Генерация вопросов по тексту

Идея

Хотим генерировать вопросы по тексту, в котором есть ответ на эти вопросы

Зачем?

- Проведение викторин по заданной тематике / генерация тестов по тексту
- Бенчмаркинг RAG-пайплайнов

Решение

Выбрал google-t5/t5-base , датасет squad 1.1

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

SQuAD 1.1 contains 100,000+ question-answer pairs on 500+ articles.

Решение

HF Training Arguments:

- learning_rate: 1e-4
- train_batch_size: 16
- eval_batch_size: 16
- gradient_accumulation_steps: 16
- optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- lr_scheduler_type: linear
- num_epochs: 3

Format:

generate question: {text}

Loss:

Деплой

HF transformers: AutoModelForSeq2SeqLM, параметры генерации:

- max_length=128
- min_length=3
- num_beams=1
- num_return_sequences=num_questions * 3
- no_repeat_ngram_size=2
- temperature=1.1
- top_k=100
- top_p=0.98
- do_sample=True
- early_stopping=True

Streamlit.app

- https://pretty-q-generator.streamlit.app/

Example text #1:

Example text #2:

Example text #3: