученные результаты генерации текстов.

# Описание задачи

Требуется разработать приложение на уровне демонстрационного прототипа, решающее здачу по генерации русскоязычного текста с использованием NLP-моделей ruGPT-3.x.

В качестве темы для текста нужно взять аннотацию КНИР в текущем семестре и текст произвольной длины, тематически связанный с выполняемой КНИР.

# Выбранные средства для разработки ПО

Для выполнения задачи сначала была выбрана модель ruGPT-light, но генерируемый текст был слишком неосмысленным. Поэтому было решено сменить модель на ruGPT-medium.

Программа разрабатывалась в VS Code на языке Python. Для работы с моделью использовались библиотеки torch и transformers. Для визуального интерфейса была взята библиотека tkinter.

# Исходные данные

В качестве исходных данных возьмём несколько текстов.

* **Аннотация КНИР в седьмом семестре:** «Тема КНИР: Рекомендательная система фильмов и сериалов, основанная на базе уже просмотренных».
* **Пример текста КНИР за шестой семестр:** «Алгоритм A\* Так как задача состоит в расчёте пути внутри здания между двумя точками, то становится очевидным, что обычный алгоритм Дейкстры будет тратить много лишнего времени на обработку ВГ, очевидно не ведущих в стороне конечной точки. В связи с этим возникает потребность не тратить время на обработку вершин, отдаляющихся от конечной точки. Именно с этой целью будет использовать алгоритм A\*. Принцип работы данного алгоритм заключается в».

# Результат

По умолчанию параметры для модели будут следующие:

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рис. 1 – Параметры для генерации по умолчанию.

Рассмотрим примеры работы программы с различными конфигурациями параметров и входных данных.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 2 – Результат запуска на тесте с аннотацией КНИР c параметрами по умолчанию.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 3 – Результат запуска на тесте с аннотацией КНИР с другой температурой.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 4 – Результат запуска на тесте с аннотацией КНИР с другим значением n-грамм.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 5 – Результат запуска на тесте с примером текста из КНИР с параметрами по умолчанию.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 6 – Результат запуска на тесте с примером текста из КНИР с другой температурой.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 7 – Результат запуска на тесте с примером текста из КНИР с другим значением n-грамм.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание

Рис. 8 – Результат запуска на тесте с последовательностью чисел.

# Анализ полученных результатов

Как видно из тестов, во всех случаях сгенерированный результат примерно сохраняет тематику изначального текста, но смысла у полученного текста почти нет. При изменении параметров генерации результат не улучшается.

Изображение выглядит как диаграмма, круг

Автоматически созданное описание

Рис. 9 – Структура датасета ruGPT-3.5.

В документации показан примерный состав датасета на котором обучалась модель (рис. 9). Из него наглядно видно, что при обучении использовался примерно 1% научных статей. Это объясняет очень низкий результат генерации текста для КНИР.

# Выводы по лабораторной работе

По представленным результатам можно понять, что модель ruGPT-3.5 (версии medium) не стоит использовать для генерации текстов на научные темы. Модель часто выдает случайную на вид последовательность символов и очень быстро отходит от изначальной темы. Различные конфигурации параметров генерации незначительно улучшают получаемый результат.

Также стоит отметить неполную документацию модели, из которой почти невозможно понять, за что отвечают параметры генерации.

# Список использованных источников

1. Документация библиотеки tkinter // Python Docs URL: <https://docs.python.org/3/library/tkinter.html> (дата обращения: 27.10.24).
2. Документация библиотеки Transformers // Hugging Face URL: [https://huggingface.co/docs/transformers/index](https://docs.python.org/3/library/tkinter.html) (дата обращения: 27.10.24).
3. Документация к моделям семейства ruGPT-3 // GitHub URL: [https://github.com/ai-forever/ru-gpts](https://docs.python.org/3/library/tkinter.html) (дата обращения: 27.10.24).