

Дисциплина: АНАЛИЗ АЛГОРИТМОВ

ЛАБОРАТОРНАЯ РАБОТА №1

ТЕМА-расстояния Левенштейна и Дамерау-Левенштейна.

Студент группы ИУ7-55,
Аминов Тимур Саидович

Содержание

Введение	2
1 Аналитическая часть	3
1.1 Описание алгоритмов	3
1.2 Вывод	3
2 Конструкторская часть	4
2.1 Разработка алгоритмов	4
2.2 Вывод	7
3 Технологическая часть	8
3.1 Требования к программному обеспечению	8
3.2 Средства реализации	8
3.3 Листинг кода	9
3.4 Тестирование	10
3.5 Сравнительный анализ потребляемой памяти	10
3.6 Оценка потребляемой памяти на 4 и 1000 символов	12
3.7 Вывод	13
4 Экспериментальная часть	14
4.1 Постановка эксперимента	14
4.2 Результаты эксперимента	14
4.3 Вывод	15
Заключение	16
Список литературы	17

Введение

Расстояние Левенштейна - минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Расстояние Левенштейна используют:

- для исправления ошибок в слове;
- для сравнения текстовых файлов утилитой diff и ей подобными;
- для сравнения ДНК в биоинформатике.

Цель работы: изучение метода динамического программирования на материале алгоритмов Левенштейна и Дамерау-Левенштейна. Задачи работы:

1. Изучение алгоритмов Левенштейна и Дамерау-Левенштейна нахождения расстояния между строками;
2. Применение метода динамического программирования для матричной реализации указанных алгоритмов;
3. Получение практических навыков реализации указанных алгоритмов: двух алгоритмов в матричной версии и одного из алгоритмов в рекурсивной версии;
4. Сравнительный анализ линейной и рекурсивной реализаций выбранного алгоритма определения расстояния между строками по затрачиваемым ресурсам (времени и памяти);
5. 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;
6. Описание и обоснование полученных результатов в отчете о выполненной лабораторной работе, выполненного как расчётно-пояснительная записка к работе.

1 Аналитическая часть

1.1 Описание алгоритмов

Обозначим простейшие действия над двумя строками:

1. D - удалить букву в одной из строк;
2. I - вставить букву в одной из строк;
3. R - заменить букву в одной из строк;
4. M - совпадение двух букв.

Каждая операция имеет свою ‘цену’ - у совпадения она 0, а у трех остальных - по 1. Задача нахождения расстояния Левенштейна между двумя строками сводится к поиску набора операций, которые нужно совершить, чтобы трансформировать одну строку в другую, причем их суммарная цена должна быть минимальной.

Расстояние Дамерау-Левенштейна отличается от расстояния Левенштейна добавлением операции транспозиции - перестановки двух соседних символов (ее цена - 1).

Пусть S_1 и S_2 — две строки над некоторым алфавитом, тогда расстояние Левенштейна между ними можно подсчитать по следующей формуле (1):

$$D(i, j) = \begin{cases} 0, & i = 0, j = 0 \\ i, & j = 0, i > 0 \\ j, & i = 0, j > 0 \\ \min(D1, D2, D3) & \end{cases} \quad (1)$$

где:

$D1 = D(i, j-1) + 1$;

$D2 = D(i-1, j) + 1$;

$D3 = D(i-1, j-1) + (0 \text{ если } \text{match}(S_1[i], S_2[j]), 1 \text{ иначе})$;

$\text{match}(a, b)$ = истина при $a = b$, ложь иначе;

$\min(A1, A2, \dots, AN)$ - минимум среди чисел $A1, A2, \dots, AN$.

В свою очередь, расстояние Дамерау-Левенштейна можно посчитать по формуле (2):

$$D(i, j) = \begin{cases} 0, & i = 0, j = 0 \\ i, & j = 0, i > 0 \\ j, & i = 0, j > 0 \\ \min(D1, D2, D3, D4), & i > 1, j > 1, a_i = b_{j-1}, a_{i-1} = b_j \\ \min(D1, D2, D3) & \end{cases} \quad (2)$$

где:

$D4 = D(i-2, j-2) + 1$, если $i > 1, j > 1$ и $a_i = b_{j-1}, a_{i-1} = b_j$.

1.2 Вывод

В данном разделе были рассмотрены формулы для нахождения расстояния Левенштейна и Дамерау-Левенштейна, а также основные применения этих расстояний.

2 Конструкторская часть

2.1 Разработка алгоритмов

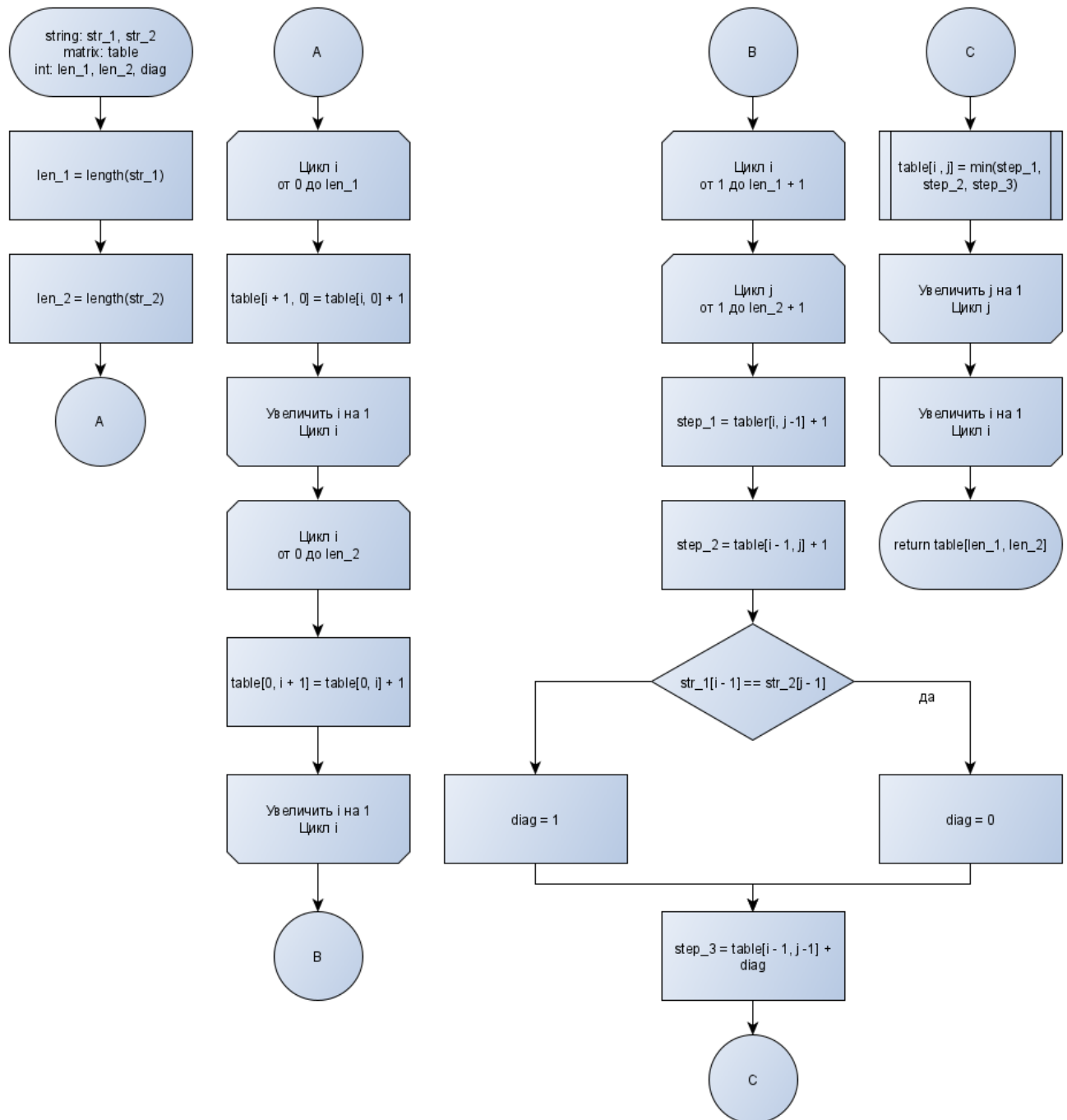


Рис. 1: Схема алгоритма нахождения расстояния Левенштейна(итеративного)

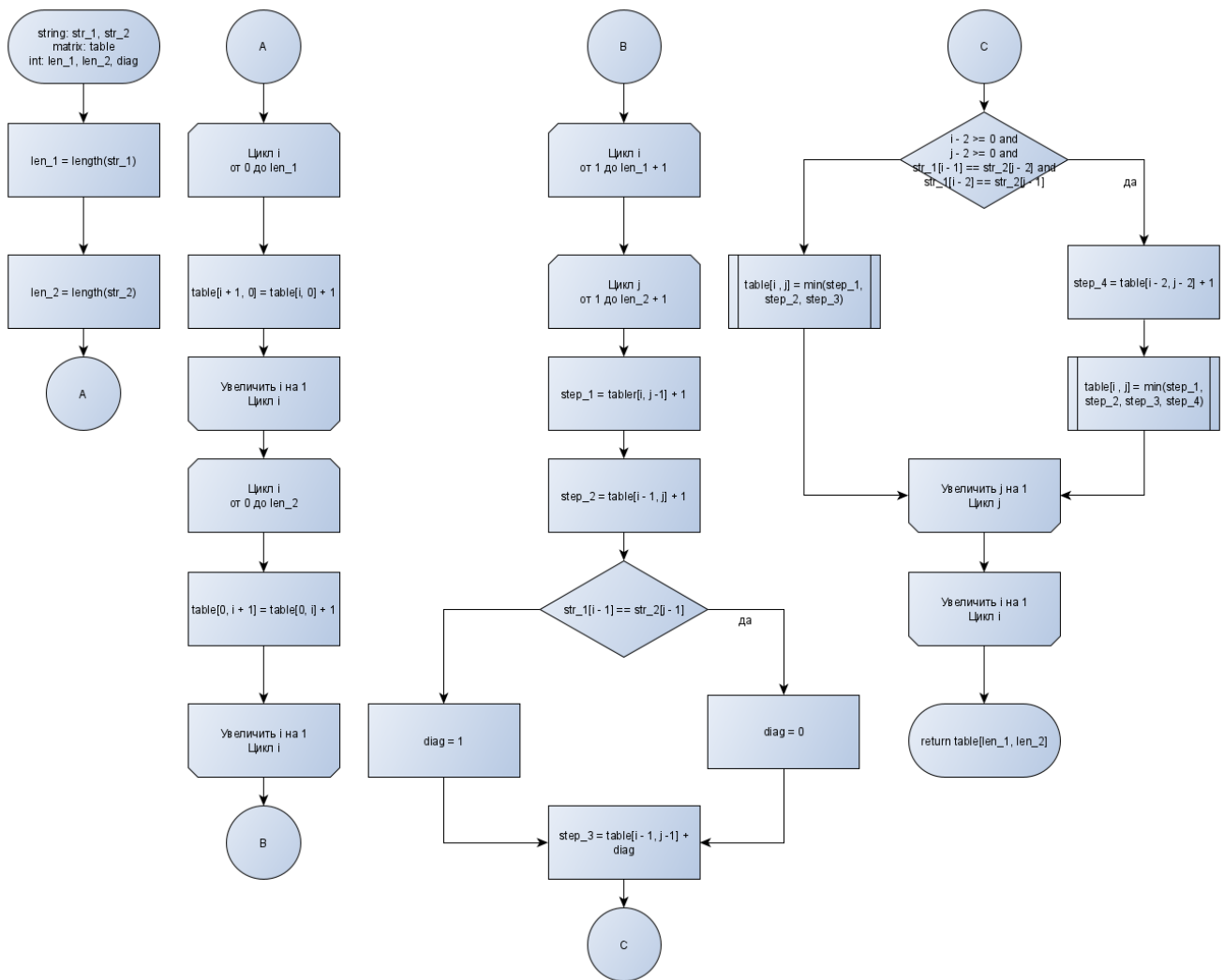


Рис. 2: Схема алгоритма нахождения расстояния Дameraу-Левенштейна(итеративного)

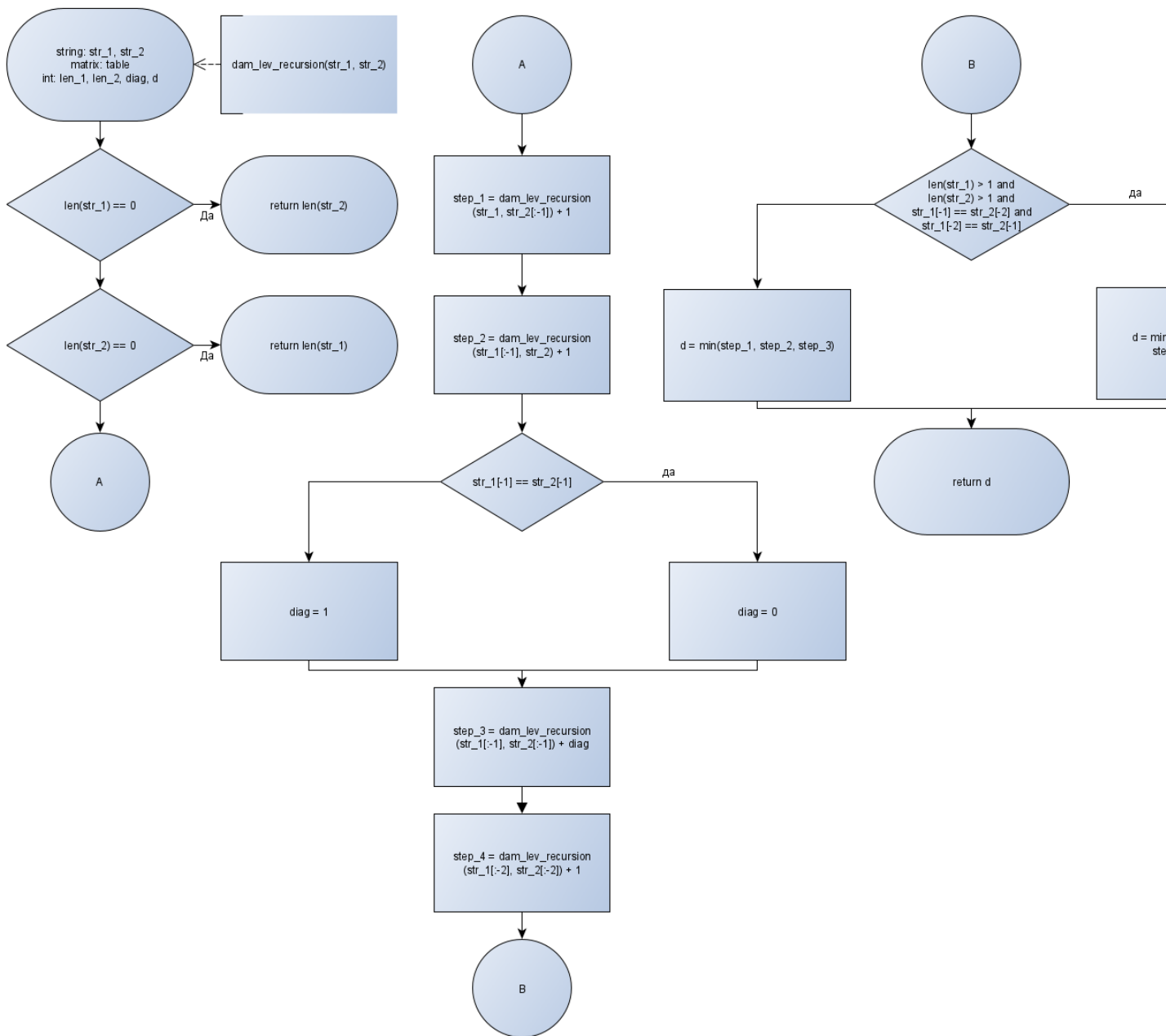


Рис. 3: Схема алгоритма нахождения расстояния Дameraу-Левенштейна(рекурсивного)

2.2 Вывод

В данном разделе были рассмотрены схемы алгоритмов нахождения расстояния Левенштейна, Дamerau-Левенштейна, а также рекурсивный алгоритм нахождения расстояния Дamerau-Левенштейна.

3 Технологическая часть

3.1 Требования к программному обеспечению

Входные данные: str1 - первое слово, str2 - второе слово.

Выходные данные: значение расстояния между двух слов.

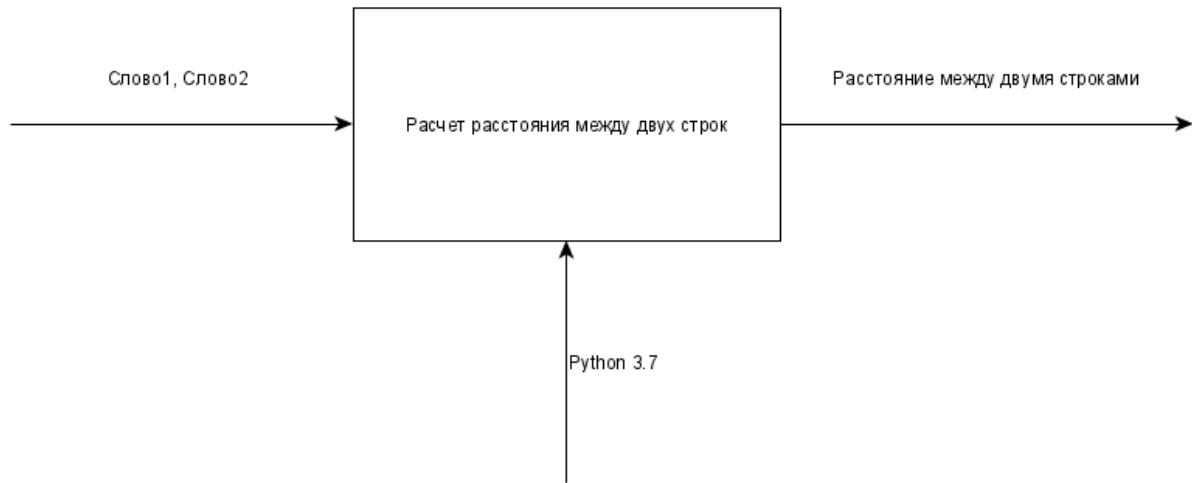


Рис. 4: IDEF0-диаграмма, описывающая алгоритм нахождения расстояния Левенштейна

3.2 Средства реализации

В качестве языка был выбран python, для вычисления памяти был использован метод `sys.getsizeof()`, возвращающий память, затрачиваемую на объект. Для замеров времени была использована функция `tick()`, приведенная ниже:

Листинг 1: Функция замера времени

```
1 unsigned long long tick(void)
2 {
3     unsigned long long d;
4     __asm__ __volatile__ ("rdtsc" : "=A" (d) );
5     return d;
6 }
```

Для подключения ее в модуль python была создана библиотека `libtick.so` и подключена с помощью модуля `ctypes`.

3.3 Листинг кода

Листинг 2: Функция нахождения расстояния Левенштейна итеративно

```
1 def lev_table(str_1, str_2):
2     len_1 = len(str_1)
3     len_2 = len(str_2)
4     table = [[0 for j in range(len_2 + 1)] for i in range(len_1 + 1)]
5
6     for i in range(len_1):
7         table[i + 1][0] = table[i][0] + 1
8
9     for i in range(len_2):
10        table[0][i + 1] = table[0][i] + 1
11
12    for i in range(1, len_1 + 1, 1):
13        for j in range(1, len_2 + 1, 1):
14            step_1 = table[i - 1][j] + 1
15            step_2 = table[i][j - 1] + 1
16
17            if str_1[i - 1] == str_2[j - 1]:
18                diag = 0
19            else:
20                diag = 1
21
22            step_3 = table[i - 1][j - 1] + diag
23
24            table[i][j] = min(step_1, step_2, step_3)
25
26    return table[len_1][len_2]
```

Листинг 3: Функция нахождения расстояния Дамерау-Левенштейна итеративно

```
1 def dam_lev_table(str_1, str_2):
2     len_1 = len(str_1)
3     len_2 = len(str_2)
4     table = [[0 for j in range(len_2 + 1)] for i in range(len_1 + 1)]
5
6     for i in range(len_1):
7         table[i + 1][0] = table[i][0] + 1
8
9     for i in range(len_2):
10        table[0][i + 1] = table[0][i] + 1
11
12    for i in range(1, len_1 + 1, 1):
13        for j in range(1, len_2 + 1, 1):
14            step_1 = table[i - 1][j] + 1
15            step_2 = table[i][j - 1] + 1
16
17            if str_1[i - 1] == str_2[j - 1]:
18                diag = 0
19            else:
20                diag = 1
21
22            step_3 = table[i - 1][j - 1] + diag
23
24            if i - 2 >= 0 and j - 2 >= 0 and str_1[i - 1] == str_2[j - 2] and str_1[i - 2]
                == str_2[j - 1]:
25                step_4 = table[i - 2][j - 2] + 1
26                table[i][j] = min(step_1, step_2, step_3, step_4)
27            else:
28                table[i][j] = min(step_1, step_2, step_3)
29
30    return table[len_1][len_2]
```

Листинг 4: Функция нахождения расстояния Дамерау-Левенштейна рекурсивно

```

1  def dam_lev_recursion(str_1, str_2):
2      if len(str_1) == 0:
3          return len(str_2)
4
5      elif len(str_2) == 0:
6          return len(str_1)
7
8      else:
9          if str_1[-1] == str_2[-1]:
10             diag = 0
11          else:
12             diag = 1
13
14             step_1 = dam_lev_recursion(str_1, str_2[:-1]) + 1
15             step_2 = dam_lev_recursion(str_1[:-1], str_2) + 1
16             step_3 = dam_lev_recursion(str_1[:-1], str_2[:-1]) + diag
17             step_4 = dam_lev_recursion(str_1[:-2], str_2[:-2]) + 1
18
19             if len(str_1) > 1 and len(str_2) > 1 and str_1[-1] == str_2[-2] and str_1[-2] ==
                str_2[-1]:
20                 d = min(step_1, step_2, step_3, step_4)
21             else:
22                 d = min(step_1, step_2, step_3)
23
24     return d

```

3.4 Тестирование

Было организовано тестирование с помощью заранее подготовленных данных в виде пары строк и расстояния по Левенштейну и Дамерау-Левенштейну. Они приведены в таблице 1:

Таблица 1. Тестовые данные для алгоритмов нахождения расстояния Левенштейна и Дамерау-Левенштейна.

s1	s2	р. Левенштейна	р. Дамерау-Левенштейна
'mod'	'rot'	2	2
'mod'	'mdo'	2	1
'mod'	'ram'	3	3
'mod'	'omr'	3	2
'modrotmod'	'modrot'	3	3
"	'modrot'	6	6
"	"	0	0
"	'a'	1	1
'a'	'b'	1	1
'modr'	'mord'	2	1

Все реальные результаты совпали с ожидаемыми из таблицы 1, что подтвердило правильность работы программы.

3.5 Сравнительный анализ потребляемой памяти

Найдем память, затрачиваемую на объекты одного вызова каждой из функций:

Таблица 2. Память, потребляемая структурами в алгоритме Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	$40 + 8 * [\text{len}(\text{str1}) + 1] + [\text{len}(\text{str1}) + 1] * [40 + 8 * (\text{len}(\text{str 2}) + 1)]$
2 строки	$2 * [49 + \text{len}(\text{str})]$
6 вспомогательных переменных(int)	168
2 счетчика (int)	56

Таблица 3. Память, потребляемая структурами в алгоритме Дамерау-Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	$40 + 8 * [\text{len}(\text{str1}) + 1] + [\text{len}(\text{str1}) + 1] * [40 + 8 * (\text{len}(\text{str2}) + 1)]$
2 строки	$2 * [49 + \text{len}(\text{str})]$
7 вспомогательных переменных (int)	196
2 счетчика (int)	56

Таблица 4. Память, потребляемая структурами в рекурсивном алгоритме нахождения расстояния Дамерау-Левенштейна за один вызов

Структура данных	Занимаемая память (байты)
4 вспомогательных переменных (int)	112
2 строки	$98 + \text{len}(\text{str1}) + \text{len}(\text{str2})$

Максимальная глубина рекурсивного вызова функции - сумма длин двух слов.

Суммируя память, затрачиваемую структурами данных в итеративном поиске расстояния Дамерау-Левенштейна по таблице 3 и принимая $\text{len}(s1) = \text{len}(s2) = l$, получим:

$$DLIMem = 8l^2 + 59l + 411, \quad (3)$$

где $DLIMem$ - суммарная память, затрачиваемая структурами данных в итеративном поиске расстояния Дамерау-Левенштейна.

Также найдем суммарную память, затрачиваемую рекурсивным поиском расстояния Дамерау-Левенштейна, принимая $\text{len}(\text{str1}) = \text{len}(\text{str2}) = l/2$ и максимальное количество одновременных вызовов - глубину вызовов (сумма длин двух строк).

$$DLRMem = 2l^2 + 420l, \quad (4)$$

где $DLRMem$ - суммарная память, затрачиваемая структурами данных в рекурсивном поиске расстояния Дамерау-Левенштейна.

Найдем значения l , при которых $DLRMem = DLIMem$:

$$8l^2 + 59l + 411 = 2l^2 + 420l$$

Вычтем из обеих частей правую, получим:

$$6l^2 - 361l + 411 = 0 \quad (5)$$

Найдем корни уравнения (5):

$$\begin{aligned} l &= \frac{361}{12} \pm \frac{\sqrt{120457}}{12} \\ l1 &= \frac{361}{12} + \frac{\sqrt{120457}}{12} \\ l2 &= \frac{361}{12} - \frac{\sqrt{120457}}{12} \end{aligned}$$

Исходя из того, что:

$$\begin{aligned} DLIMem > DLRMem &\Leftrightarrow 6l^2 - 361l + 411 > 0 \\ DLIMem < DLRMem &\Leftrightarrow 6l^2 - 361l + 411 < 0 \end{aligned}$$

Получаем:

$$\begin{aligned} DLIMem > DLRMem &\Leftrightarrow l \in (-\infty, l1) \cup (l2, +\infty) \\ DLIMem < DLRMem &\Leftrightarrow l \in (l1, l2) \end{aligned}$$

Из того, что $l1 \approx 1.09$, а $l2 \approx 59.08$ делаем вывод, что при длине строк в отрезке от 0 до 1 и в интервале от 59 до бесконечности итеративная реализация поиска расстояния Дамерау-Левенштейна потребляет больше памяти, чем рекурсивная, а в отрезке от 2 до 59 рекурсивная реализация более затратна по памяти. На рисунке 5 приведена зависимость потребляемой памяти в различных реализациях поиска расстояния Дамерау-Левенштейна.

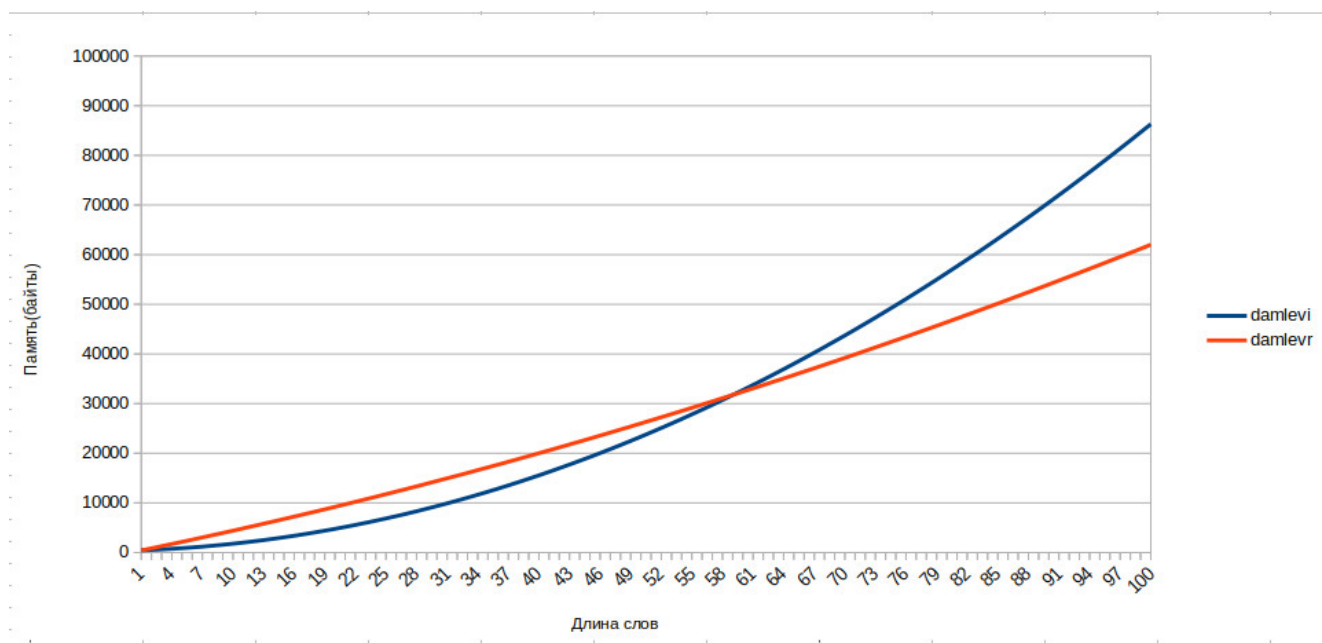


Рис. 5: Память, потребляемая реализациями алгоритма поиска расстояния Дамерау-Левенштейна

3.6 Оценка потребляемой памяти на 4 и 1000 символах

Оценим алгоритмы на словах длиной 4 символа:

Таблица 5. Память, потребляемая структурами в алгоритме нахождения расстояния Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	480
2 строки	106
5 вспомогательных переменных(int)	140
2 счетчика (int)	56
Всего	780

Таблица 6. Память, потребляемая структурами в алгоритме нахождения расстояния Дамерау-Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	480
2 строки	106
6 вспомогательных переменных(int)	168
2 счетчика (int)	56
Всего	808

Таблица 7. Память, потребляемая структурами в рекурсивном алгоритме нахождения расстояния Дамерау-Левенштейна

Структура данных	Занимаемая память (байты)
4 вспомогательных переменных (int)	$140 * 8(\text{максимальная глубина вызовов}) = 896$
2 строки	$106 * 4/2(\text{Усредненное значение}) = 212$
Всего	1108

Оценим алгоритмы на словах длиной 1000 символов:

Таблица 8. Память, потребляемая структурами в алгоритме нахождения расстояния Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	8 064 096
2 строки	106
5 вспомогательных переменных(int)	140
2 счетчика (int)	56
Всего	8064398

Таблица 9. Память, потребляемая структурами в алгоритме нахождения расстояния Дамерау-Левенштейна

Структура данных	Занимаемая память (байты)
Матрица	8 064 096
2 строки	106
6 вспомогательных переменных (int)	168
2 счетчика (int)	56
Всего	8064426

Таблица 10. Память, потребляемая структурами в рекурсивном алгоритме нахождения расстояния Дамерау-Левенштейна

Структура данных	Занимаемая память (байты)
4 вспомогательных переменных (int)	$140 * 2000$ (максимальная глубина вызовов) = 280000
2 строки	$2098 * 1000 / 2$ (Усредненное значение) = 1049000
Всего	1329000

По таблицам 5,6,8 и 9 мы видим, что и при длине слов 4, и при длине слов 1000 итеративные реализации алгоритмов нахождения расстояния Левенштейна и Дамерау-Левенштейна сравнимы по потребляемой памяти, причем разница между ними всегда 28 байт. В свою очередь, рекурсивная реализация Дамерау-Левенштейна при 4 потребляет больше памяти, а при 1000 - меньше, что видно в таблицах 6, 7, 9 и 10.

3.7 Вывод

В технологической части были предоставлены реализации алгоритмов нахождения расстояния Левенштейна в итеративной форме, а также нахождения расстояния Дамерау-Левенштейна в итеративной и рекурсивной формах. Помимо этого, были предоставлены результаты тестов на правильность данных реализаций, сравнительный анализ потребляемой памяти всех реализаций на слов длиной 4 и 1000. Было выявлено, что и при длине слов 4, и при длине слов 1000 итеративные реализации алгоритмов нахождения расстояния Левенштейна и Дамерау-Левенштейна сравнимы по потребляемой памяти, причем разница между ними всегда 28 байт (3 и 0.0001 процента при длине слова 4 и 1000 соответственно). Рекурсивная реализация же при длине слов 4 потребляет на 37 процентов больше памяти, в то время как при длине слов 1000 она потребляет меньше памяти на 84 процента. По зависимостям потребляемой памяти от длин строк (рис. 5) мы видим, что рекурсивная реализация более затратна на промежутке от 1 до 59, что вызвано тем, что в функцию всегда передается копия строки, из-за того, что в python нет указателей, чтобы передавать их в функции вместо копий, иначе бы рекурсивная реализация была всегда эффективнее по памяти.

4 Экспериментальная часть

4.1 Постановка эксперимента

Должны быть произведены замеры времени работы каждого из алгоритмов при длинах строк от 2 до 7. Каждый тест должен быть проведен 100 раз на 100 случайных строках. Таким образом тест для одной длины строки проводится 10000 раз. Как результат должно быть взято среднее значение для уменьшения роли случайных факторов в итоге.

4.2 Результаты эксперимента

Ниже приведены графики со сравнением требуемого процессорного времени на итеративные алгоритмы поиска расстояния Левенштейна и Дameraу-Левенштейна (рис 6) с длинами строк 2-7 и со сравнением времени, затрачиваемого на поиск расстояния Дameraу-Левенштейна в итеративном и рекурсивном вариантах. lev tab - поиск расстояния Левенштейна в итеративной форме, dam lev tab - поиск расстояния Дameraу-Левенштейна в итеративной форме, dam lev rec - поиск расстояния Дameraу-Левенштейна в рекурсивной форме, на оси абсцисс откладывается длина строк, на оси ординат - затрачиваемое время в секундах.

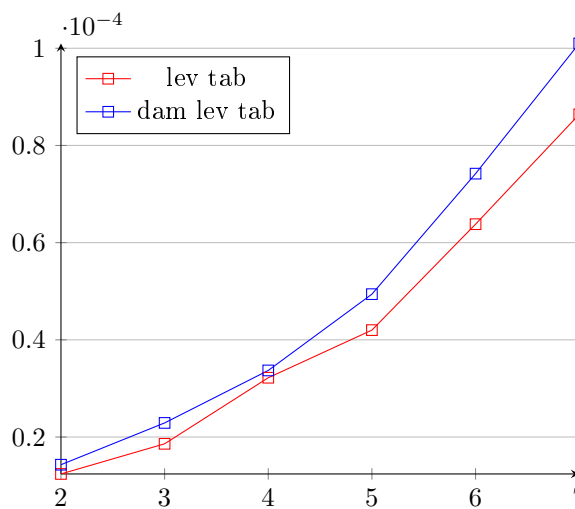


Рис. 6: сравнение времени на поиск Левенштейна и Дameraу-Левенштейна в итеративной форме

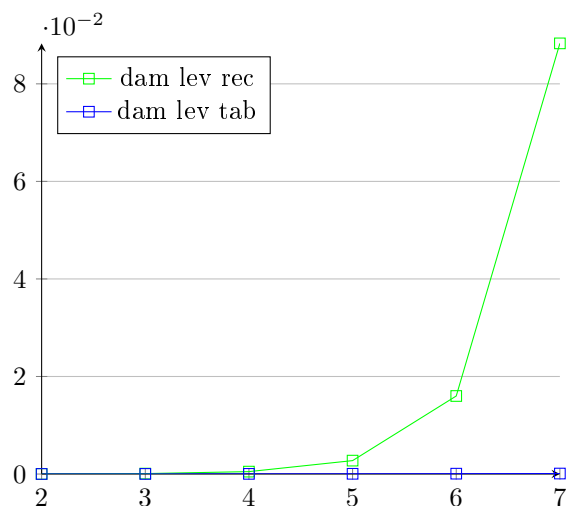


Рис. 7: сравнение времени на поиск Дameraу-Левенштейна в итеративной и рекурсивной формах

Таблица 11. Время, затрачиваемое различными алгоритмами на обработку строк длин от 2 до 9(в секундах).

Длина	Левенштейна	м. Дамерау-Левенштейна	р. Дамерау-Левенштейна
2	1.24e-05	1.43e-05	1.83e-05
3	1.86e-05	2.29e-05	9.10e-05
4	3.22e-05	3.37e-05	4.95e-04
5	4.20e-05	4.94e-05	2.75e-03
6	6.38e-05	7.42e-05	1.60e-02
7	8.64e-05	1.01e-04	8.83e-02

4.3 Вывод

По первому графику видно, что временные затраты на итеративный алгоритмы Левенштейна и Дамерау-Левенштейна сравнимы, но при этом алгоритм Дамерау-Левенштейна всегда медленнее. Из второго графика мы замечаем то, что рекурсивный алгоритм Дамерау-Левенштейна на порядки более затратный по времени, чем итеративный, начиная с длины строки в 5.

Заключение

В ходе данной лабораторной работы мною были реализованы алгоритмы Левенштейна в матричной форме и Дамерау-Левенштейна в матричной и рекурсивной форме. В ходе проверки на временные затраты было выявлено, что матричные реализации алгоритма Левенштейна и Дамерау-Левенштейна сравнимы по затрачиваемым процессорным ресурсам при длинах слов от 2 до 9 (разница между ними растет от 11 до 25 процентов в пользу Левенштейна). По используемой памяти разница меньше: от 0 до 3 процентов. Также было выявлено, что начиная со строк длиной в 3 символа рекурсивный вариант Дамерау-Левенштейна на порядки более затратный по процессорному времени, чем матричный (от 12.6 до 1.3×10^4 раз). Это вызвано тем, что в рекурсивном виде алгоритм одни и те же расчеты производит по несколько раз, так как расстояние между ними и теми же промежуточными словами может быть запрошено в нескольких независимо вызванных функциях, в то время как в матричном варианте все промежуточные расчеты записываются в матрицу и не пересчитываются. Говоря о памяти: рекурсивная форма нахождения расстояния Дамерау-Левенштейна при длине строки в 1000 на 84 процента меньше памяти, хотя при длине строки 4 потребляет на 34 процента больше памяти. Рекурсивная форма потребляет больше памяти на интервале от 1 до 59, при всех остальных значениях длин строк итеративная реализация более затратна по памяти.

Список литературы

- [1] Дж. Макконнелл. Анализ алгоритмов. Активный обучающий подход.-М.:Техносфера, 2009.
- [2] Нечёткий поиск в тексте и словаре // [Электронный ресурс]. Режим доступа: <https://habr.com/ru/post/114997/> (дата обращения: 2.10.19).
- [3] Нечеткий поиск, расстояние левенштейна алгоритм // [Электронный ресурс]. Режим доступа: <https://steposleep.ru/antanarivo-106/> (дата обращения: 2.10.19).