Конкурс по текстовой релевантности

Ира Букреева

- обработка документов
- обработка запросов
- ранжирование

обработка документов

- лемматизация - батчами по 10 текстов стеммеру Mystem
- удалила стоп-слова из текстов документов: словарь nltk + еще 112 добавленных руками, всего 264
- сначала брала просто просто тексты и заголовок, потом еще заголовки различных уровней
- 3 разных словаря

обработка запросов

много поисковых расширений

- исправление опечаток
- синонимы
- транслиты
- ошибочные написания
- более сложные синонимы: почему -> причины
- класс запроса
- разные части речи
- в аббревиатуры и обратно
-

ранжирование

- BM25

итоговая формула: score = 3.0 * titlescore + 1.2 * textscore + 1.0 * headersscore

- gensim.models.doc2vec

обучение на тайтлах: 37971 тайтл

лучший скор: переранжирование документов на основе косинусного расстояния между эмбеддингом запроса(без поисковых расширений) и заголовка документа с учетом скора BM25

у 143 документов нет заголовков