

Генерация анекдотов

Малашенко Борис
Шахвалиева Юлиана
Шувалова Полина





Критерий качества

Юмор – субъективен

Анекдот хороший, если:

- на заданную тему**
- лексически и логически**
- правильный текст на русском**
- языке**

Данные

|0|1|Расскажи анекдот про key_words |1|-|
«анекдот»|</s>

- Ключевые слова: библиотека [natasha](#)
- Сегментация
- Анализ морфологии и синтаксиса
- Лемматизация

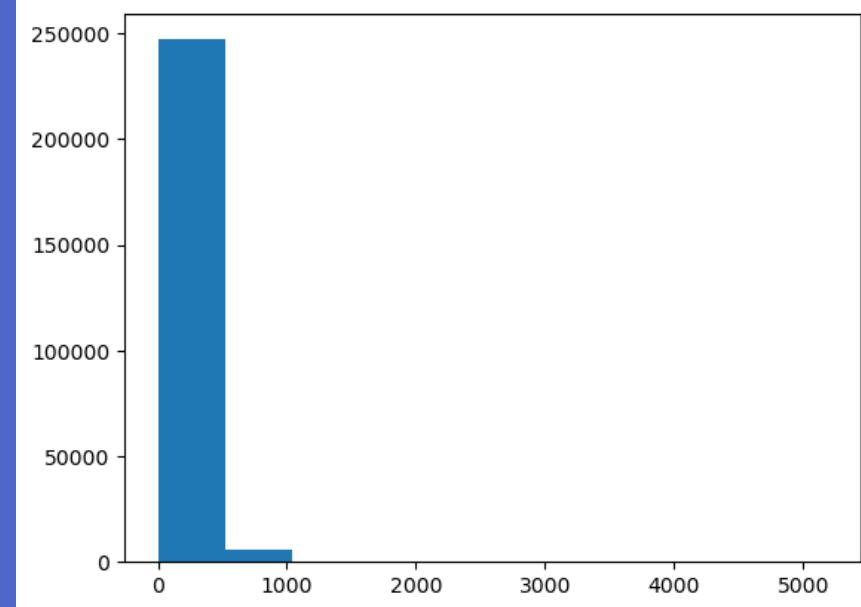
253k

анекдотов

LSTM

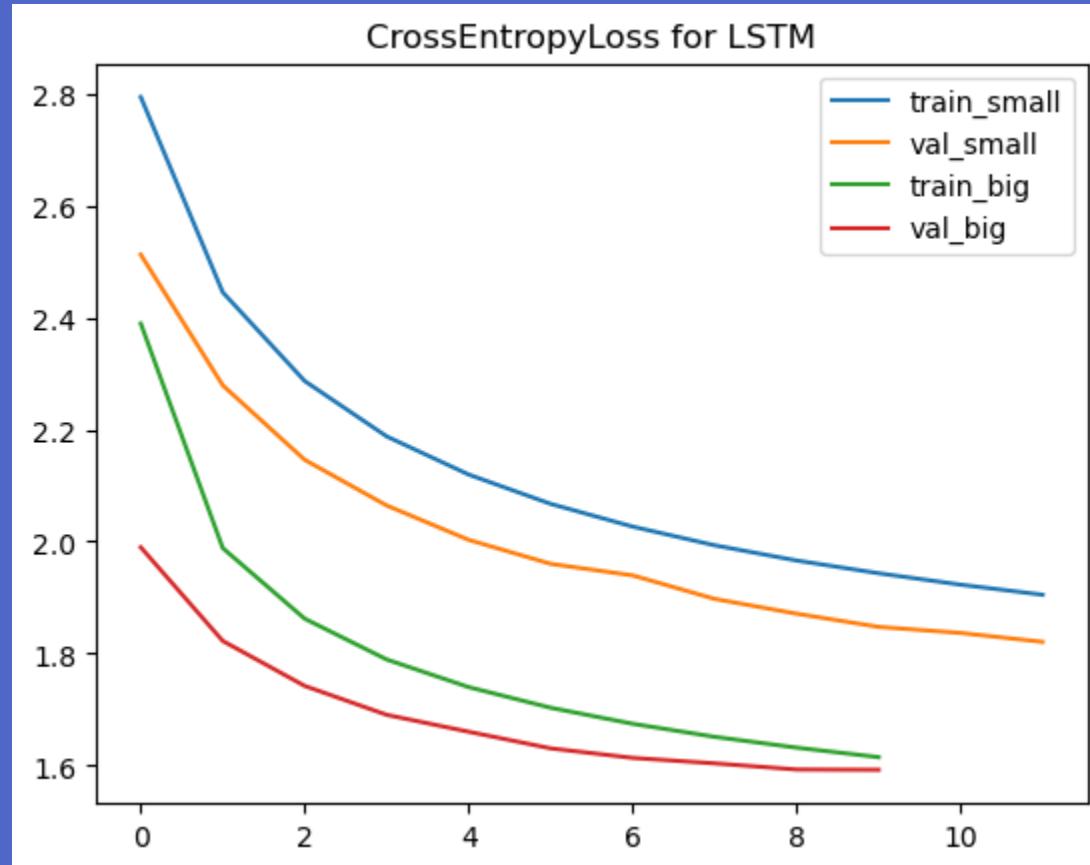
CharTokenizer

- Словарь – уникальные символы из датасета
- Max_length = 512
- Задача модели: предсказание $[i + \text{seq_len} + 1]$ - того токена.
- seq_len = 30
- ReLU, Dropout(0.2), CrossEntropyLoss



Архитектура LSTM

- **Small**
 - 2 слоя LSTM размерности 512
 - 2 линейных слоя
- **Big**
 - 3 слоя LSTM размерности 1024
 - 2 линейных слоя
- **1 эпоха ~30 минут**
- **10 эпох**



Генерация LSTM

- Предсказание следующего символа, опираясь на пользовательский ввод (softmax)
- Для регулирования случайности введен гиперпараметр k, на который умножается слой перед softmax
- Чем больше k, тем меньше случайности
k= 1
- Чем меньше k, тем больше случайности
k= 5

`/lstm` Вовочка приходит медледь на капешку, а то дам правону когда ты беременная'Эк, поктор ззBSa*il' не может дырый, так еще на меня он просто застевляется и спрашивает:

- Чдеть, посмотрите у нас еще чашечку сделали, где крестьянный!
- Ооже нет, - зу: Нету - о том смотрю локоль кончаешь бернонет басенькое атпичинки, то открыл у тебя лопы!

Песэ

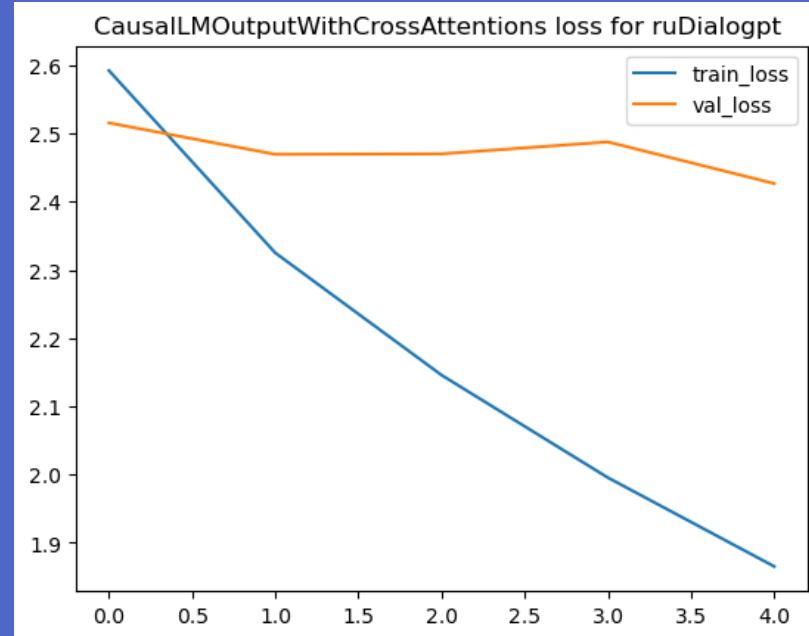
`/lstm` Вовочка приходит в постель и спрашивает:- А что такое слово?- Нет, не понял, а ты подарил себе по половому соседу приходит в комнату и спрашивает:

- А что такое просто под баловом подарили на столе по пороге стоит мужик и подходит к столу. Подходит к нему на столе стоит половина в половине стола на работу и поставил его в постель. Мн подходит к нему и говорит:
- Не понял, ты же просто не подарю свою половину секса не подарить.[SEP]

ruDialogPT

Предобученная на русских диалогах модель GPT_2

- Трансформер с 1,5 млрд параметрами
- Цель обучения: предсказать следующее слово в тексте
- Temperature – случайность генерации
- **1 эпоха ~2 часа 10 минут**
- **5 эпох**

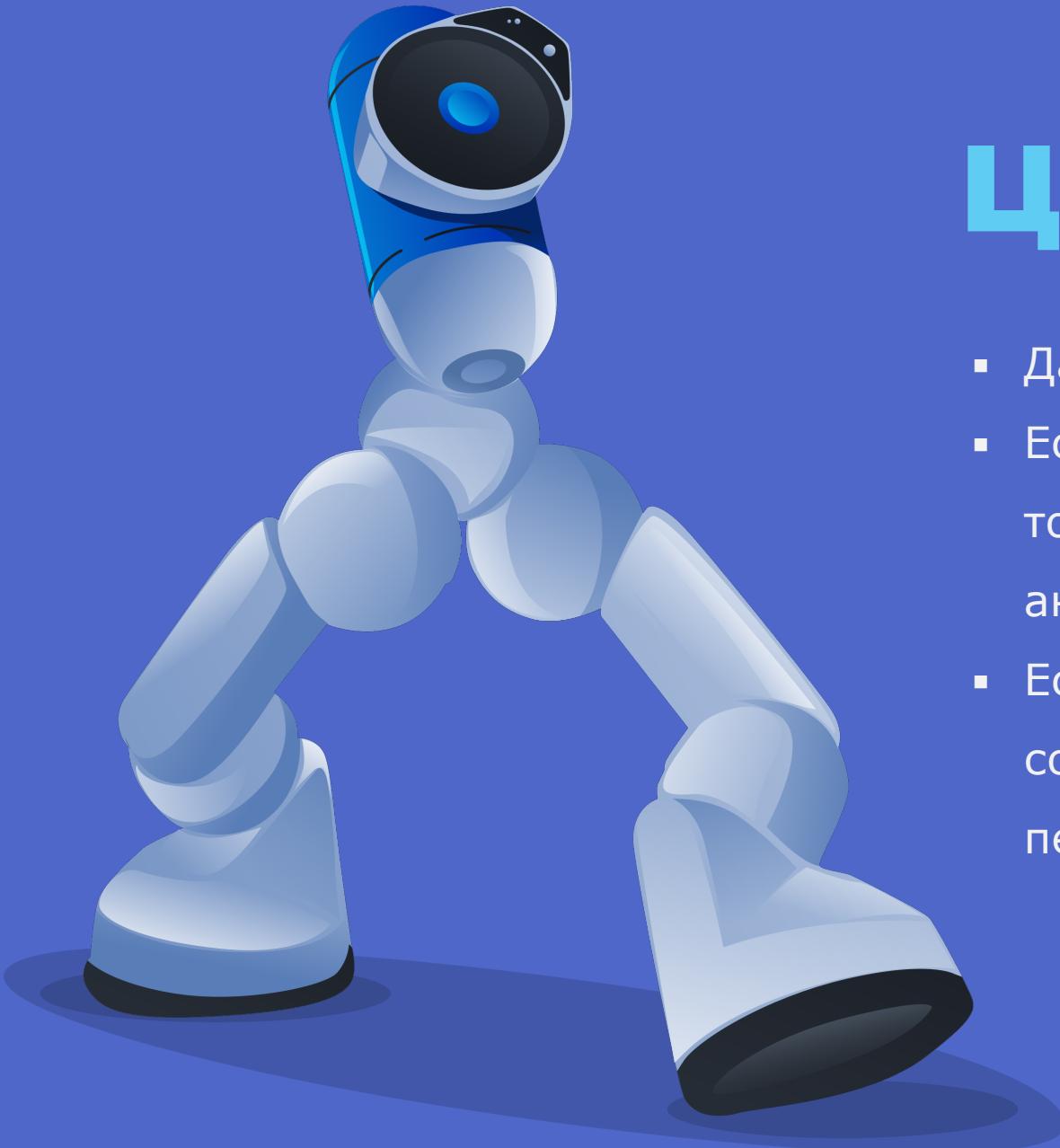


Штирлиц выстрелил в упор, упор упал.

Кроссовки - это не обувь, это - диагноз.

Смех - это лекарство. Но только до тех пор, пока не заразишься им от кого-нибудь.

15:48

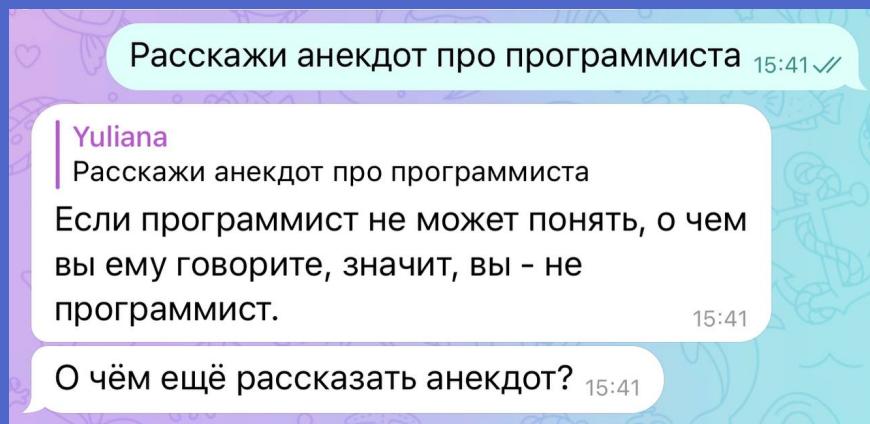
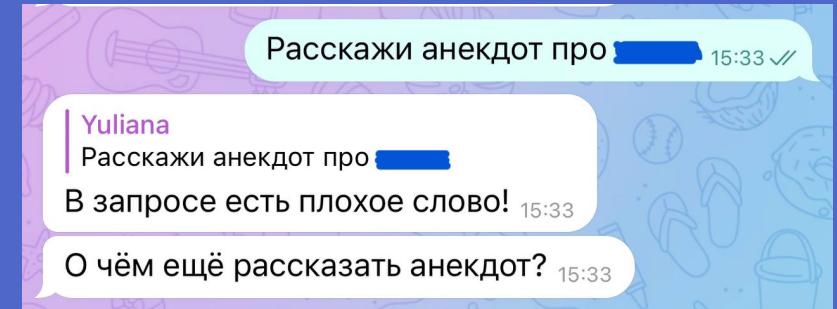


Цензура

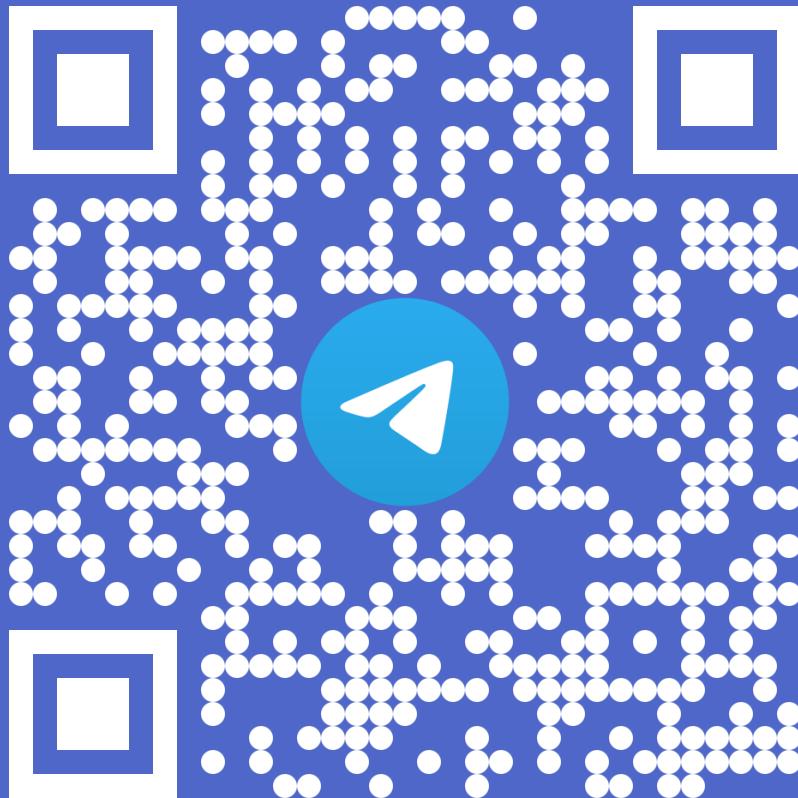
- Датасет из стоп-слов (288)
- Если в запросе присутствует стоп-слово, то модель не станет генерировать анекдот
- Если модель сгенерировала анекдот, содержащий стоп-слова, то она перегенерирует его

Бот Telegram

- Библиотека [aiogram](#)
- Асинхронный режим
- Алгоритм работы:
 - получаем запросы от пользователей
 - передаём в заранее загруженные на GPU модели
 - получаем ответы



/
/text генерация анекдота от GPT
/lstm <text> генерация анекдота от LSTM
/change_param изменение параметров генерации
/params список текущих параметров



Спасибо за внимание!