Text relevance

Калашников Дмитрий

Что было сделано

- Анализ
- Предобработка
- Предрасчет
- TF*IDF score
- BM25 score
- BM25f score
- Fasttext + косинусное расстояние
- Линейная комбинация BM25f и fasttext
- **Score: 0.635**

Предрасчет

- docs.tsv.gz ~15ГБ
- Распакованный вариант ~60ГБ
- Для ускорения обработки: распределения контента по шардам
- Предрасчет IDF для каждого терма, TF для каждого терма в документе - отдельно для заголовка и тела документа
- Предрасчет средней длины документа, количества документов.
- Предрасчет fasttext-эмбеддингов для каждого документа - отдельно заголовка и заголовка + тела.
- Для запросов - расширение запросов до замены наиболее вероятной раскладки

Модели

- TF*IDF
- BM25
- BM25f - взвешивание термов в зависимости от зоны
- Вес заголовка - 2, вес тела - 1
- **Fasttext**, косинусное расстояние между запросами и документами, запросами и заголовками документов
- Линейная комбинация fasttext и bm25f
- Отбор кандидатов fasttext, ранжирование bm25f

Что еще хотелось сделать

- Учитывать пассажи в bm25f
- Честный spellchecker
- Честный булев поиск по индексу
- Translater

Спасибо!