**О Т Ч Е Т**

о выполнении лабораторной работы № 2

по дисциплине «Распределенная обработка информации»

Вариант 14

Выполнил:

Чернов А.А. \_\_\_\_\_\_\_\_\_\_

(ПОДПИСЬ)

Проверил:

Мальцев Р.В. \_\_\_\_\_\_\_\_\_\_

(ПОДПИСЬ)

МОСКВА 2019

Оглавление

[1 Задание 3](#_Toc22561679)

[2 Теоретическая справка 4](#_Toc22561680)

[3 Спецификации разработанных программ 6](#_Toc22561681)

[3.1 Программное средство «Генератор» 6](#_Toc22561682)

[3.2 Программное средство «Однопоточное приложение» 6](#_Toc22561683)

[4 Алгоритм решения задачи 8](#_Toc22561684)

[5 Тестирование программных средств 11](#_Toc22561685)

[5.1 Проверка корректности реализованных приложений 11](#_Toc22561686)

[5.2 Проведение экспериментов 14](#_Toc22561687)

[5.3 Результаты экспериментов 15](#_Toc22561688)

[5.4 Анализ полученных результатов 17](#_Toc22561689)

# Задание

Для заданного каталога построить индекс файлов, представляющий собой отсортированный массив пар (<символ>; <полное имя файла>). Где <символ> – любой из символов встречающийся в сокращенном имени файла (т.е. имени без пути к нему). Подумать над вариантом алгоритма формирующим пары <подстрока>; <полное имя файла>), где максимальная длина подстроки задается при запуске алгоритма построения. Также подумать над вариантом формирования пар со всеми встречающимися в именах файлов вариантами подстрок. Построить индекс файлов виде древовидной структуры.

# Теоретическая справка

Поисковый индекс — структура данных, которая содержит информацию о документах и используется в поисковых системах. Индекси́рование, совершаемое поисковой машиной — процесс сбора, сортировки и хранения данных с целью обеспечить быстрый и точный поиск информации. Создание индекса включает междисциплинарные понятия из лингвистики, когнитивной психологии, математики, информатики и физики. Веб-индексированием называют процесс индексирования в контексте поисковых машин, разработанных, чтобы искать веб-страницы в Интернете.

Популярные поисковые машины сосредотачиваются на полнотекстовой индексации документов, написанных на естественных языках. Мультимедийные документы, такие как видео и аудио и графика, также могут участвовать в поиске.

Метапоисковые машины используют индексы других поисковых сервисов и не хранят локальный индекс, в то время как поисковые машины, основанные на кешированных страницах, долго хранят как индекс, так и текстовые корпусы. В отличие от полнотекстовых индексов, частично-текстовые сервисы ограничивают глубину индексации, чтобы уменьшить размер индекса. Большие сервисы, как правило, выполняют индексацию в заданном временном интервале из-за необходимого времени и затрат на обработку, в то время как поисковые машины, основанные на агентах, строят индекс в масштабе реального времени.

# Спецификации разработанных программ

## Программное средство «Многопоточное приложение»

Пользовательский интерфейс созданного программного средства реализован следующим образом:

multithread --input inputFilesName --output outputFileName.csv

где multithread – имя исполняемого файла разработанной программы; --input – ключ, определяющий следующий аргумент командной строки как общую часть для имени группы входных файлов, обрабатываемых данных; inputFilesName – общая часть для имени группы входных файлов; --output – ключ, определяющий следующий аргумент командной строки как имя выходного файла с результатами измерений в формате CSV; outputFileName.csv – имя выходного файла. Кроме перечисленных допускается наличие других аргументов командной строки, они должны следовать после обязательных аргументов, перечисленных выше.

# Алгоритм решения задачи

Программа multithread требует дополнительные параметры – конфигурационный файл и длина подстрок, на которые необходимо разбить имена входных файлов. Ее результатом является набор выходных файлов с результатами разбиения и файл, содержащий время обработки каждого из входных файлов.

Программа реализована следующим образом. В начале работы при помощи функции std::thread::hardware\_concurrency() определяется число ядер компьютера, на котором запущена программа. Затем общий список полных путей файлов для создания индекса делится на cores частей (где cores - количество ядер процессора). Каждая порция данных обрабатывается параллельно и независимо функциями build\_subline\_index и build\_symbol\_index. По окончанию обработки словари, полученные в результате работы функций, сливаются в один общий (для строкового и символьного индексов - соответственно). Для того, чтобы не вызывать конструктор копирования контейнеров при вызове каждой функции, организована передача по-ссылке.

# Тестирование программных средств

## Проверка корректности реализованных приложений

Были сравнены выходные значения обработки одних и тех же данных при помощи monothread и multithread. Результаты соответствуют ожиданиям.

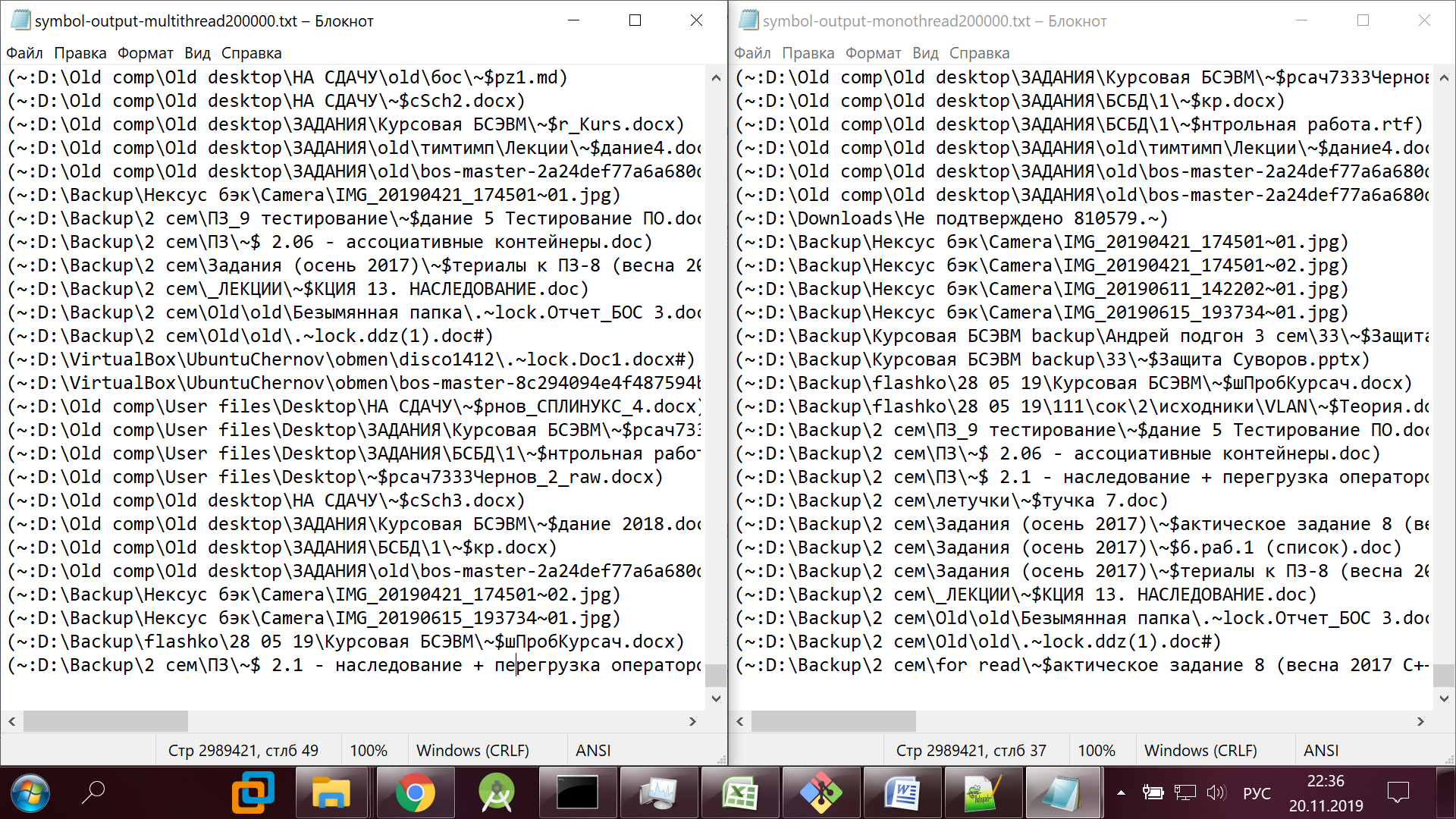


Рисунок 1 – входные monothread и multithread символьный индекс

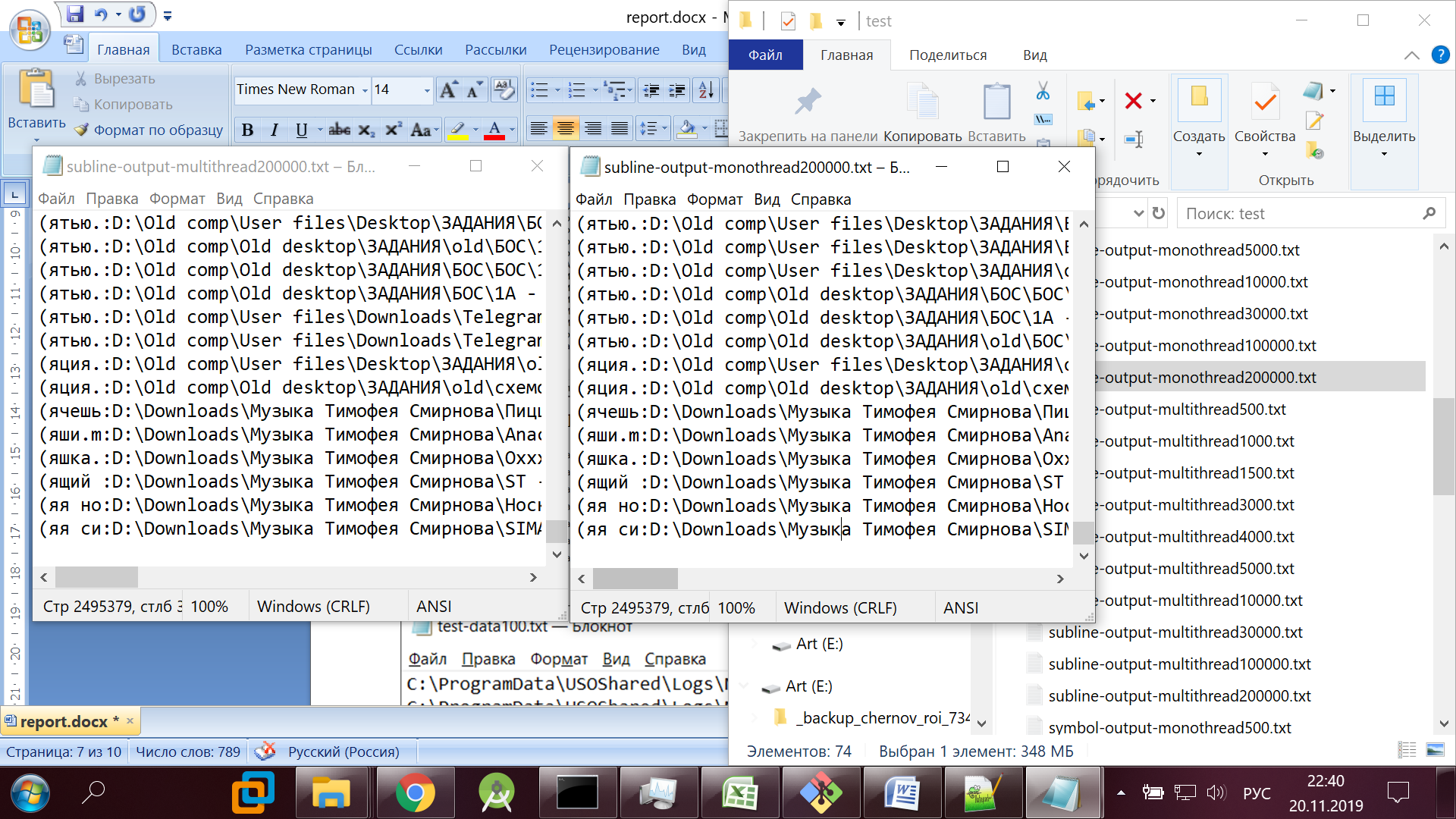


Рисунок 2 – входные monothread и multithread строковый индекс

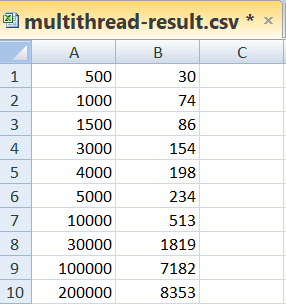


Рисунок 3 – время обработки данных

## Проведение экспериментов

Был проведен эксперимент для большого количества данных. Для программы generator был указан путь «D:/» и сгенерирован большой объем входных данных.

## Результаты экспериментов

Результат выполнения monothread и обработки данных был представлен в таблице 1.

|  |  |
| --- | --- |
| Количество строк | Время (мс) |
| 500 | 30 |
| 1000 | 74 |
| 1500 | 86 |
| 3000 | 154 |
| 4000 | 198 |
| 5000 | 234 |
| 10000 | 513 |
| 30000 | 1819 |
| 100000 | 7182 |
| 200000 | 8353 |

Таблица 1 – результат эксперимента

## Анализ полученных результатов

В ходе лабораторной работы были реализованы генерация и обработка больших объемов данных. На скорость обработки влияет загруженность компьютера другими процессами и общее количество программ, запущенных в данный момент. Для ускорения обработки возможно использование библиотеки Qt с дальнейшей оптимизацией.

На графике наблюдается близкая к логарифмической зависимость. На графике наблюдается, что при больших объемах данных заметна эффективность многопоточного алгоритма. На маленьких объемах расхождения несущественны из-за необходимости слияния результатов обработки. Временные результаты могут отличаться, в зависимости от длины подстрок разбиения и заданной директории для построения индекса.

Рисунок 4 – оценка сложности