### 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

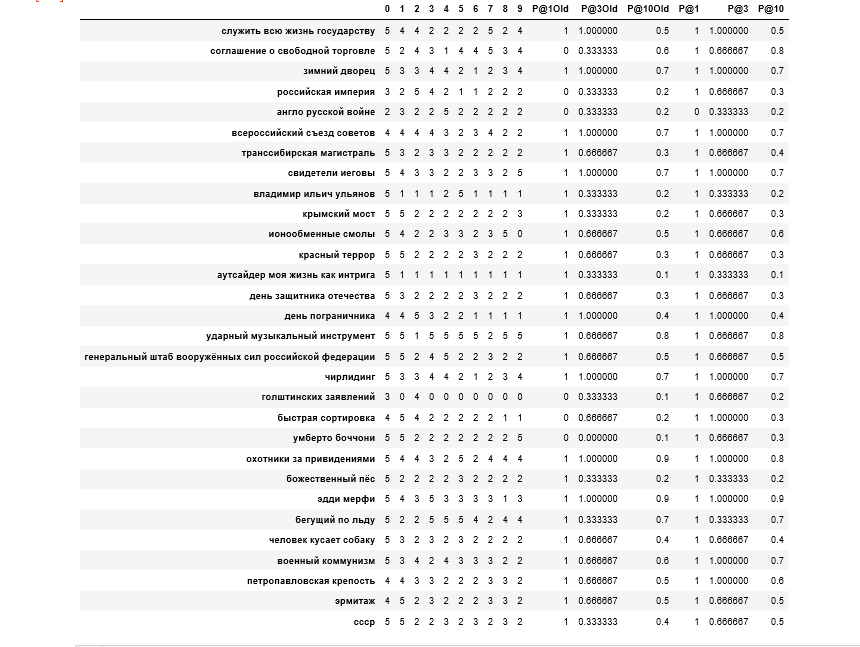
В поиск будет добавлен поиск по зоне заголовков документов. Индекс зоны представляет собой обратный индекс- слова в заголовках являются токенами и им в соответствие ставятся список постингов. Для данной лабораторной работы он вынесен в отдельный файл

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | … |  |  |  |  |  |  |  | … |  |  |
| 4 байта – длина токена-слова, содержащегося в заголовке (N) | | | | N байт - токен | | | 4 байта – длина списка документов (М) | | | | М\*4 байт – список doc\_id документов | | | | |

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

Где Практическим путем было выявлено, что оптимальное значение для этой награды является 0.1.

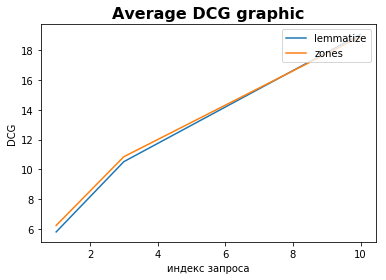
Оценим качество поиска после введенных изменений:











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)