Домашнее задание 5

В данном домашнем задании Вам предстоит обучить нейронную модель лучше ранжировать документы на последней стадии поиска.

1. Датасет

Для обучения вам предоставляется файл *train.tsv* в формате аналогичном *queries.tsv*, используйте эти пары (запрос, документ) в качестве положительных примеров. Отрицательные примеры подберите на своё усмотрение. Используйте полученные в предыдущих домашних работах тексты статей википедии как входные данные для документа.

2. Обучение

Используйте готовую предобученную BERT-оподобную модель в качестве основы. Дообучите её на задачу бинарной классификации - предсказания того, является ли документ релевантным запросу.

Рекомендую пользоваться библиотекой transformers (документация). Она поддерживает pytorch и tensorflow, содержит репозиторий предобученных моделей.

Рекомендую использовать colab с настройками среды GPU, чтобы ускорить обучение и применение.

3. Поиск

Используйте предсказания модели для переранжирования лучших N документов, полученных при поиске BM25. Сравните качество полученного алгоритма поиска при разных значения N (10, 25, 50).

~~Дополнительно~~

4. Оптимизация применения (+1 балл)

Обучите BERT-опободную нейронную модель в сиамской схеме, аналогичной DSSM, когда запросная и документная часть считаются независимо, и связываются только в конце через косинус.

Предпрочитайте векторные представления для запросов и документов, требуемых для оценки качества.

Сравните качество и **скорость** полученной модели с предыдущим пунктом.