Московский авиационный институт
Факультет прикладной математики и физики

Лабораторная работа №9

по курсу:
«Информационный поиск»
по теме:
«Зонный поиск»
2 семестр

Студент: Ахмед С. Х.
Преподаватель: Калинин А. Л.
Группа: 8О-106М

Москва, 2019 г

Постановка задачи

Необходимо добавить в поисковый индекс информацию о зонах, в которых встретились термины. Как минимум, нужно сделать отдельные зоны для заголовков документов. Так же, необходимо учесть эти зоны в ранжировании, причём таким образом, чтобы поиск стал искать лучше. В отчёте нужно привести:

• Побитовое описание индекса с зонами.

• Формулу ранжирования, подобранные веса.

• Оценку качества поиска после внедрения зон.

Есть ли запросы, по которым качество ухудшилось? Почему? Что можно сделать, чтобы качество поиска по ним улучшилось, а по остальным запросам –не ухудшилось бы?

Ход работы

В поиск будет добавлен поиск по зоне заголовков документов. Индекс зоны представляет собой обратный индекс- слова в заголовках являются токенами и им в соответствие ставятся список постингов. Для данной лабораторной работы он вынесен в отдельный файл

					…								…		
4 байта – длина токена-слова, содержащегося в заголовке (N)				N байт - токен			4 байта – длина списка документов (M)				M*4 байт – список doc_id документов				

Формула, по которой осуществлялось ранжирование документа с учетом зон следующая:

$$TF - IDF(term, doc) = TF - IDF(term, doc) + Rewards\big[term[idx]\big]$$

Где $Rewards\big[term[idx]\big]$ является наградой за вхождение термина в заголовок документа. Практическим путем было выявлено, что оптимальное значение для этой награды является 0.1.

Оценим качество поиска после введенных изменений:

	0	1	2	3	4	5	6	7	8	9	P@1Old	P@3Old	P@10Old	P@1	P@3	P@10
служить всю жизнь государству	5	4	4	2	2	2	2	5	2	4	1	1.000000	0.5	1	1.000000	0.5
соглашение о свободной торговле	5	2	4	3	1	4	4	5	3	4	0	0.333333	0.6	1	0.666667	0.8
зимний дворец	5	3	3	4	4	2	1	2	3	4	1	1.000000	0.7	1	1.000000	0.7
российская империя	3	2	5	4	2	1	1	2	2	2	0	0.333333	0.2	1	0.666667	0.3
англо русской войне	2	3	2	2	5	2	2	2	2	2	0	0.333333	0.2	0	0.333333	0.2
всероссийский съезд советов	4	4	4	4	3	2	3	4	2	2	1	1.000000	0.7	1	1.000000	0.7
транссибирская магистраль	5	3	2	3	3	2	2	2	2	2	1	0.666667	0.3	1	0.666667	0.4
свидетели иеговы	5	4	3	3	2	2	3	3	2	5	1	1.000000	0.7	1	1.000000	0.7
владимир ильич ульянов	5	1	1	1	2	5	1	1	1	1	1	0.333333	0.2	1	0.333333	0.2
крымский мост	5	5	2	2	2	2	2	2	2	3	1	0.333333	0.2	1	0.666667	0.3
ионообменные смолы	5	4	2	2	3	3	2	3	5	0	1	0.666667	0.5	1	0.666667	0.6
красный террор	5	5	2	2	2	2	3	2	2	2	1	0.666667	0.3	1	0.666667	0.3
аутсайдер моя жизнь как интрига	5	1	1	1	1	1	1	1	1	1	1	0.333333	0.1	1	0.333333	0.1
день защитника отечества	5	3	2	2	2	2	3	2	2	2	1	0.666667	0.3	1	0.666667	0.3
день пограничника	4	4	5	3	2	2	1	1	1	1	1	1.000000	0.4	1	1.000000	0.4
ударный музыкальный инструмент	5	5	1	5	5	5	5	2	5	5	1	0.666667	0.8	1	0.666667	0.8
генеральный штаб вооружённых сил российской федерации	5	5	2	4	5	2	2	3	2	2	1	0.666667	0.5	1	0.666667	0.5
чирлидинг	5	3	3	4	4	2	1	2	3	4	1	1.000000	0.7	1	1.000000	0.7
голштинских заявлений	3	0	4	0	0	0	0	0	0	0	0	0.333333	0.1	1	0.666667	0.2
быстрая сортировка	4	5	4	2	2	2	2	2	1	1	0	0.666667	0.2	1	1.000000	0.3
умберто боччони	5	5	2	2	2	2	2	2	2	5	0	0.000000	0.1	1	0.666667	0.3
охотники за привидениями	5	4	4	3	2	5	2	4	4	4	1	1.000000	0.9	1	1.000000	0.8
божественный пёс	5	2	2	2	2	3	2	2	2	2	1	0.333333	0.2	1	0.333333	0.2
эдди мерфи	5	4	3	5	3	3	3	3	1	3	1	1.000000	0.9	1	1.000000	0.9
бегущий по льду	5	2	2	5	5	5	4	2	4	4	1	0.333333	0.7	1	0.333333	0.7
человек кусает собаку	5	3	2	3	2	3	2	2	2	2	1	0.666667	0.4	1	0.666667	0.4
военный коммунизм	5	3	4	2	4	3	3	3	2	2	1	0.666667	0.6	1	1.000000	0.7
петропавловская крепость	4	4	3	3	2	2	2	3	3	2	1	0.666667	0.5	1	1.000000	0.6
эрмитаж	4	5	2	3	2	2	2	3	3	2	1	0.666667	0.5	1	0.666667	0.5
ссср	5	5	2	2	3	2	3	2	3	2	1	0.333333	0.4	1	0.666667	0.5

Сравнивая результаты можно сказать, что результаты улучшились, вверх пошли те документы, которые содержать больше полезной информации. Некоторые запросы поднять так и не удалось(в данном случае оценка ведется по точности). В них преимущественно различные не несущие смысловой нагрузки части речи. Это можно решить построив дерево синтаксического разбора и с помощью него осуществлять фильтрацию несущественных конструкций(не несущих смысловую нагрузку)