В качестве домашнего задания ответьте на следующие вопросы:

1. В чем принципиальные отличия item-item рекомендаций от ALS?

ALS – это прежде всего коллаборативная фильтрация, а item-item - content-based детерминированный алгоритм.

Мы проговаривали это на уроке: для ALS важен факт взаимодействия с товаром, услугой или чем-то еще, что мы рекомендуем. В Item-item это в явном виде может и не быть.

ALS использует эмбеддинги данных о пользователе и товаре. В рекомендациях item-item используется информация о товаре и подбирается исходя из сходства пользователей и товаров

2. Приведите 3 примера весов user-item матрицы для задачи рекомендаций товаров

* оценки товаров, выставляемые пользователями
* количество комментариев
* репост о товаре в фейсбуке, инстаграме и т.д
* отзыв о товаре с оценкой

3. Какие ограничения есть у ALS? (Тип информации, линейность/нелинейность факторов и т д)

- по типу информации на входе: только факт взаимодействия (обсуждали на семинаре)

- линейна относительно факторов

- отсутствие персонализации (как гипотеза – является ли?)

- не учитывает влияние нестационарных составляющих временного ряда

- однородность и качество входных данных

- имеет одинаковые веса ошибок предсказаний.

4. Мы рассматривали bm\_25\_weight. Опишите, как оно работает. Как сделать рекомендации только на основе bm\_25? (Можно и нужно пользоваться любыми источниками, приложите на них ссылки)

<https://ru.wikipedia.org/wiki/Okapi_BM25>

Как это работает расписано в статье на хабре:

<https://habr.com/ru/post/162937/>

Там же есть ссылка на документ с расчетом bm25.

📌 Алгоритм BM25 пришел на смену TF-IDF. Он более сложный и его суть заключается в оценке текста на странице, основываясь на количестве и месторасположении ключевых слов.

📌 BM25 — функция ранжирования документов по их релевантности поисковому запросу (на сколько часто встречается слово в каждом документе коллекции).

📌 Алгоритм BM25F (field=зона) учитывает не только текст, но и его отдельные участки или зоны:

1. Заголовок документа Title.  
2. Текст документа, но с рядом модификаций:  
— стандартный вариант  
— с «угасающим» весом от начала документа (приоритет отдается вхождениям в начале текста)  
— только по точным вхождениям  
— по части слов из запроса (скажем, по парам слов)  
3. По текстовым заголовкам и фрагментам с акцентами (strong и т.п.).  
4. По текстам анкоров исходящих с документа ссылок (вероятно, внутренних).  
5. По URL-адресу документа (вспоминаем про важность ЧПУ).

Каждый участок текста имеет свою значимость для ранжирования страницы.

📌 При анализе текстовой релевантности документа используются и другие формулы, например:

• Наличие/отсутствие точных вхождений многословного запроса в документе.  
• Взаимное расположение слов из запроса (пар слов) в документе.  
• Сочетание Title и текста документа.  
• Соответствие языковой модели.  
• и др.

❗ Итог: понимая алгоритм BM25F, становится понятно, почему нужно оптимизировать разные участки документа

Оригинал:

<https://mindubaev.com/100-dnej/bm25bm25f/>