КОНКУРС ПО ТЕКСТОВОЙ РЕЛЕВАНТНОСТИ

Салихова Кария

kaggle: Kariya07

ПРОБЛЕМА 1. КАК ПОСТРОИТЬ ИНДЕКС?

- Запросы русскоязычные => можно извлекать только русские слова (или нет?)
- Можно избавиться от цифр (или нет?)
- Сбор ссылок (как обработать?)
- Для BM25F надо сразу отделить части документов

Инструменты: BeautifulSoup, pymorphy2

ПРОБЛЕМА 2. КАК ОТБИРАТЬ РЕЛЕВАНТНЫЕ ДОКУМЕНТЫ?

- Искать документы, в которых есть слова из запроса. Не набирается даже 10 документов.
- Опечатки (расстояние Левенштейна)
- Раскладка
- Сокращения
- Синонимы

} из интернет-словарей

BM25 + ПРИЗНАКИ:

- Близость слов из запроса друг к другу
- Близость слов из запроса к началу документа
- Количество ссылок, ведущих на данную страницу
- Нормализуем все признаки
- Ранжируем по сумме признаков + BM25 с весами (1.0, 1.0, 1.0, 2.0).

BM25F

- Сглаженный IDF:

$$\mathrm{IDF(w)} = \log \frac{N - n(w) + 0.5}{n(w) + 0.5}$$

- Стоп-слова (nltk: english + russian)
- Боремся с отрицательными IDF: вводим нижнюю границу значений

LDA

- Топ-10 документов по версии BM25F ранжируем с помощью LDA.
- Отдельная модель для каждого запроса.
- 5 тем
- Или наоборот: сначала ранжируем LDA, потом BM25F.

Инструменты: gensim

Спасибо за внимание!