МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика»

Лабораторная работа №*7*
по курсу «Информационный поиск»

«Булев индекс»

Выполнил: *Ермаков Ярослав Валерьевич*
Группа: *М8О-407Б-22*
Преподаватели: *А.А. Кухтичев*

Москва, *2025*

Описание

Целью работы является построение обратного индекса, пригодного для булева поиска, по подготовленному корпусу документов. Индекс строится в собственном бинарном формате и предназначен для дальнейшего расширения в следующих лабораторных работах. Помимо обратного индекса формируется прямой индекс, содержащий ссылки на документы и их заголовки, чтобы по идентификаторам документов можно было восстанавливать человекочитаемую выдачу. В качестве термов используются токены после нормализации и, при включённом режиме, после стемминга.

Журнал выполнения и решение проблем

При построении индекса основная сложность заключалась в ограничениях на структуры данных: нельзя использовать ассоциативные контейнеры, поэтому нельзя напрямую “копить” словарь термов в map/unordered_map. Для решения использована внешняя сортировка: из документов формируются пары “терм-doc_id”, далее они сортируются по частям и сливаются в итоговый индекс. На практике возникла проблема согласования идентификаторов документов: метаданные из выгрузки содержали строковый идентификатор, а индексу требовались плотные числовые doc_id. Это было решено приведением метаданных к формату, где doc_id соответствует номеру документа в списке корпуса. После этого индекс успешно строится на полном наборе документов.

Исходный код

Индексатор реализован на C++ и формирует три файла: прямой индекс docs.bin, словарь terms.bin и постинги postings.bin. В docs.bin хранятся URL и заголовки документов, в terms.bin - список термов и смещения на соответствующие постинги, а postings.bin содержит последовательности doc_id для каждого терма. Формат бинарный и расширяемый за счёт заголовков с magic и version. Для построения используется внешняя сортировка: данные разбиваются на несколько “run”-файлов, каждый run сортируется, затем выполняется k-way merge, в ходе которого формируются словарь и постинги.

Выводы

В ходе работы построен булев индекс по корпусу из 58866 документов. В процессе внешней сортировки было создано 10 промежуточных run-файлов, итоговое число уникальных термов составило 376586, а размер файла постингов - 79223688 байт.

Полученный индекс соответствует требованиям: использован собственный бинарный формат, создан прямой индекс для выдачи, обеспечена возможность расширения формата в будущих лабораторных работах.