Домашнее задание 5

Датасет Miracl представляет собой мультиязычный корпус документов, предназначенный для задач информационного поиска. Для данного задания была выбрана русскоязычная часть данного датасета,которая содержит **9 543 918 документов**.
Основной целью обработки было создание чистого и релевантного поднабора данных, оптимального для задач обучения моделей.

2. Этапы предобработки

2.1. Фильтрация по длине текста

На первом этапе из датасета удалялись документы, не соответствующие следующим критериям:

• **Слишком короткие тексты:** документы длиной менее **50 слов**.

• **Слишком длинные тексты:** документы длиной более **100 000 слов**.

После фильтрации было удалено **708 809 документов**.

Результат: на данном этапе оставлено **8 835 109 документов**.

2.2. Дедубликация

Для удаления дублирующихся документов применялась техника **MinHash** с использованием **n-грамм** (где **n = 13**). Такой подход позволил эффективно находить схожие документы.

2.3. Фильтрация на основе перплексии

Для дальнейшей очистки данных использовалась языковая модель **GPT-2**. На основе перплексии фильтровались тексты.

2.4. Исключение схожих документов между train и test

После разделения данных на **обучающую (train)** и **тестовую (test)** выборки была проведена дополнительная фильтрация. Целью было удаление из **train**-выборки документов, похожих на те, что присутствуют в **test**. Это предотвратило утечку данных из тестовой выборки в обучающую.

Результат: итоговый размер датасета составил **7 580 594 документа**.

3. Итоговый результат

Полученный очищенный датасет опубликован на платформе HuggingFace: **ru-miracl-cleaned**.