МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ ИМ. Н. Э. БАУМАНА

УДК ______
№ госрегистрации ______
Инв. №______

УТВЕРЖДАЮ

Преподаватель

____

«______» _____________________ 2019 г.

ДИСЦИПЛИНА АНАЛИЗ АЛГОРИТМОВ
ОТЧЁТ
ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

Расстояние Левенштейна
(промежуточный)

Студент __________________ Ф.М. Набиев

Преподаватели Л.Л. Волкова, Ю.В. Строганов

Москва, 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Целью данной работы является изучение динамического программирования на материале алгоритмов поиска расстояний Левенштейна и Дамерау-Левенштейна.

Данные алгоритмы решают проблему нахождения редакционного расстояния между двумя строками. Редакционное расстояние определяется количеством некоторых операций, необходимых для превращения одного слова в другое, а так же стоимостью этих операций.

Для достижения поставленной цели необходимо решить следующие задачи:

— изучение алгоритмов Левенштейна и Дамерау-Левенштейна нахождения расстояния между строками;

— применение метода динамического программирования для матричной реализации указанных алгоритмов;

— получение практических навыков реализации указанных алгоритмов: двух алгоритмов в матричной версии и одного из алгоритмов в рекурсивной версии;

— сравнительный анализ линейной и рекурсивной реализаций выбранного алгоритма определения расстояния между строками по затрачиваемым ресурсам (времени и памяти);

— экспериментальное подтверждение различий во временной эффективности рекурсивной и нерекурсивной реализаций выбранного алгоритма определения расстояния между строками при помощи разработанного программного обеспечения на материале замеров процессорного времени выполнения реализации на варьирующихся длинах строк;

— описание и обоснование полученных результатов в отчете о выполненной лабораторной работе, выполненного как расчётно-пояснительная записка к работе.

1 Аналитический раздел

В данном разделе будет определена теоретическая база, необходимая для реализации поставленных задач.

1.1 Описание алгоритмов

Данные алгоритмы основываются на применении формул Левенштейна и Дамерау-Левенштейна. Рассмотрим эти формулы подробнее.

1.1.1 Расстояния Левенштейна

Расстояние Левенштейна между двумя строками - это минимальная сумма произведений количества операций вставки, удаления и замены одного символа, необходимых для превращения одной строки в другую, на их стоимость.

Вышеописанные операции имеют следующие обозначения:

— I ($insert$) - вставка;

— D ($delete$) - удаление;

— R ($replace$) - замена;

При этом $cost(x)$ есть обозначение стоимости некоторой операции x. Будем считать, что символы в строках нумеруются с первого. Пусть S_1 и S_2 - две строки с длинами N и M соответственно. Тогда расстояние Левенштейна D(M, N) вычисляется по формуле (1.1):

$$D(i,j) = \begin{cases} 0, & i = 0, j = 0 \\ i * cost(D), & j = 0, i > 0 \\ j * cost(I), & i = 0, j > 0 \\ min(& \\ D(i,j-1) + cost(I), & \\ D(i-1,j) + cost(D), & j > 0, i > 0 \\ D(i-1,j-1) + mrcost(S_1[i], S_2[j]) & \\ ) & \end{cases} \tag{1.1}$$

где $min(a, b, c)$ возвращает наименьшее значение из a, b, c; а $mrcost(x_1, x_2)$ - 0, если символы x_1, x_2 совпадают, и $cost(R)$ иначе.

1.1.2 Расстояние Дамерау-Левенштейна

Определение расстояния Дамерау-Левенштейна аналогично определению расстояния Левенштейна с учётом новой операции - перестановки соседних символов (транспозиции). Соответственно, обозначения операций:

— I ($insert$) - вставка;

— D ($delete$) - удаление;

— R ($replace$) - замена;

— T ($transpose$) - перестановка соседних символов.

При тех же обозначениях имеем формулы (1.2) и (1.3):

$$D(i,j) = \begin{cases} min(A, D(i-2, j-2) + cost(T), & i > 1, j > 1, \\ & S_1[i] = S_2[j-1], \\ & S_1[i-1] = S_2[j] \\ A & \text{Иначе} \end{cases} \tag{1.2}$$

где A:

$$A = \begin{cases} 0, & i = 0, j = 0 \\ i * cost(D), & j = 0, i > 0 \\ j * cost(I), & i = 0, j > 0 \\ min(& \\ D(i,j-1) + cost(I), & \\ D(i-1, j) + cost(D), & j > 0, i > 0 \\ D(i-1, j-1) + mrcost(S_1[i], S_2[j]) & \\ ) & \end{cases} \tag{1.3}$$

1.2 Вывод

Очевидно, формулы Левенштейна и Дамерау-Левенштейна имеют различную прикладную направленность. Если вторая рассчитана больше на слова, набранные человеком, то первая - нет, так как транспозиция фактически не является тривиальной операцией, и её наличие это условность, требуемая контекстом применения.

2 Конструкторский раздел

В данном разделе будет проведена конкретизация поставленных задач, составлены и проанализированы алгоритмы.

2.1 Модель

IDEF0 модель задачи вычисления редакционного расстояния приведена на рисунке 2.1.

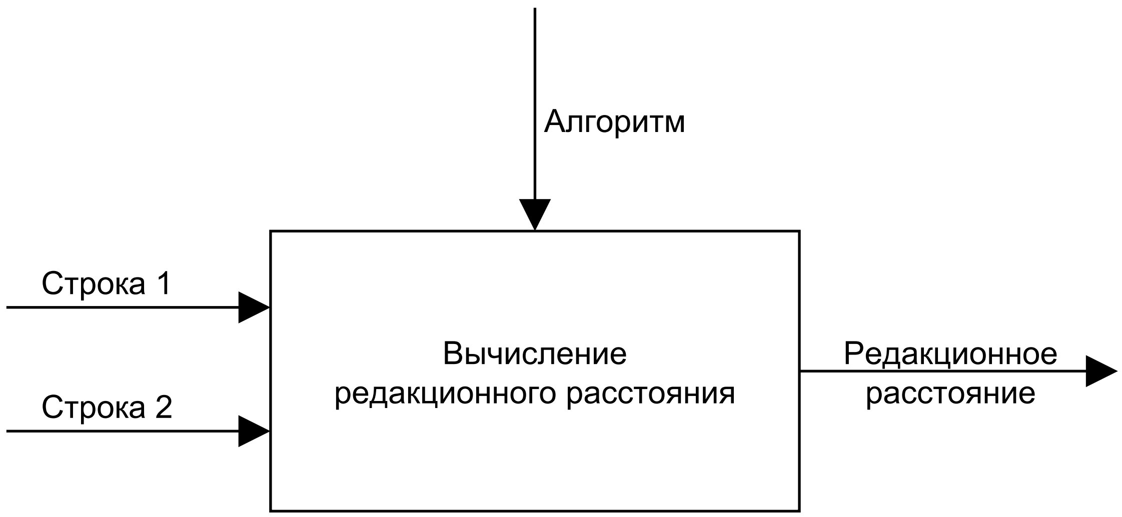

Рисунок 2.1 — IDEF0 модель

2.2 Разработка алгоритмов

Для непосредственной реализации вышеописанных алгоритмов важно иметь их некоторые упрощённые визуальные представления, так как чтение таких представлений упрощает написание кода. Подходящим для этого вариантом визуализации являются схемы алгоритмов.

2.2.1 Алгоритм Вагнера-Фишера

Алгоритм Вагнера-Фишера является матричной реализацией поиска расстояния Левенштейна. Схема данного алгоритма приведена на рисунках 2.2 и 2.3.

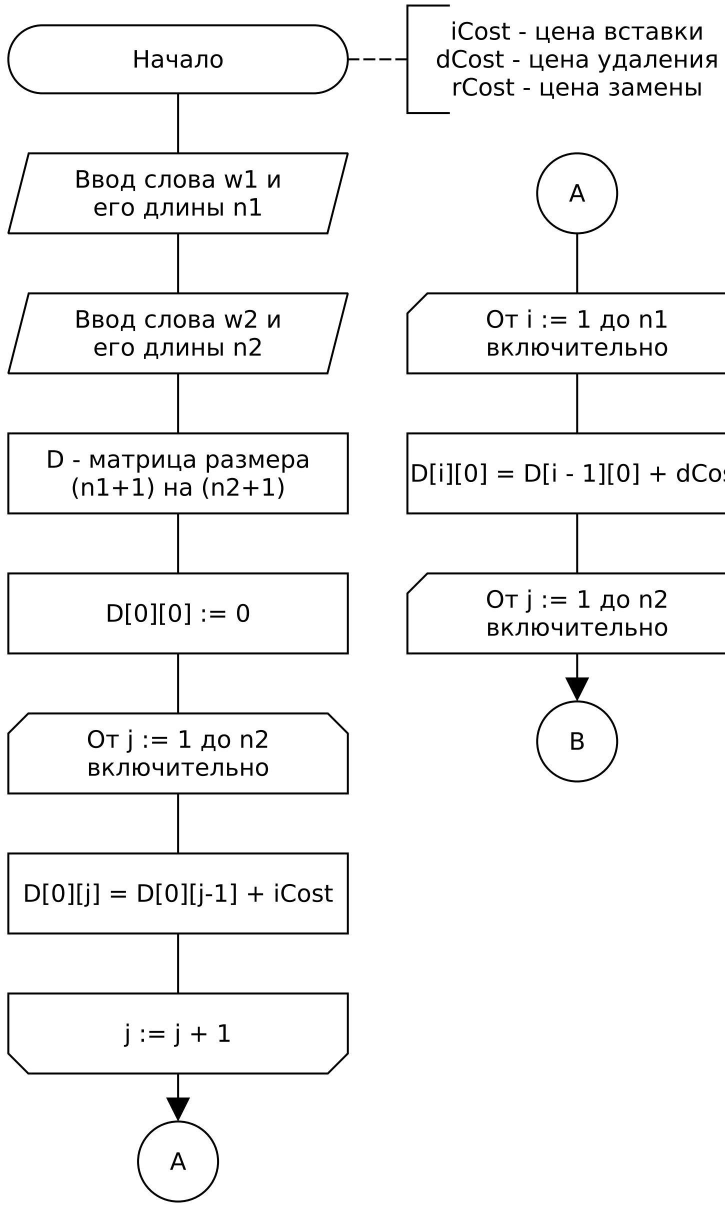

Рисунок 2.2 — Алгоритм Вагнера-Фишера, часть 1

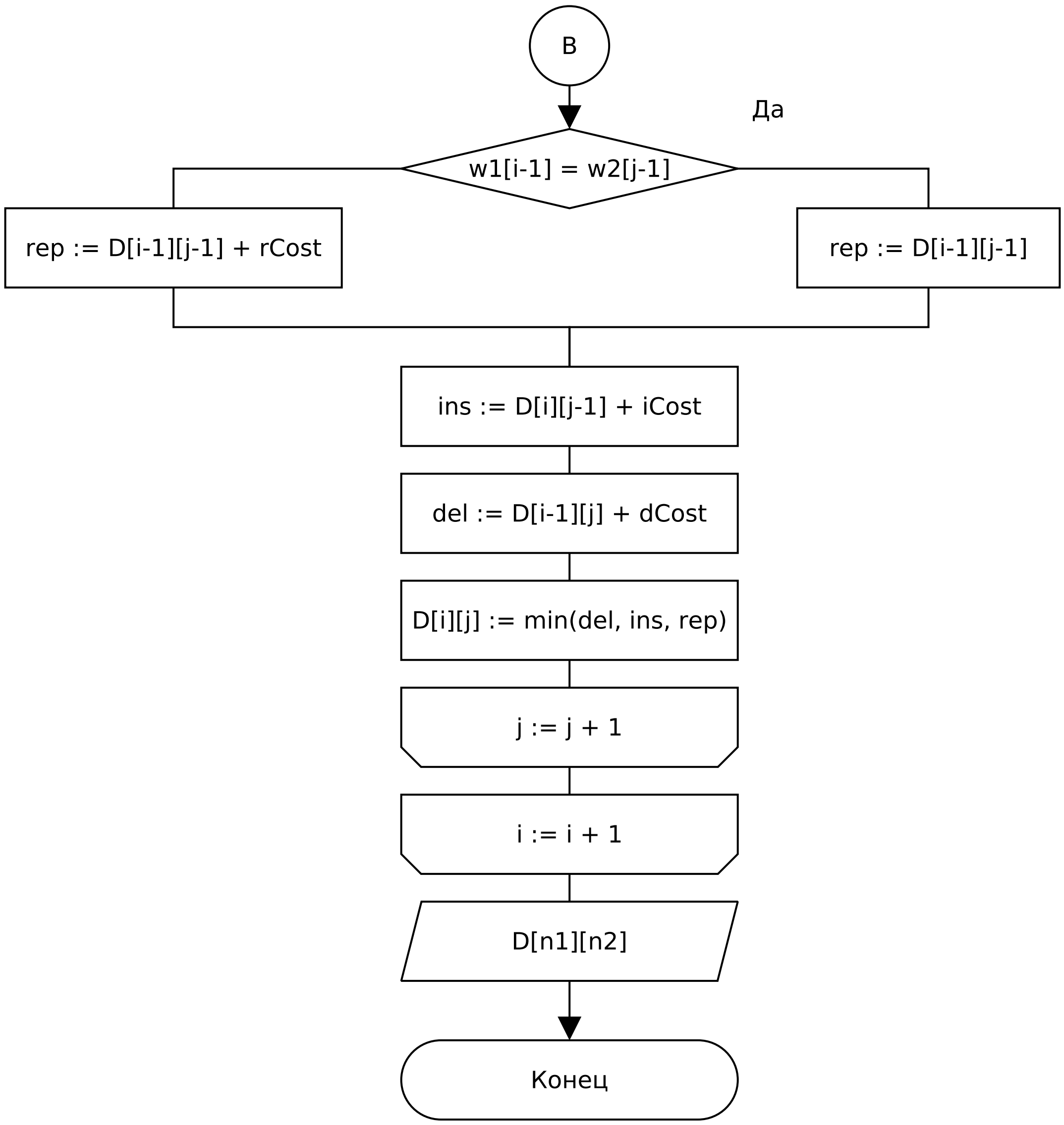

Рисунок 2.3 — Алгоритм Вагнера-Фишера, часть 2

2.2.2 Матричный алгоритм Дамерау-Левенштейна

Матричный алгоритм Дамерау-Левенштейна представляет из себя модификацию алгоритма Вагнера-Фишера, в котором происходит дополнительная проверка на возможность проведения операции транспозиции. Схема данного алгоритма приведена на рисунках 2.4 и 2.5

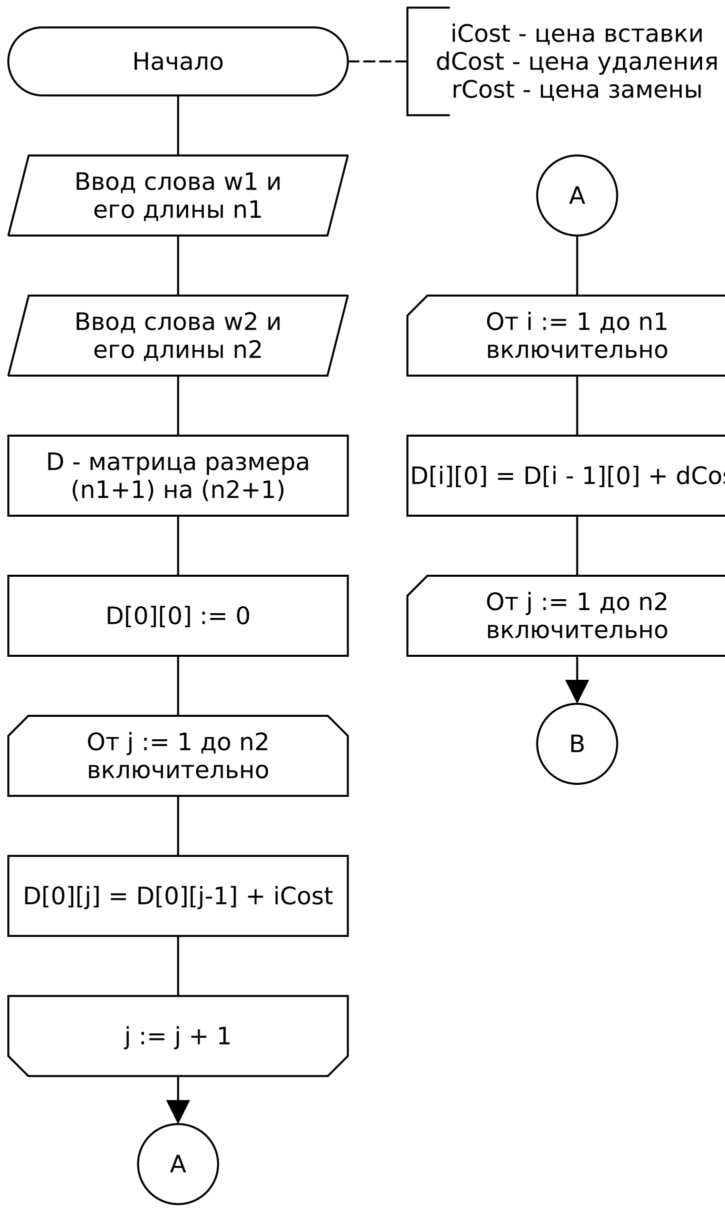

Рисунок 2.4 — Матричный алгоритм Дамерау-Левенштейна, часть 1

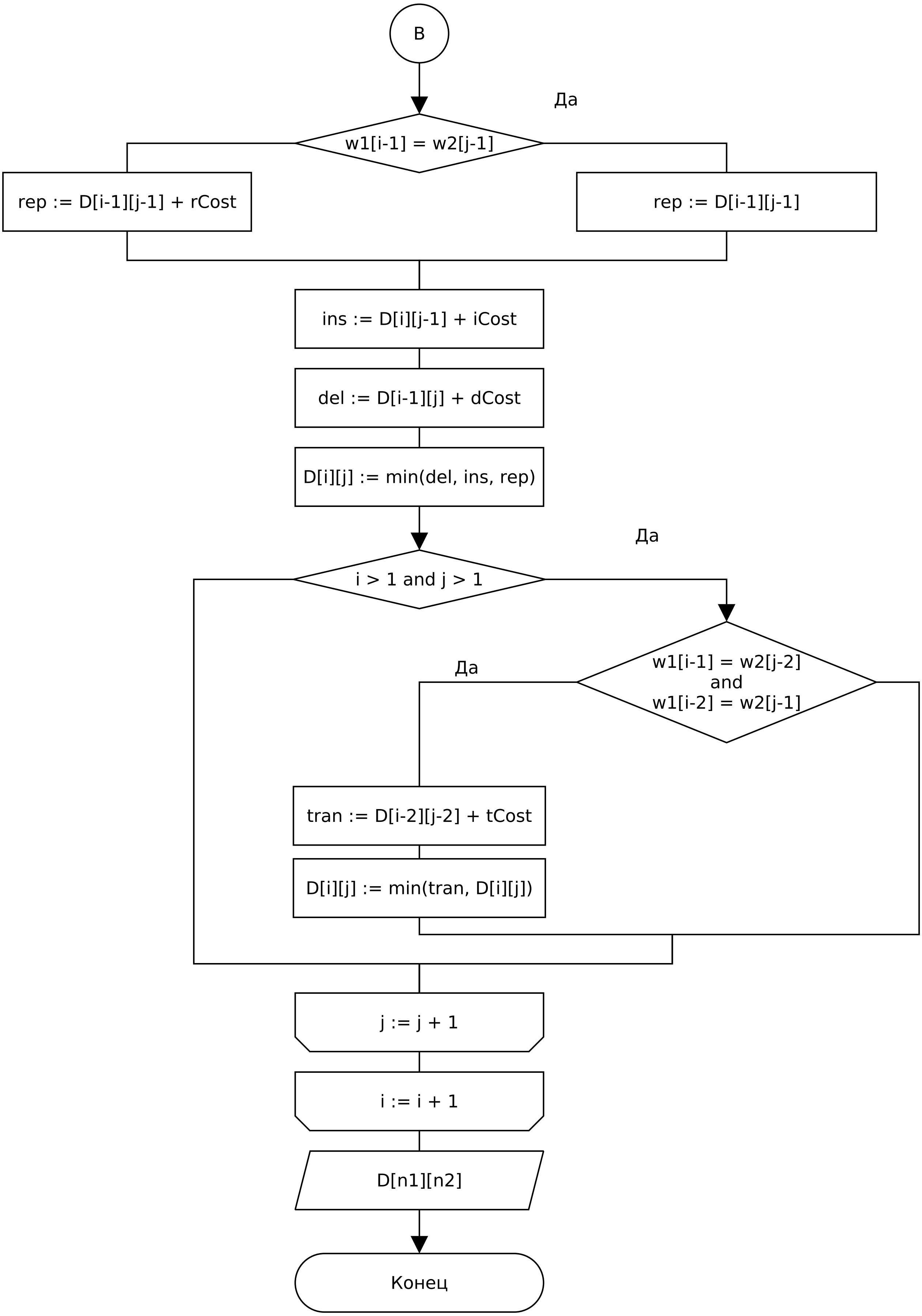

Рисунок 2.5 — Матричный алгоритм Дамерау-Левенштейна, часть 2

2.2.3 Рекурсивный алгоритм Дамерау-Левенштейна

Суть рекурсивного алгоритма Левенштейна состоит в сведении поиска редакционного расстояния до тривиального случая, когда длина хотя бы одного из слов равна 0. Отличие алгоритма Левенштейна от алгоритма Дамерау-Левенштейна было описано ранее. Схема данного алгоритма приведена на рисунках 2.6, 2.7 и 2.8.

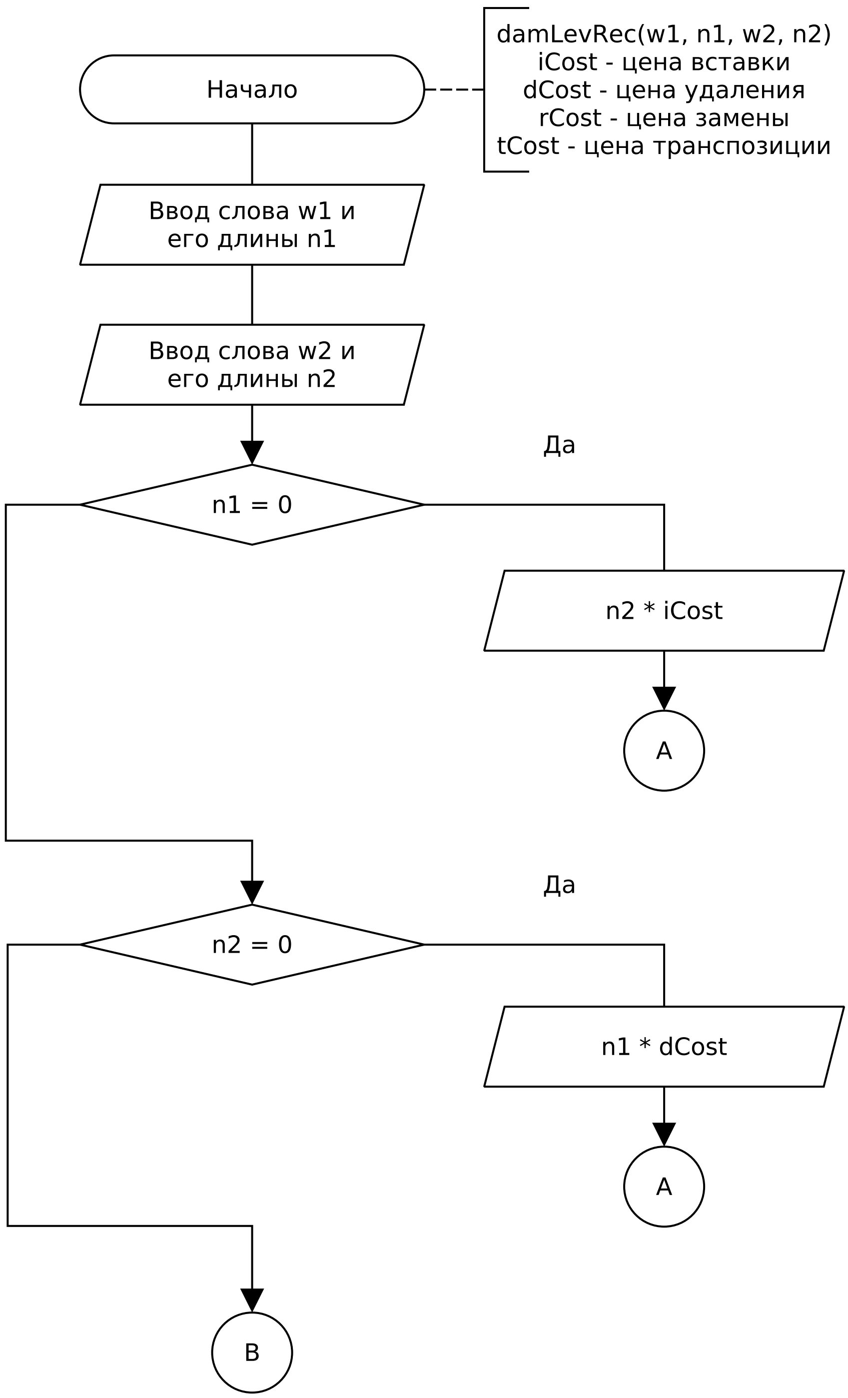

Рисунок 2.6 — Рекурсивный алгоритм Дамерау-Левенштейна, часть 1

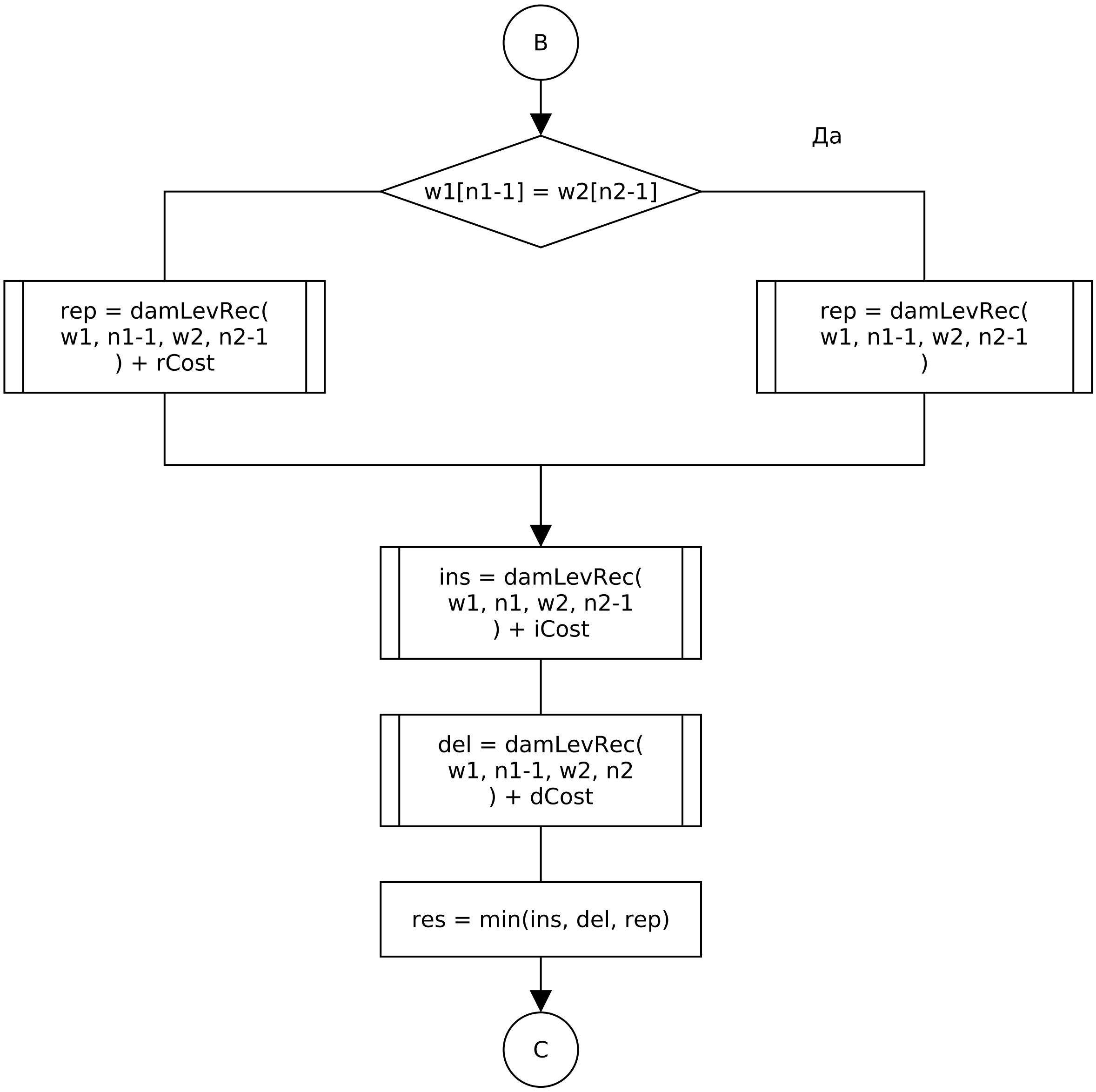

Рисунок 2.7 — Рекурсивный алгоритм Дамерау-Левенштейна, часть 2

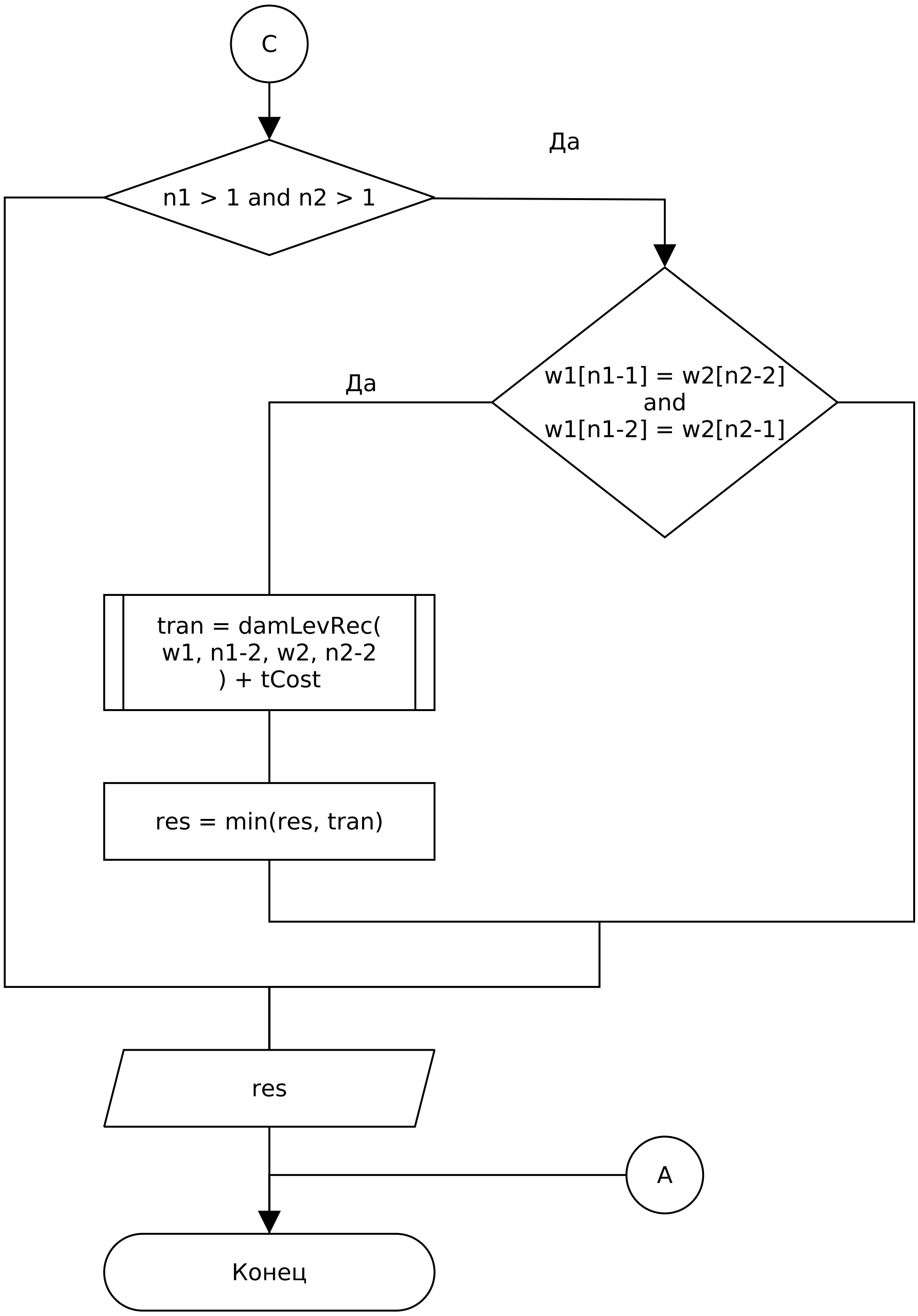

Рисунок 2.8 — Рекурсивный алгоритм Дамерау-Левенштейна, часть 3

2.3 Сравнительный анализ реализаций

Для поиска редакционного расстояния можно применять как рекурсивные алгоритмы, так и матричные. Чтобы сделать вывод об эффективности того или иного алгоритма, проведём их анализ. Для упрощения задачи рассмотрим только алгоритмы поиска расстояния Левенштейна, так как объём затрачиваемых ресурсов этими алгоритмами прямо коррелирует с объёмом ресурсов, затрачиваемых алгоритмами поиска расстояния Дамерау-Левенштейна.

2.3.1 Оценка сложности

Произведем оценку общей сложности рекурсивного алгоритма. В рассмотренной реализации присутствуют 3 точки входа в рекурсию, условием выхода из рекурсии является равенство длины хотя бы одной из строк нулю. Из этого можно сделать вывод, что рекурсивный алгоритм Левенштейна в худшем случае имеет общую сложность $O(3^n)$, где n - максимальная длина обрабатываемых слов.

Что касается матричной реализации, её задача сводится к полному обходу матрицы размера $(n_1+1)\cdot(n_2+1)$, где n_1 и n_2 - длины обрабатываемых строк. Следовательно, данный алгоритм имеет сложность $O(n_1 \cdot n_2)$.

2.3.2 Оценка памяти

Память, затрачиваемая на выполнение рассматриваемых алгоритмов зависит от используемых типов данных и соглашении о вызовах. В качестве примера, будем считать, что обрабатываемые строки передаются в функции по указателю, размер одного символа составляет 1 байт, целочисленного типа - 4 байта, указателя - 8 байт, используется cdecl (параметры функции передаются через стек, в который так же помещаются значения адреса возврата и указателя на верхушку текущего стекового кадра).

Как было показано ранее, в рекурсивной реализации происходит $3^n+3^{n-1}+...+3^0$ вызовов функции в худшем случае. Кроме того, важно то, что схема вызовов рекурсивного алгоритма имеет древовидную структуру, в которой для того же худшего случая глубина дерева достигает $log_3(3^n) = n$. При этом вызовы, находящиеся на одном уровне дерева, не могут обрабатываться одновременно, то есть для всех вызовов одного уровня необходимо столько памяти, сколько нужно одному такому вызову. Допустим, что в аргументах функции передаются два указателя на обрабатываемые строки и 2 целочисленные переменные, означающие длины этих строк. Без учета возможного использования локальных переменных, имеем следующие затраты памяти:

$$n \cdot (8 + 8 + 8 + 4 + 8 + 4) = 40 \cdot n \tag{2.1}$$

В случае матричного алгоритма Дамерау-Левенштейна будем считать, что функция использует матрицу размера $(n_1 + 1) \cdot (n_2 + 1)$, указатель на начало этой матрицы и два целочисленных счётчика для циклов. Тогда имеем следующие затраты памяти:

$$\begin{aligned} &40 + (n_1 + 1) \cdot (n_2 + 1) \cdot 4 + 8 + 4 + 4 = \\ &56 + 4 \cdot (n_1 + 1) \cdot (n_2 + 1) \end{aligned} \tag{2.2}$$

2.4 Вывод

Таким образом, можно сделать вывод о том, что матричная реализация алгоритма поиска расстояния Левенштейна работает гораздо быстрее, чем рекурсивная, но и потребляет памяти так же гораздо больше.

3 Технологический раздел

В данном разделе будут составлены требования к программному обеспечению, выбраны средства реализации и определены тестовые данные.

3.1 Требования к программному обеспечению

Требования к вводу:

— на вход подаются два слова;

— каждое слово завершается символом переноса строки;

— пробел может быть в составе слова;

— одна и та же буква в верхнем и нижнем регистрах считается как разные символы;

— пустое слово допускается.

Требования к выводу:

— редакционное расстояние;

— в случае матричного алгоритма выводить матрицу, полученную в ходе вычисления расстояния.

Требования к программе:

— выбор алгоритма происходит через аргументы командной строки путём передачи его номера:

1) алгоритм Вагнера-Фишера;

2) матричный алгоритм Дамерау-Левенштейна;

3) рекурсивный алгоритм Дамерау-Левенштейна.

3.2 Средства реализации

Для реализации программы вычисления редакционного расстояния мной был выбран язык программирования C++. В рамках текущей задачи данный язык программирования имеет ряд существенных преимуществ:

— Статическая типизация;

— Близость к низкоуровневому C при наличии многих возможностейвысоко уровненных языков;

— Встроенная библиотека std::chrono, позволяющая измерять процессорное время.

3.3 Листинги кода

Реализации алгоритмов Вагнера-Фишера, матричного и рекурсивного Дамерау-Левенштейна приведены в листингах 3.1, 3.2, 3.3 соответственно.

Листинг 3.1 — Расстояние Левенштейна (матричная реализация)

```
template<typename _Word_t>
int WagnerFischer<_Word_t>::distance(_Word_t w1, int n1,
                                     _Word_t w2, int n2) {
    int **D = Util::createMatrix<int>(n1 + 1, n2 + 1);
    D[0][0] = 0;

    for (int j = 1; j <= n2; j++) {
        D[0][j] = D[0][j - 1] + insertCost;
    }

    for (int i = 1; i <= n1; i++) {
        D[i][0] = D[i - 1][0] + deleteCost;

        for (int j = 1; j <= n2; j++) {
            int replaceCost = this->replaceCost;

            if (w1[i - 1] == w2[j - 1]) {
                replaceCost = 0;
            }

            D[i][j] = std::min(std::min(D[i - 1][j] + deleteCost,
                                        D[i][j - 1] + insertCost),
                               D[i - 1][j - 1] + replaceCost);
        }
    }

    int res = D[n1][n2];
    delete[] D;

    return res;
}
```

Листинг 3.2 — Расстояние Дамерау-Левенштейна (матричная реализация)

```
template<typename _Word_t>
int DamerauLevenshtein<_Word_t>::distance(_Word_t w1, int n1,
                                          _Word_t w2, int n2) {
    int **D = Util::createMatrix<int>(n1 + 1, n2 + 1);
    D[0][0] = 0;

    for (int j = 1; j <= n2; j++) {
        D[0][j] = D[0][j - 1] + insertCost;
    }

    for (int i = 1; i <= n1; i++) {
        D[i][0] = D[i - 1][0] + deleteCost;

        for (int j = 1; j <= n2; j++) {
            int replaceCost = this->replaceCost;

            if (w1[i - 1] == w2[j - 1]) {
                replaceCost = 0;
            }

            D[i][j] = std::min(std::min(D[i - 1][j] + deleteCost,
                                        D[i][j - 1] + insertCost),
                               D[i - 1][j - 1] + replaceCost);

            if (i > 1 && j > 1) {
                if (w1[i - 1] == w2[j - 2] && w1[i - 2] == w2[j - 1]) {
                    D[i][j] = std::min(D[i - 2][j - 2] + transposeCost,
                                       D[i][j]);
                }
            }
        }
    }

    int res = D[n1][n2];
    delete[] D;

    return res;
}
```

Листинг 3.3 — Расстояние Дамерау-Левенштейна (рекурсивная реализация)

```
template<typename _Word_t>
int DamerauLevenshteinRecursive<_Word_t>::distance(_Word_t w1, int n1,
                                                     _Word_t w2, int n2) {
    if (n1 == 0) {
        return n2 * insertCost;
    }
    if (n2 == 0) {
        return n1 * deleteCost;
    }

    bool isSame = w1[n1 - 1] == w2[n2 - 1];

    int res = std::min(std::min(distance(w1, n1 - 1, w2, n2) + deleteCost,
                                distance(w1, n1, w2, n2 - 1) + insertCost),
                       distance(w1, n1 - 1, w2, n2 - 1) + replaceCost *
                                                         !isSame);

    if (n1 > 1 && n2 > 1) {
        if (w1[n1 - 1] == w2[n2 - 2] && w1[n1 - 2] == w2[n2 - 1]) {
            res = std::min(distance(w1, n1 - 2, w2, n2 - 2) +
                           transposeCost, res);
        }
    }

    return res;
}
```

3.4 Описание тестирования

Для тестирования программы были подготовлены данные, представленные в таблице 3.1.

Таблица 3.1 — Тестовые данные

№	Строка 1	Строка 2	Ожидаемое расстояние Левенштейна	Ожидаемое расстояние Дамерау-Левенштейна
1	some	any	4	4
2		nothing	7	7
3			0	0
4	bashrc	bashcr	2	1
5	bus	BuS	2	2
6	electricity	city	7	7
7	powerful	powerless	4	4
8	grow	flow	2	2
9	rise	rice	1	1
10	legal	illegal	2	2
11	same	same	0	0

3.5 Вывод

Таким образом, были сформулирован требования к реализуемой программе, выбран язык программирования C++ в качестве основного инструмента разработки, подготовлены тестовые данные для проверки корректности работы итоговой программы.

4 Исследовательский раздел

В данном разделе будет продемонстрирована и проанализирована работа разработанной программы поиска редакционного расстояния на основе формул Левенштейна и Дамерау-Левенштейна.

4.1 Примеры работы

Рассмотрим примеры работы программы. На рисунках 4.1, 4.2, 4.3, 4.4 представлены случаи с двумя переставленными соседними буквами, пропущенной буквы, пустыми словами, и полностью разными словами соответственно.

```
# ./fn_aa_lab_01 1
bashrc
bashcr
0 1 2 3 4 5
1 0 1 2 3 4
2 1 0 1 2 3
3 2 1 0 1 2
4 3 2 1 0 1
5 4 3 2 1 1
2
```

```
# ./fn_aa_lab_01 2
bashrc
bashcr
0 1 2 3 4 5
1 0 1 2 3 4
2 1 0 1 2 3
3 2 1 0 1 2
4 3 2 1 0 1
5 4 3 2 1 1
1
```

```
# ./fn_aa_lab_01 3
bashrc
bashcr
1

  ~/Documents/Repos:
/bmstu/AlgoesAnalys:
01/build  master !
 #
```

Рисунок 4.1 — Транспозиция

```
# ./fn_aa_lab_01 1
lab1
lab
0 1 2
1 0 1
2 1 0
3 2 1
1
```

```
# ./fn_aa_lab_01 2
lab1
lab
0 1 2
1 0 1
2 1 0
3 2 1
1
```

```
# ./fn_aa_lab_01 3
lab1
lab
1

  ~/Documents/Repos
/bmstu/AlgoesAnalys
01/build  master !
```

Рисунок 4.2 — Пропуск одной буквы

```
# ./fn_aa_lab_01 1


0
```

```
# ./fn_aa_lab_01 2


0
```

```
# ./fn_aa_lab_01 3


0
```

Рисунок 4.3 — Пустые слова

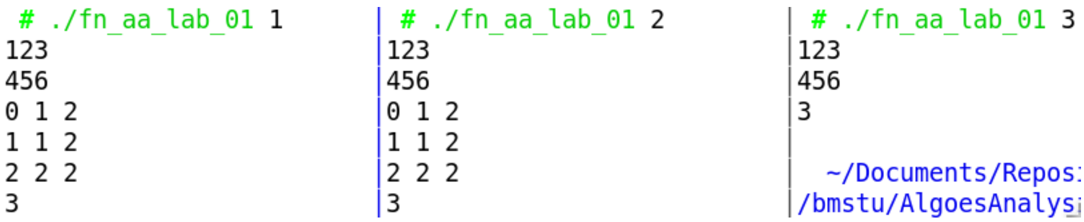

Рисунок 4.4 — Совершенно разные слова

4.2 Результаты тестирования

Тестирование всех трёх реализаций алгоритмов прошло успешно. Результаты тестов представлены в таблицах 4.1, 4.2, 4.3.

Таблица 4.1 — Результаты тестирования алгоритма Вагнера-Фишера

№	Строка 1	Строка 2	Расстояние Левенштейна	Ожидаемое расстояние Левенштейна
1	some	any	4	4
2		nothing	7	7
3			0	0
4	bashrc	bashcr	2	2
5	bus	BuS	2	2
6	electricity	city	7	7
7	powerful	powerless	4	4
8	grow	flow	2	2
9	rise	rice	1	1
10	legal	illegal	2	2
11	same	same	0	0

Таблица 4.2 — Результаты тестирования рекурсивного алгоритма Дамерау-Левенштейна

№	Строка 1	Строка 2	Расстояние Дамерау-Левенштейна	Ожидаемое расстояние Дамерау-Левенштейна
1	some	any	4	4
2		nothing	7	7
3			0	0
4	bashrc	bashcr	1	1
5	bus	BuS	2	2
6	electricity	city	7	7
7	powerful	powerless	4	4
8	grow	flow	2	2
9	rise	rice	1	1
10	legal	illegal	2	2
11	same	same	0	0

Таблица 4.3 — Результаты тестирования рекурсивного алгоритма Дамерау-Левенштейна

№	Строка 1	Строка 2	Расстояние Дамерау-Левенштейна	Ожидаемое расстояние Дамерау-Левенштейна
1	some	any	4	4
2		nothing	7	7
3			0	0
4	bashrc	bashcr	1	1
5	bus	BuS	2	2
6	electricity	city	7	7
7	powerful	powerless	4	4
8	grow	flow	2	2
9	rise	rice	1	1
10	legal	illegal	2	2
11	same	same	0	0

4.3 Эксперименты по замеру времени

Чтобы подтвердить вывод об оценке сложности алгоритмов поиска редакционного расстояния, проведём эксперименты по замеру времени и построим графики зависимости времени выполнения данных алгоритмов от длины обрабатываемых слов.

4.3.1 Эксперимент 1

На рисунке 4.5 приведён график сравнения алгоритма Вагнера-Фишера и матричного алгоритма Дамерау-Левенштейна. Для этого эксперимента было сгенерировано 100 пар полностью не совпадающих строк с диапазоном длин от 10 до 1000. Как видно, законы изменения времени выполнения этих алгоритмов практически одинаковы и отличаются лишь на некоторый постоянный коэффициент.

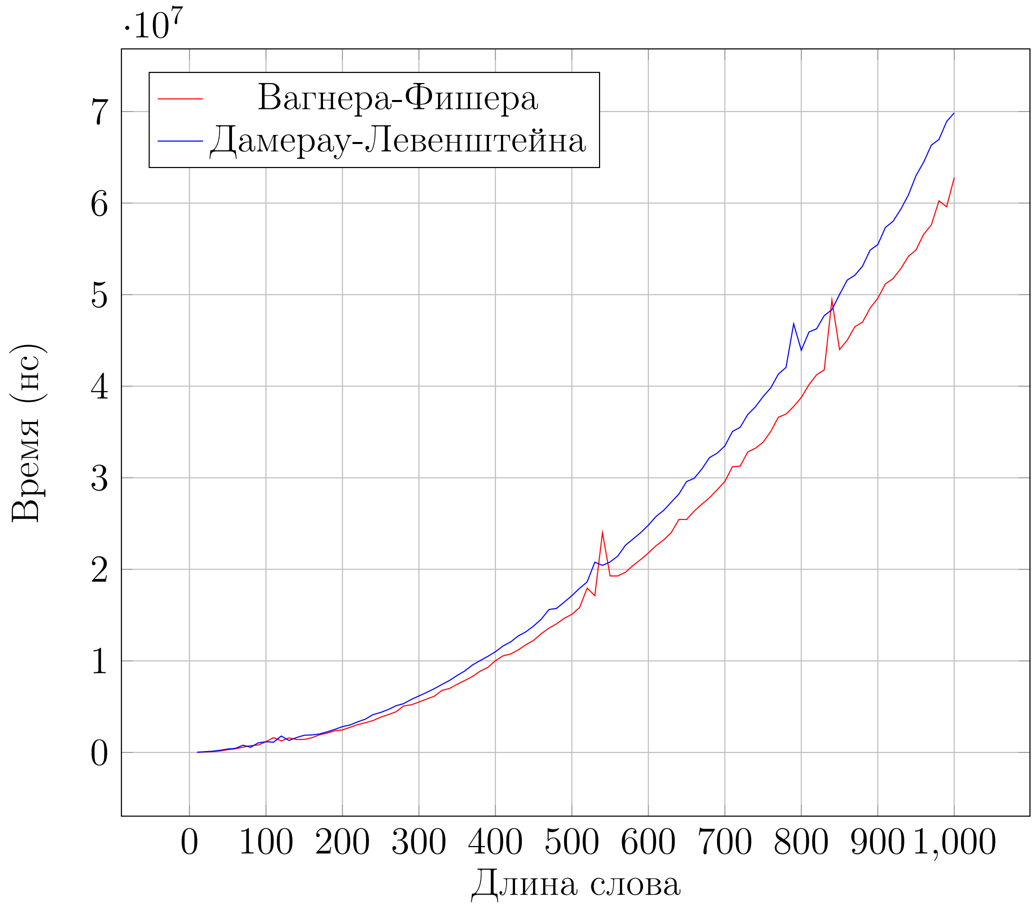

Рисунок 4.5 — График сравнения алгоритма Вагнера-Фишера и матричного алгоритма Дамерау-Левенштейна

4.3.2 Эксперимент 2

На рисунке 4.6 приведён график сравнения рекурсивного и матричного алгоритмов нахождения расстояния Дамерау-Левенштейна. Для этого эксперимента было сгенерировано 10 пар полностью различных слов с диапазоном длин от 1 до 10. Количество времени, необходимого для выполнения рекурсивного алгоритма, растёт экспоненциально, в то время как сложность матричного алгоритма имеет квадратичный рост, что наглядно изображено на рисунке 4.5.

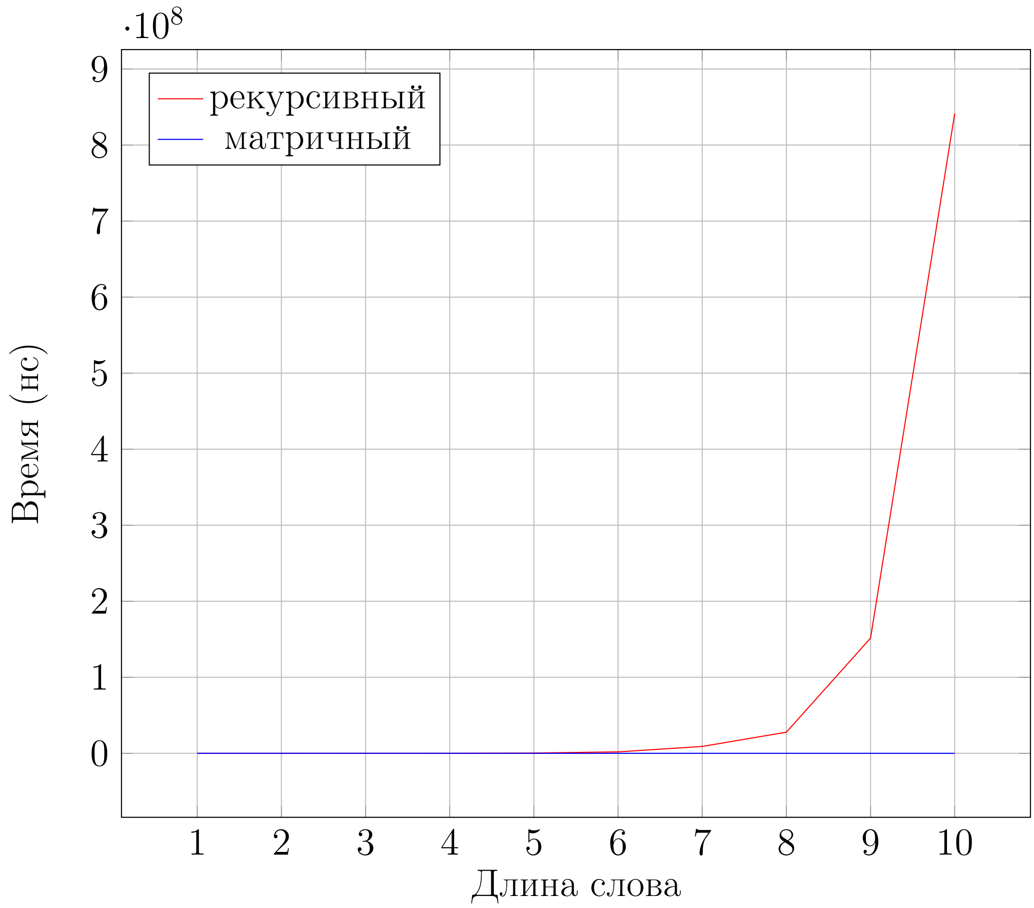

Рисунок 4.6 — График сравнения рекурсивного и матричного алгоритмов Дамерау-Левенштейна

4.4 Вывод

Как итог, была подтверждена корректная работоспособность реализованной программы нахождения расстояний Левенштейна и Дамерау-Левенштейна и доказаны тезисы, составленные в результате анализа этих алгоритмов.

ЗАКЛЮЧЕНИЕ

В ходе выполнения данной лабораторной работы мной был изучен метод динамического программирования на материале алгоритмов поиска редакционного расстояния. Кроме того, были изучены непосредственно алгоритмы поиска редакционного расстояния, проведён их анализ и сравнение, успешно реализована и протестирована программа, осуществляющая этот поиск, проведены эксперименты, в ходе которых были подтверждены полученные в ходе анализа тезисы.