Семинарское задание №8
Сжатие Индекса

Мосиенко Константин Викторович

2019

Вам даны три программы:

1. dump.py - читает хранилище, которое мы использовали для скачивания википедии, и распечатывает текст статей в стандартный поток вывода.

2. indexer - читает со стандартного потока ввода данные в формате вывода dump.py и строит инвертированный индекс, сохраняя его в файлы в текущей папке.

3. searcher - использует индекс из текущей папки для обработки запросов, которые считывает со стандартного потока ввода.

Вам необходимо:

1. Прочитать код программ и понять, как они работают.

2. Модифицировать indexer и searcher так, чтобы постинг листы сжимались varint-ом.

3. Построить график распределения степеней сжатия постинг листов. Как изменился размер файла index?

4. Привести 10 термов, для которых было достигнуто самое хорошее сжатие. Какие термы сжались хуже всего? Почему? (В данном пункте стоит рассматривать только термы, которые встретились не менее чем в 100 документах)

Мне присылайте код и отчёт.