МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)
Факультет информационных технологий и прикладной математики
Кафедра вычислительной математики и программирования

Отчет по лабораторной работе №9
«Зонный поиск»
по курсу
«Информационный поиск»

Группа: М80-108М-17
Выполнил: Забарин Н.И.
Преподаватель: А.Л. Калинин

Москва, 2018

Задание

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 зонам
- 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?

Решение

Добавим индексу коэффициент для TF-IDF и проиндексируем заголовки в отдельный индекс. Что бы отличать файлы разных видов просто изменим префикс имени файла. Поскольку изначально поиск затачивался на множество небольших индексов, то никаких манипуляций с битовым форматом производиться не будет.

Заголовки статей зачастую не содержат всех слов запроса, поэтому по заголовкам имеет смысл выполнять только нечеткий поиск, с довольно высоким коэффициентом.

Влияние на поиск

По сути влиять на поиск по заголовкам я могу только изменяя коэффициент индекса. Поэкспериментировав с этим числом, сделал следующие выводы:

- коэффициент больше 20 ломает результаты поиска и зачастую находит бред с каким нибудь словом в заголовке.
- слишком низкий коэффициент не влияет на поиск.
- адекватное влияние в интервале 10-15, топ 50 результатов меняется не очень сильно, но порядок становится более релевантным.

Сравнение со списком запросов из лабораторной работы №8:

- from which books bible composed (0/0/0/1)
- What? Where? When? (всего один результат)
- game (0/1/2/5)
- russian aircraft factories (0/0/0/0)
- without me the nation are not complete (0/0/0/0)
- name of inhabitants of embankments (0/0/0/0)
- where they crowned Nicholas 2 (1/1/1/1)
- assistant prosecutor (0/0/0/0)
- kings of gas (ничего не найдено)
- administrative code (0/1/2/2)
- organic compounds (1/1/1/1)
- game of thrones (ничего не найдено)
- Sigmund Freud (0/1/2/2)
- Barack Obama (1/3/4/6)
- World War (1/2/2/3)
- Cristiano Ronaldo (1/2/-/-)
- Star Wars (0/1/1/1)
- Indigenous australian (0/0/0/0)
- Web scraping (0/0/0/0)
- Illuminati (1/1/2/3)
- September 11 attacks (почти все результаты вида «List of minor planets: 102001–103000»)
- List of The Big Bang Theory episodes (один странный результат)
- Princess Charlotte of Cambridge (1/1/-/-)
- when is mothers day (0/0/0/0)
- how to solve a rubix cube (1/1/1/1)

- how to write a cover letter (0/0/0/0)
- what is amazon prime (0/0/0/0)
- what is my spirit animal (0/0/0/0)
- what does og mean (0/0/0/0)
- is pluto a planet (ничего не найдено)

Выводы

Поиск по заголовкам с настроенным коэффициентом положительно влияет на запросы состоящие из одного слова-названия, если по этой теме есть множество документов, хороший пример: «Библия». В целом улучшает порядок выдачи, но слабо влияет на состав выдачи.