Задание:

Необходимо дообучить многоязычную модель BERT (bert-base-multilingual-cased) на русскоязычной части датасета TyDi QA для вопросно-ответного поиска. Данные приложены в папке. Для сдачи задания достаточно получить F1=65.

Что может помочь при выполнении домашнего задания (1-3 ссылки самые полезные):

1. <https://huggingface.co/transformers/preprocessing.html> — как выполняется токенизация и препроцессинг
2. <https://huggingface.co/transformers/master/custom_datasets.html> — как работать со сторонними датасетами
3. <https://colab.research.google.com/drive/1uSlWtJdZmLrI3FCNIlUHFxwAJiSu2J0-#scrollTo=uJ7PRx2dKqFN> — ноутбук про то, как использовать уже готовую модель для вопросно-ответного поиска. Здесь можно подсмотреть, как привести предсказания модели в “человеческий” вид
4. <https://github.com/huggingface/transformers/tree/master/notebooks> — ноутбуки от huggingface, как дообучать и использовать модели
   1. <https://colab.research.google.com/github/huggingface/notebooks/blob/master/examples/question_answering.ipynb> — ноутбук для вопросно-ответного поиска. Тут более сложный код для предобработки данных и извлечения ответов.
5. <https://github.com/allenai/bi-att-flow/blob/master/squad/evaluate-v1.1.py> — официальный скрипт для вычисления F1 и Exact Match для датасета SQuAD
6. ner\_rubert.ipynb — моя попытка дообучить русскоязычную модель для задачи извлечения именованных сущностей

Важно:

1. Из Google Colab можно сохранять/загружать модели/данные с гугл диска. Выполните следующий код в ячейке, чтобы примонтировать диск:

from google.colab import drive

drive.mount('/content/gdrive')

1. При работе с нейронными сетями включите использование GPU в Google Colab: Среда выполнения -> Сменить среду выполнения