Отчёт по обратным индексам.

Обратный индекс - это простейшая форма индексации документов, позволяющая выполнять логические запросы к текстовым данным. Например, у нас есть несколько текстовых файлов, и мы хотим выяснить, какие из них содержат все или некоторые термы. Для этого мы разделяем полученную поисковую строку на термы, убираем знаки препинания, и после проходимся по всем документам и ищем эти термы в них, если найдём, то получаем индексы документов. Также в рамках лабораторной работы требуется реализация And поиска (Все термы присутствуют в документе), Or (Хотя бы одна терма из запроса в документе), Not (ID документов, в которых нет ни одного терма из запроса). Все данные документов (Содержание и ID) будут сжаты алгоритмом Хаффмена.

Строки будут загружаться из csv файла random\_string.csv, после в коде будет проводиться каждый вид поиска с результатом, а время работы засекаться объектами DataTime и записываться в файл Inverted\_Index\_time.txt.

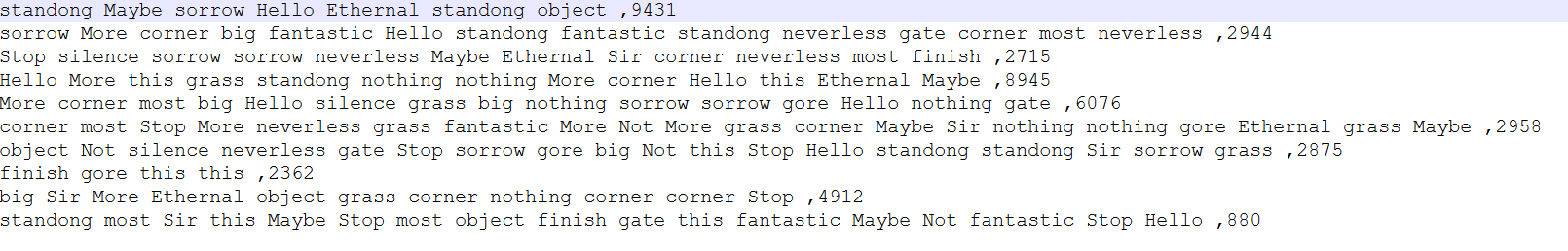


Рисунок 1 - Пример входных данных.

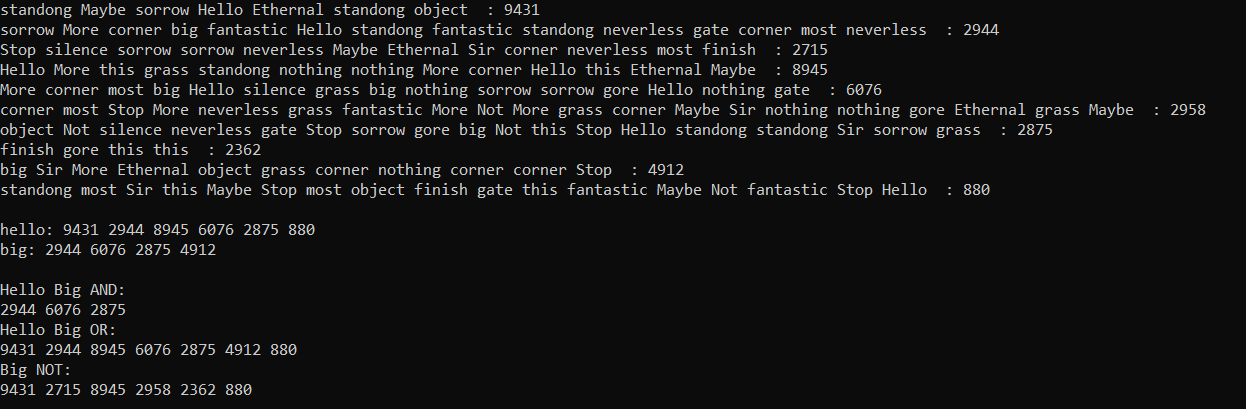
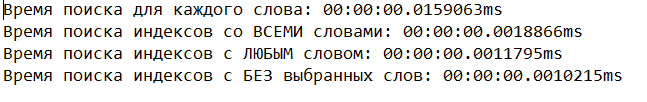
В классе InvertedIndex реализованы поля all\_documents (хранит документы в зашифрованном виде) и trees\_of\_documents (Хранят деревья документов, чтобы кодировать и декодировать их). В методе Add создаются деревья и кодируются ID и строки, после записываются в поля. Метод SearchEachWord ищет для каждого слова из за проса индексы документов и выводит их в словарик Dictionary<string, List<int>>. SearchAnd ищет документы со всеми термами, SearchOr если есть хоть одна терма, SearchNot если нет в документе терм

Рисунок 2 - Пример вывода программыРисунок 3 - Пример времени работы поиска.

Реализация сжатия Хаффмена приведена в отдельном файле класса HuffmanCompression.cs.