# Ranking long tail queries

ShcherbakovaElena[Technosphere]

24 февраля 2021

# Подготовительная работа

- Исправление текстов запросов с помощью спеллчекера.;
- Получение заголовков документов и их нормализация;
- Работа с документами: получение файлов для каждого документа;
- Работа с документами: получение файлов, в котором для каждого запроса все тела документов, относящихся к нему.

# Признаки

- Универсальный кодировщик предложений кодирует текст в высокоразмерные векторы, которые можно использовать для классификации текста, семантического сходства, кластеризации и других задач на естественном языке. С помощью universal-sentence-encoder-multilingual-large кодировались запрос и отвечающий документ, находилась косинусная близость между ними.
- Модель gensim.models.doc2vec и косинусная близость заголовков, запросов.
- Натренирована модель fasttext gensim для заголовков и запросов, взята косинусная близость векторов.
- TfidfVectorizer применялся к заголовкам и запросам с различным параметром ngram-range, далее бралась мера близости.
- Пассажи разной длины для заголовков документов и запросов.

# Признаки

- BM25Plus, BM25L, BM25Okapi для части слов запросов и заголовков. Вычиислялась косинусная близость.
- Отношение числа кликов к числу показов, CTR, и вариации.
- Среднее время, проведенное на странице.
- Среднее кол-во кликов до/после перехода к тек. странице.
- Логарифмы количества показов/кликов.
- Модель SDBN...

Лучший итоговый score - 0.77296.

Спасибо за внимание!