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

Отчет по лабораторной работе №8
«Оценка качества поиска»
по курсу
«Информационный поиск»

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

Москва, 2018

Задание

Необходимо оценить качество поиска, сравнить выдачи с двумя альтернативами. Измерить P@1, P@3, P@5 и P@10, приветствуется использование дополнительных метрик качества.

Придумать и проанализировать 20 запросов отражающих интересы пользователей, определить сильные и слабые стороны поисковой системы, наметить улучшения.

Решение

Оценка будет проводиться на выборке из 1e6 случайных статей. Сравнивать будем с встроенным поиском в Википедию и поиском Google(site:en.wikipedia.org). Считаю такое сравнение некорректным за счет сильного отличия корпусов документов по которым производится поиск, мой корпус меньше в 8,5 раз. Сравнение проводится скриптом, который выполняет один и тот же запрос во всех поисковых системах и сравнивает сколько результатов выдачи моей системы присутствуют в эталонных выдачах, сравниваются топ-100 результатов. Список запросов(P@1/P@3/P@5/P@10):

- from which books bible composed (1/1/1/1)
- What? Where? When? (0/0/0/0)
- game (1/3/4/8)
- russian aircraft factories (0/1/1/1)
- 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 (1/1/1/2)
- assistant prosecutor (0/0/0/0)
- kings of gas (ничего не найдено)
- administrative code (1/1/1/1)
- organic compounds (1/2/3/4)
- game of thrones (ничего не найдено)
- Sigmund Freud (0/1/2/2)
- Barack Obama (1/1/2/3)
- World War (0/0/1/1)
- Cristiano Ronaldo (1/2/2/2)
- Star Wars (1/3/4/4)
- 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 (0/0/0/0)
- when is mothers day (0/0/0/0)
- how to solve a rubix cube (0/0/0/0)
- 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)

Большинство запросов дают мало совпадений, до двух совпадений. Лучшие результаты:

- «from which books bible composed», 6 совпадений,
- «game», 4 совпадения,
- «organic compounds», 4 совпадения,
- «Illuminati», 10 совпадений

Выводы

Лучше всего поисковая система работает с запросами где есть 1-2 слова по которым выполняется поиск. Среди статей в системе остается много «мусорных» документов. Для полноценного сравнения поисковых систем необходимо строить индекс по всей википедии.