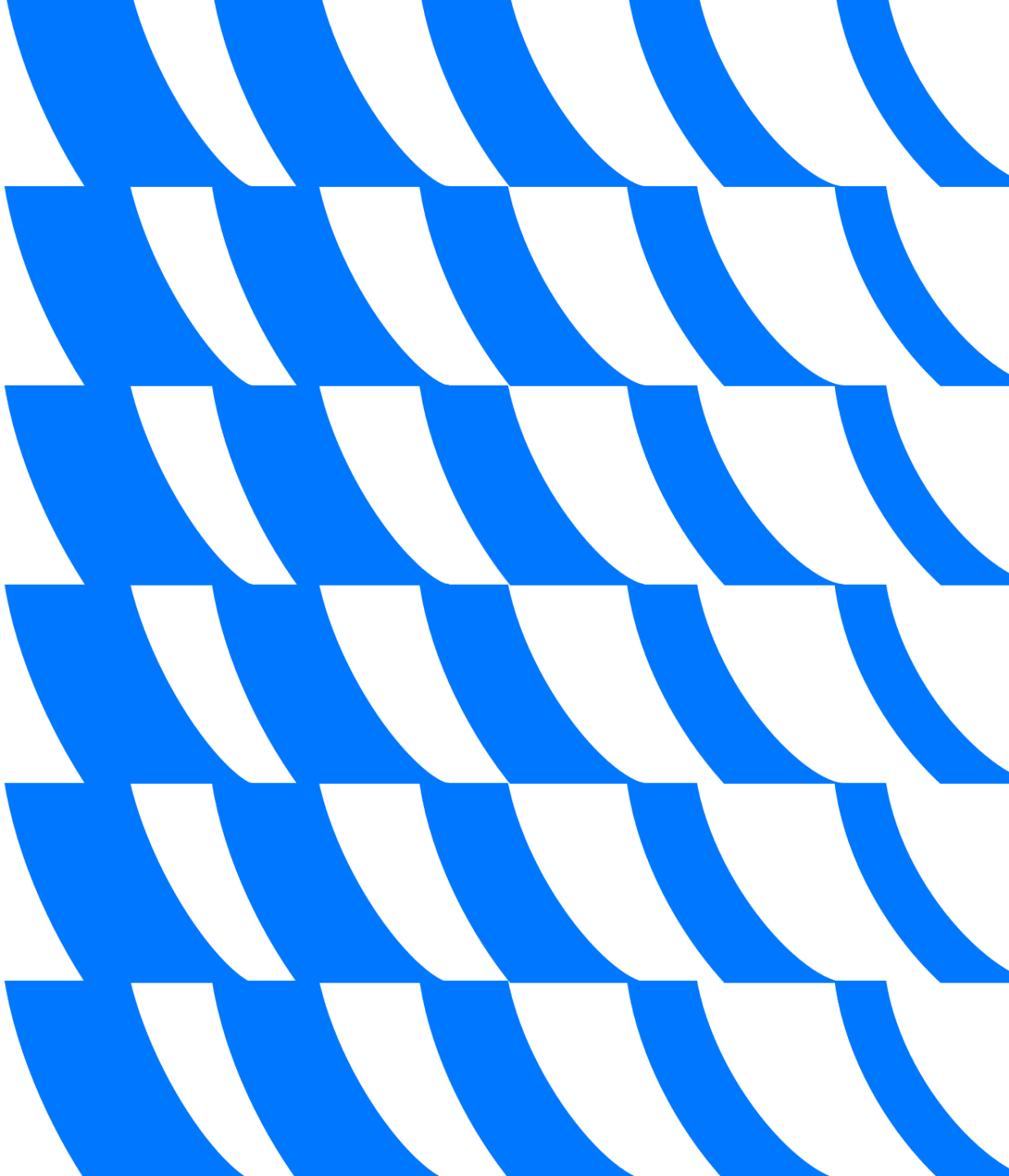
Генерация фейковых текстов веб документов по запросу пользователя

Ермолаев Никита

образование

Описание проекта



Задача проекта

Основная задача: создать модель генерации текстов основанную на **Transformer** подобных архитектурах

Оформление:

(Минимум) Python библиотека с возможностью запуска основного функционала (обучение, тестирование, генерация) из командной строки.

(Опционально) Сервис с веб-интерфейсом. Который, например, принимает запрос и генерирует простую html страницу с текстом документа.

Ключевые метрики

Глобально: насколько успешно сгенерированные документы смогут обмануть модели ранжирования

Ключевая метрика: средняя позиция сгенерированных документов в выдаче среди реальных документов с точки зрения моделей ранжирования (bm25, bert’ы)

Дополнительный критерий: уязвимость к детекции. Надо показать, что нет тривиального способа ее обнаружить.

Например:

1) Текст из набора ключевых слов – плохой (у него будет маленькое правдоподобие, он не похож на естественный)
2) Текст из повтора запрос 10 раз подряд – плохой (посчитаем кол-во уникальных слов относительно длины)

Данные

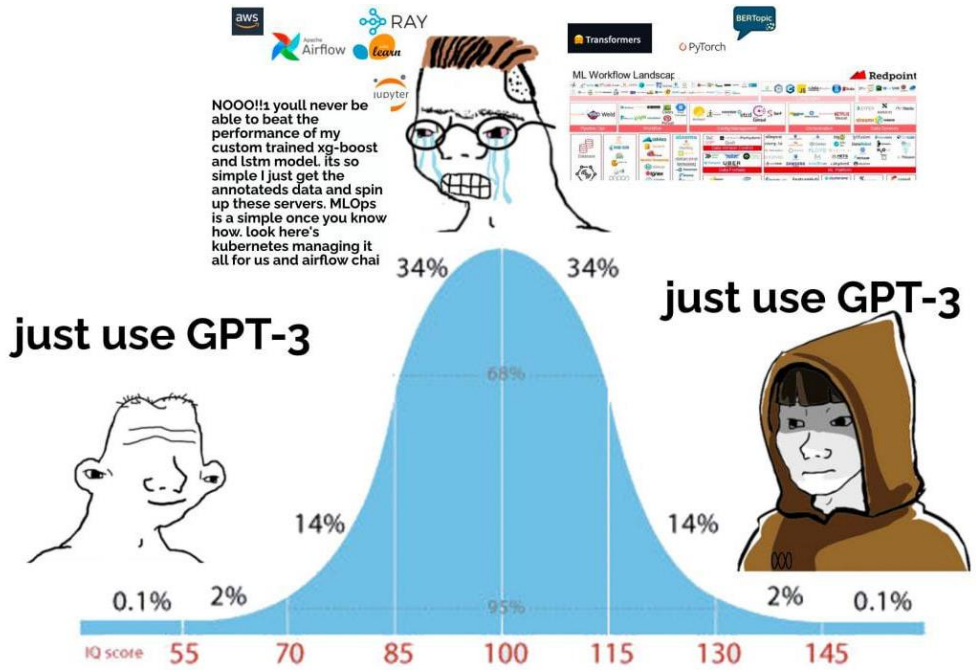
Веб-документ:

Query:
Title:
Meta:
Body:
Queries:

```
('0 00 дом muzono net raim feat artur adil скачать',
 'Скачать Дом - RaiM feat. Artur & Adil — Mussic.kz',
 'Здесь можете бесплатно скачать Дом - RaiM feat. Artur & Adil — Mussic.kz. Новинки казахских песен, ежедневное обновление',
  '» RaiM feat. Artur & Adil - Дом 2018 RaiM feat. Artur & Adil - Дом 2018 Дата: 7-12-2018, 18:48 Скачать mp3 Скачиваний: 3 145
RaiM feat. Adil - Роза 2018 RaiM feat. Adil - Роза 2018 RaiM feat. Artur - Карамель 2018 RaiM feat. Artur - Карамель 2018 RaiM
feat. Adil - Тая 2018 RaiM feat. Adil - Тая 2018 Artur & Adil (DDrecords) - В долину хочу 2018 Artur & Adil (DDrecords) - В дол
ину хочу 2018 Raim & Artur - Догола 2018 Raim & Artur - Догола 2018 Raim & Artur feat. Zhenis - Дискотека из 90 2018 Raim & Art
ur feat. Zhenis - Дискотека из 90 2018 Raim & Artur - Лава 2018 Raim & Artur - Лава 2018 BN x Raim & Artur - Лучший 2018 BN x R
aim & Artur - Лучший 2018 Ninety One - Men Emes 2019  Әбдіжаппар Әлқожа - Жәудір мұң 2019  Әбдіжаппар Әлқожа – Факт 201
9  Төреғали Төреәлі – Түнгі вальс 2019  Әсет Қарабалин - Екеумізге не болды 2019  Қайрат Нұртас – Өзің ғана 2019&nb
sp Жазира Байрбекова – Қазақ қызы әдемі 2019  Қуандық Рахым – Неке жүзік 2019  Ернар Айдар & Зарина Омарова – Мөлдіреге
н-ай 2019  Беркут & Наташа Королева - Қылықты қыз 2019  2019 © mussic.kz, Все права защищены. Мы не несём ответственнос
ти за содержание. Бесплатные произведения предназначены исключительно для предварительного ознакомительного прослушивания. Авто
рские права принадлежат авторам. После прослушивания песен вы должны удалить их со своего компьютера, смартфона, планшета или л
юбого другого устройства на которое был загружен файл в течение следующих 24 часов.',
'raim дом скачать;raim feat artur adil дом;raim artur дом;дом raim artur feat adil')
```

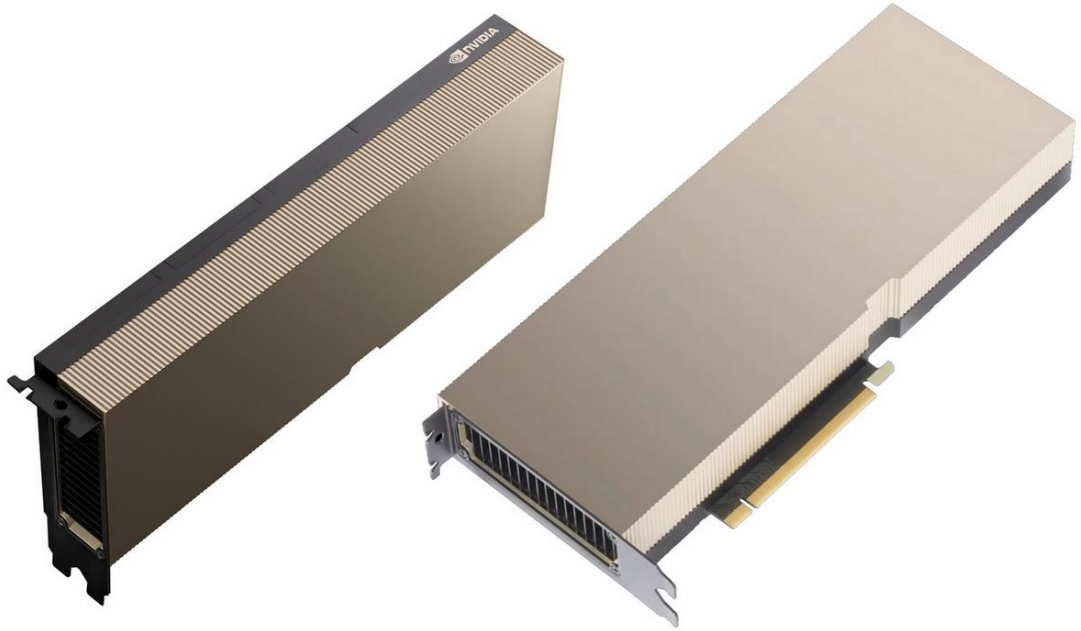
Что можно попробовать

- Завести инференс больших open-source моделей в zero|few-shot режиме
- Дотюнить ru|m-GPT*, ru|m-T5* модели
- Может даже RL?



Ресурсы

Чем больше тем лучше…



…но без GPU никуда

Требования к участникам

- Знание методов NLP
- Опыт применения DL + знание pytorch
- Опыт применения трансформеров в задачах NLP

Какие плюсы?

Для вас:

- Разобраться с задачей генерации текстов
- Получить опыт работы с трансформерами
- Сделать интересный проект

Для нас:

- Лучше понять устройство наших данных
- Понять как работают наши модели, найти слабые места
- Понять как детектить сгенерированные текст

Вопросы?