



Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Московский государственный технический университет имени
Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ «Информатика и системы управления»

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»

Отчет по лабораторной работе №7 по курсу "Анализ алгоритмов"

Тема Алгоритмы поиска в словаре

Студент Зайцева А.А.

Группа ИУ7-52Б

Преподаватели Волкова Л.Л.

Москва — 2021 г.

Оглавление

Введение	3
1 Аналитическая часть	4
1.1 Алгоритм полного перебора	4
1.2 Алгоритм бинарного поиска	4
1.3 Алгоритм поиска в сегментированном словаре	5
1.4 Вывод из аналитической части	6
2 Конструкторская часть	7
2.1 Структура используемого словаря	7
2.2 Схема алгоритма полного перебора	7
2.3 Схема алгоритма бинарного поиска	9
2.4 Схема алгоритма поиска в сегментированном словаре	10
2.5 Вывод из конструкторской части	13
3 Технологическая часть	14
3.1 Требования к ПО	14
3.2 Выбор средств реализации	14
3.3 Листинги кода	14
3.4 Тестирование	17
3.5 Вывод из технологического раздела	17
4 Исследовательская часть	18
4.1 Технические характеристики	18
4.2 Пример работы программы	18
4.3 Количество сравнений	19
4.4 Сравнение трудоемкостей	35
4.5 Сравнение времени выполнения реализаций алгоритмов	38
4.6 Вывод из исследовательской части	41
Заключение	42
Список использованной литературы	43

Введение

Словарь (англ. dictionary, map) — абстрактный тип данных, позволяющий хранить набор значений, обращение к которым происходит по ключам. Ключи должны допускать сравнение друг с другом. Примеры словарей достаточно разнообразны. Например, обычный толковый словарь хранит определения слов (являющиеся значениями), сопоставленные с самими словами (являющимися ключами), а банковская база данных может хранить данные клиентов, сопоставленные с номерами счетов.

Одной из основных операций в словаре является поиск значения по ключу. Словари могут содержать большое количество элементов, поэтому вопрос скорости поиска в них является важным [1].

Целью данной работы является разработка эффективного алгоритма поиска по словаре.

В рамках выполнения работы необходимо решить следующие задачи:

- 1) изучить три алгоритма поиска в словаре: алгоритм полного перебора, алгоритм бинарного поиска и алгоритм поиска в сегментированном словаре;
- 2) разработать и реализовать изученные алгоритмы;
- 3) провести сравнительный анализ трудоёмкости реализаций алгоритмов на основе теоретических расчетов (в среднем; в лучшем и худшем случаях);
- 4) провести сравнительный анализ процессорного времени выполнения реализаций алгоритмов на основе экспериментальных данных;
- 5) провести сравнительный анализ количества сравнений с заданным ключом, необходимых для поиска каждым алгоритмом значения по ключу и для определения отсутствия заданного ключа в словаре.

1 Аналитическая часть

В данном разделе будет приведена теория, необходимая для разработки и реализации трех алгоритмов поиска в словаре: алгоритма полного перебора, алгоритма бинарного поиска и алгоритма поиска в сегментированном словаре.

1.1 Алгоритм полного перебора

Алгоритм полного перебора сводится к последовательному прохождению по всем ключам словаря и их сравнению с заданным ключом.

Этот алгоритм считается методом «грубой силы» [2]. Зато он может производиться в неотсортированном словаре, а добавление новых элементов в такой словарь не вызывает затруднений – их можно добавлять на любую позицию. Дополнительных затрат по памяти также не требуется.

1.2 Алгоритм бинарного поиска

Двоичный поиск (бинарный поиск) [3] — алгоритм поиска объекта (в нашем случае - значения) по заданному признаку (ключу) в множестве объектов, упорядоченных по тому же самому признаку (по ключу), принцип работы которого заключается в том, что на каждом шаге множество объектов делится на две части и в работе остаётся та часть множества, где находится искомый объект.

Данный алгоритм работает за логарифмическое время, дополнительных затрат по памяти не требует. Однако алгоритм бинарного поиска подразумевает, что массив ключей отсортирован. Это вызывает трудности при добавлении новых элементов в словарь, особенно если словарь хранится в постоянной памяти (например, на диске).

Алгоритм вставки элемента в упорядоченный массив заключается в последовательном анализе элементов массива, начиная с последнего и, при необходимости, сдвиге анализируемого элемента вправо на одну позицию, для того, чтобы освободить место для вставляемого элемента. Сдвиги про-

водятся до тех пор, пока не будет найдено место для вставляемого элемента, соответствующее его значению [4].

1.3 Алгоритм поиска в сегментированном словаре

Словарь можно разбить на сегменты так, чтобы все элементы с некоторым общим признаком попадали в один сегмент. Признаком для строк может быть первая буква, для чисел – остаток от деления. Элементы внутри одного сегмента можно также разбивать на подсегменты и так далее.

Затем сегменты упорядочиваются по значению частотной характеристики так, чтобы к элементам с наибольшей частотной характеристикой был самый быстрый доступ (в случае поиска полным перебором они должны быть ближе к началу). Такой характеристикой может послужить, например, размер сегмента.

Вероятность обращения к определенному сегменту равна сумме вероятностей обращений к его ключам, то есть $P_i = \sum_j p_j$, где P_i - вероятность обращения к i -ому сегменту, p_j - вероятность обращения к j -ому элементу, который принадлежит i -ому сегменту. Если обращения ко всем ключам равновероятны, то можно заменить сумму на произведение: $P_i = N \cdot p$, где N - количество элементов в i -ом сегменте, а p - вероятность обращения к произвольному ключу.

Таким образом, для поиска ключа в словаре необходимо найти сегмент, которому принадлежит этот ключ (например, полным перебором), а затем осуществить поиск ключа в пределах найденного сегмента. Чтобы поиск в пределах сегмента осуществлялся эффективнее, можно упорядочить ключи в каждом сегменте и реализовать бинарный поиск.

Сегментированный словарь позволяет уменьшить количество сравнений, необходимых для поиска в словаре. Однако он требует предобработки произвольного словаря для превращения его в сегментированный, а также дополнительной памяти под такую организацию. Объем дополнительной памяти определяется количеством выделяемых сегментов и подсегментов.

Также из-за необходимости поддержания структуры словаря усложня-

ется операция добавления в словарь новых элементов. Необходимо определить, к какому сегменту относится новый ключ, если сегмент разбит на подсегменты – соответствующий подсегмент, а затем найти позицию внутри сегмента (подсегмента), в которую этот ключ необходимо вставить для сохранения упорядоченности. Также после добавления нового ключа сегмент (подсегмент) становится больше, поэтому необходимо проверить упорядоченность сегментов (подсегментов) по частотной характеристике и пересортировать их в случае необходимости.

1.4 Вывод из аналитической части

Были рассмотрены идеи и материалы, необходимые для разработки и реализации трех алгоритмов поиска в словаре: алгоритма полного перебора, алгоритма бинарного поиска и алгоритма поиска в сегментированном словаре.

2 Конструкторская часть

В данном разделе будут приведены структура используемого словаря и схемы рассмотренных в предыдущем разделе алгоритмов поиска в словаре. Во всех алгоритмах возвращаемое значение -1 означает, что заданного ключа нет в словаре.

2.1 Структура используемого словаря

Словарь состоит из пар вида $\langle \text{en} - \text{rus} \rangle$, где en - слово на английском языке, rus - его перевод на русский.

2.2 Схема алгоритма полного перебора

На рисунке 2.1 приведена схема алгоритма полного перебора поиска в словаре.



Рисунок 2.1 – Схема алгоритма полного перебора

2.3 Схема алгоритма бинарного поиска

На рисунке 2.2 приведена схема алгоритма бинарного поиска в словаре. Предполагается, что передаваемый словарь `my_dict` лексикографически упорядочен по возрастанию ключей.

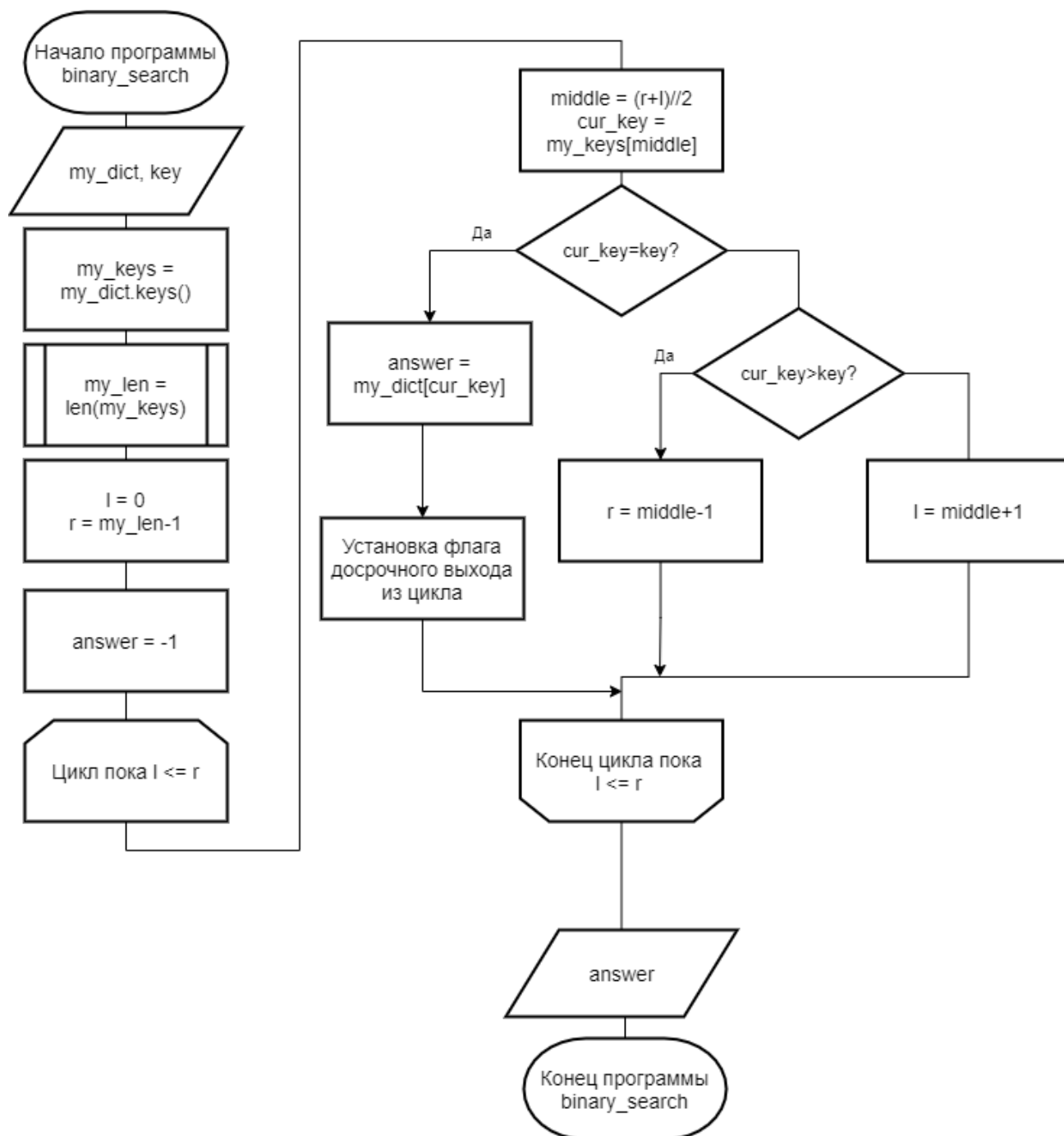


Рисунок 2.2 – Схема алгоритма бинарного поиска

2.4 Схема алгоритма поиска в сегментированном словаре

На рисунке 2.3 приведена схема создания словаря, сегментированного по первым буквам ключей (все буквы английского алфавита). Такой подход к сегментации выбран, так как он является наиболее нативным для используемого словаря. Предполагается, что обращения ко всем ключам равновероятны, поэтому для ускорения поиска нужного сегмента полным перебором сегментированный словарь упорядочен по убыванию количества пар (ключ: значение) в нем. Пары внутри одного сегмента упорядочены лексикографически по возрастанию ключей для возможности реализации бинарного поиска в пределах сегмента.

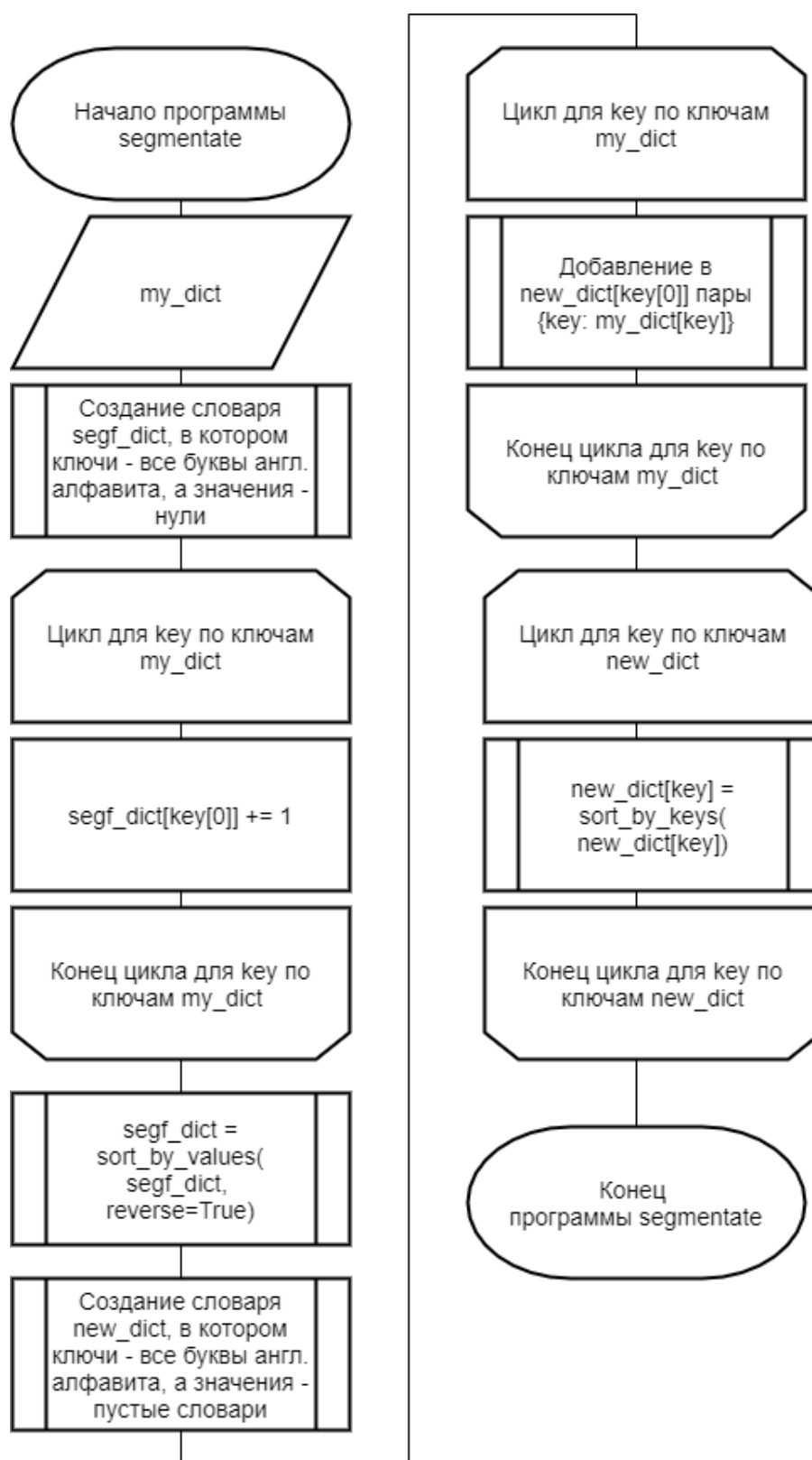


Рисунок 2.3 – Схема создания сегментированного словаря

На рисунке 2.4 приведена схема поиска в созданном сегментированном словаре. Сначала полным перебором осуществляется поиск нужного сегмента, а затем – бинарный поиск ключа в выбранном сегменте.

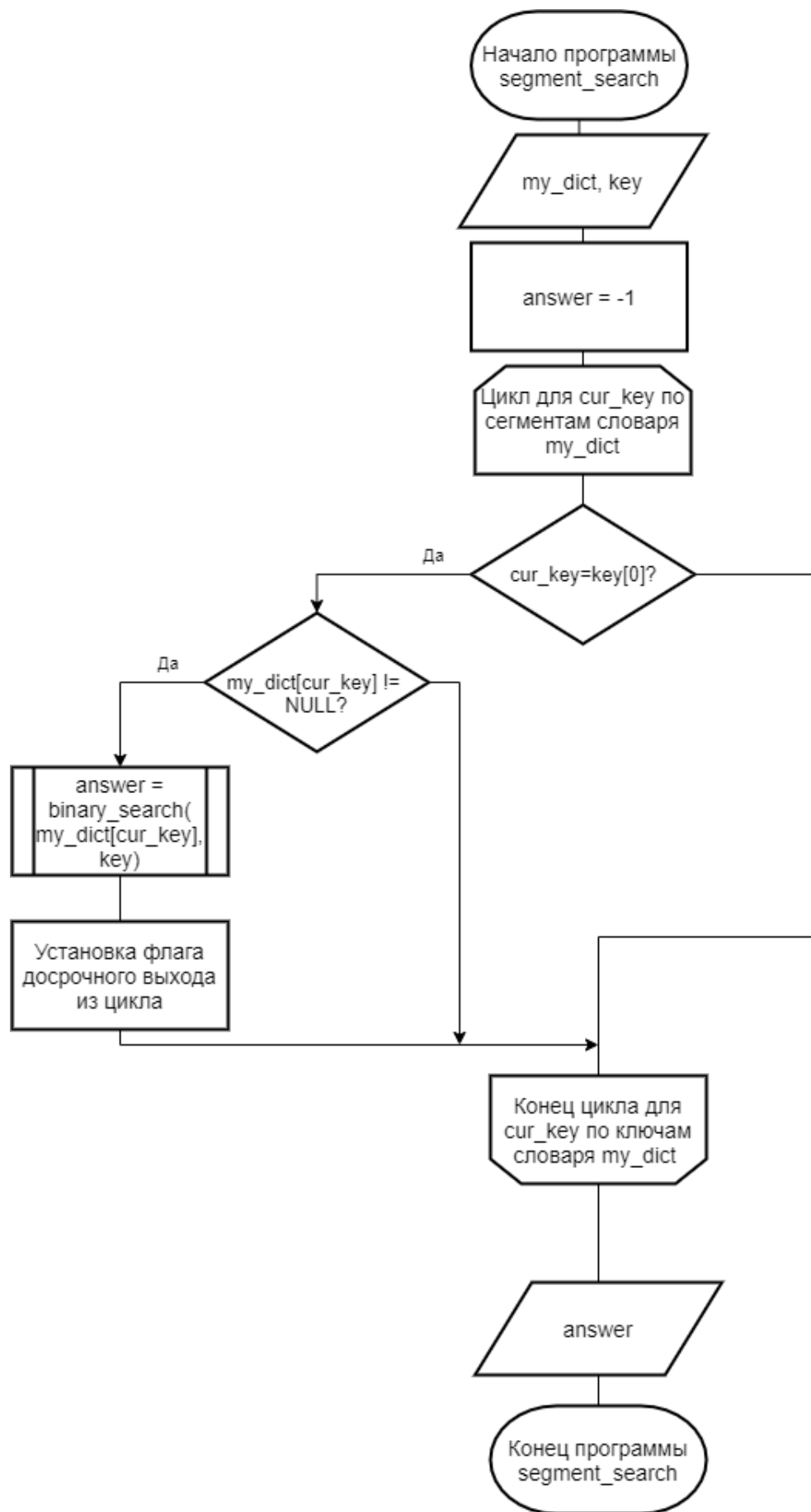


Рисунок 2.4 – Схема поиска в сегментированном словаре

2.5 Вывод из конструкторской части

Были приведены схемы разрабатываемых алгоритмов поиска в словаре: алгоритма полного перебора, алгоритма бинарного поиска и алгоритма поиска в сегментированном слолваре.

3 Технологическая часть

В данном разделе производится выбор средств реализации, приводятся требования к ПО, листинги реализованных алгоритмов поиска в словаре, а также результаты их тестирования.

3.1 Требования к ПО

На вход программе подается ключ, на выходе ожидается значение, найденное в словаре по заданному ключу каждым реализованным алгоритмом, или значение -1, если такого ключа в словаре нет.

3.2 Выбор средств реализации

В качестве языка программирования для реализации данной лабораторной работы был выбран язык Python [5]. Он позволяет быстро реализовывать различные алгоритмы без выделения большого времени на проектирование структуры программы и выбор типов данных.

Кроме того, в Python есть библиотека `time`, которая предоставляет функцию `process_time` для замера процессорного времени [6].

В качестве среды разработки выбран PyCharm. Он является кросс-платформенным, а также предоставляет удобный и функциональный отладчик и средства для рефакторинга кода, что позволяет быстро находить и исправлять ошибки [7].

3.3 Листинги кода

В листинге 3.1 представлена реализация алгоритма полного перебора поиска в словаре.

Листинг 3.1 – Алгоритм полного перебора

```
1 def full_search(my_dict, key):
2     for cur_key in my_dict.keys():
3         if cur_key == key:
4             return my_dict[cur_key]
5     return -1
```

В листинге 3.2 представлена реализация алгоритма бинарного поиска в словаре.

Листинг 3.2 – Алгоритм бинарного поиска в словаре

```
1 def binary_search(my_dict, key):
2     my_keys = list(my_dict.keys())
3     my_len = len(my_keys)
4     l = 0
5     r = my_len - 1
6
7     while l <= r:
8         middle = (r + l) // 2
9         cur_key = my_keys[middle]
10
11        if cur_key == key:
12            return my_dict[cur_key]
13        elif cur_key > key:
14            r = middle - 1
15        else:
16            l = middle + 1
17
18    return -1
```

В листинге 3.3 представлена реализация функции для создания сегментированного словаря.

Листинг 3.3 – Функция для создания сегментированного словаря

```
1
2 def segmentate(my_dict):
3     seg_frequency_dict = {letter: 0 for letter in "
4         abcdefghijklmnopqrstuvwxyz"}
5     for key in my_dict.keys():
6         seg_frequency_dict[key[0]] += 1
7
8     seg_frequency_dict = sort_by_values(seg_frequency_dict, reverse=
9         True)
10
11     new_dict = {letter: dict() for letter in seg_frequency_dict.keys
12         ()}
13     for key in my_dict.keys():
14         new_dict[key[0]].update({key: my_dict[key]})
15     for key in new_dict:
16         new_dict[key] = sort_by_keys(new_dict[key])
17
18     return new_dict
```

В листинге 3.4 представлена реализация алгоритма поиска в сегментированном словаре.

Листинг 3.4 – Алгоритм поиска в сегментированном словаре

```
1 def segment_search(my_dict, key):
2     for k in my_dict:
3         if key[0] == k:
4             if my_dict[k]:
5                 return binary_search(my_dict[k], key)
6             else:
7                 return -1
8     return -1
```


3.4 Тестирование

В таблице 3.1 приведены функциональные тесты для алгоритмов поиска в словаре. Все тесты пройдены успешно этим каждым алгоритмом.

Таблица 3.1 – Функциональные тесты

Ключ	Словарь	Ожидание	Результат
"few"	{}	-1	-1
"few"	{"few": "неск.", "little": "немн."}	"неск."	"неск."
"much"	{"few": "неск.", "little": "немн."}	-1	-1

3.5 Вывод из технологического раздела

Был произведен выбор средств реализации, приведены требования к ПО, реализованы и протестированы алгоритмы поиска в словаре.

4 Исследовательская часть

4.1 Технические характеристики

Технические характеристики устройства, на котором выполнялось тестирование:

- операционная система: Windows 10;
- оперативная память: 16 Гб;
- процессор: Intel® Core™ i5-8259U;
- количество ядер: 4;
- количество логических процессоров: 8.

Во время тестирования ноутбук был включен в сеть питания и нагружен только встроенными приложениями окружения и системой тестирования.

4.2 Пример работы программы

На рисунке 4.1 приведен пример работы программы.

```
Loaded 2 items
Словарь:
{'few': 'несколько', 'little': 'немного'}
Input key: few
Python search: несколько
Full search: несколько
Binary search: несколько
Segment search: несколько

Process finished with exit code 0
```

Рисунок 4.1 – Пример работы программы

4.3 Количество сравнений

На рисунках 4.2-4.7 приведены гистограммы с количеством сравнений (ось y), которые потребовались рассматриваемым реализациям алгоритмов поиска по ключам в словаре для того, чтобы найти нужный ключ (ось x). Все замеры производились в случае словаря, состоящего из 31 элемента. Для каждой реализации приведены две гистограммы, построенные по одним и тем же данным, но в первой ключи на оси x расположены в том порядке, в котором они хранились в словаре, а во второй – упорядочены по убыванию потребовавшегося количества сравнений.

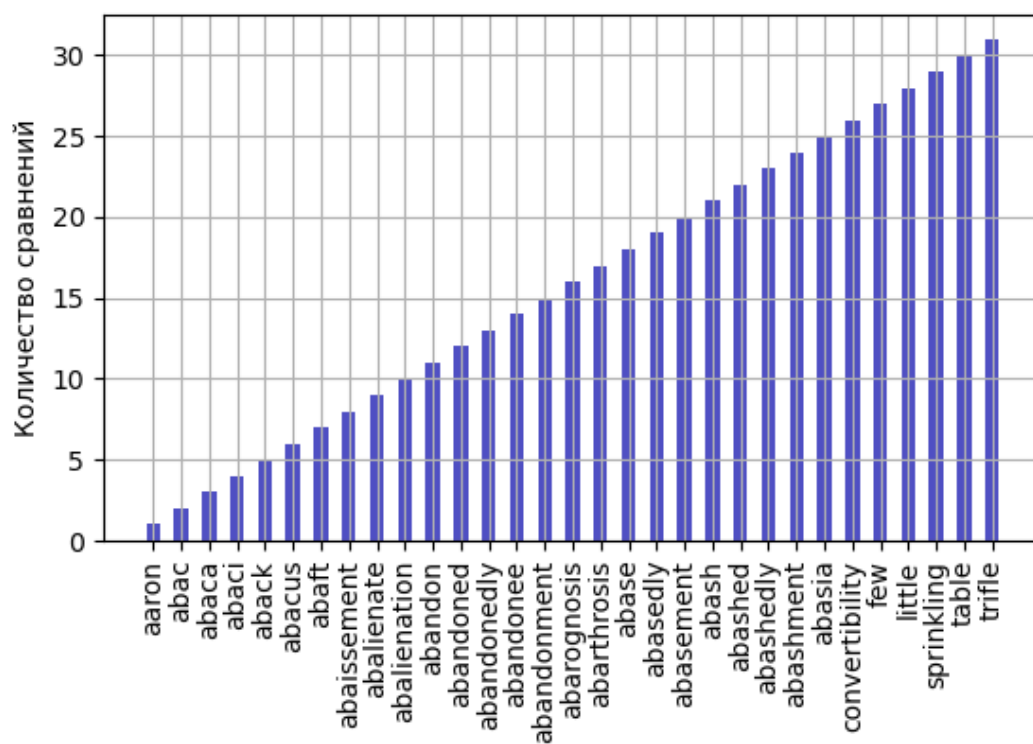


Рисунок 4.2 – Количество сравнений: полный перебор, сортировка по расположению

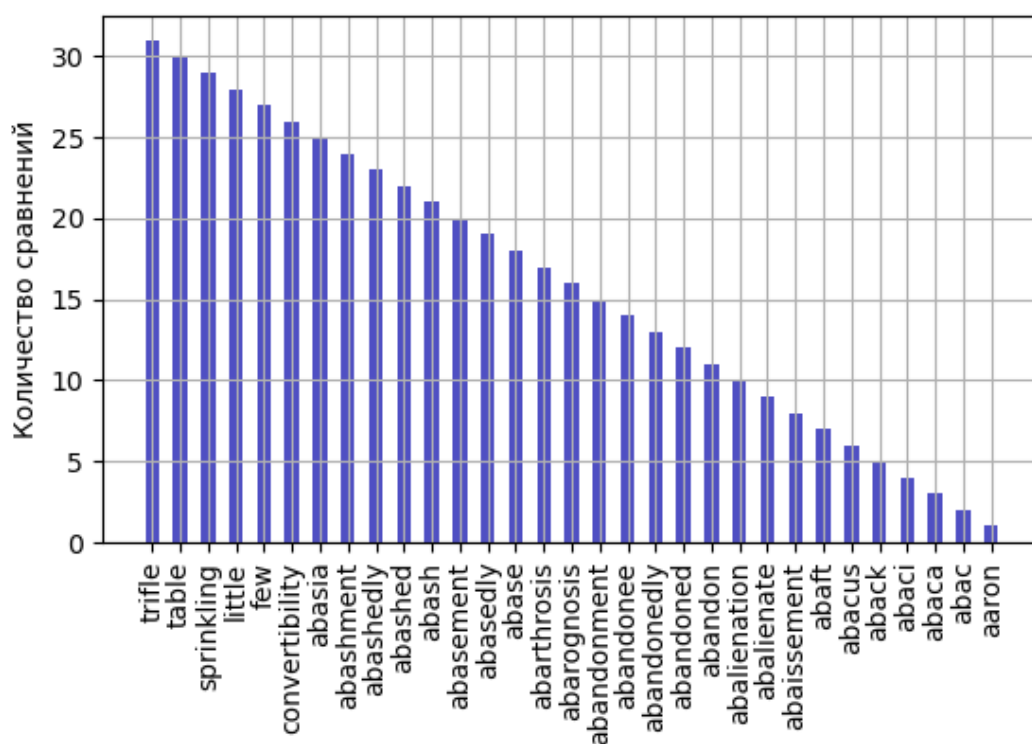


Рисунок 4.3 – Количество сравнений: полный перебор, сортировка по сравнениям

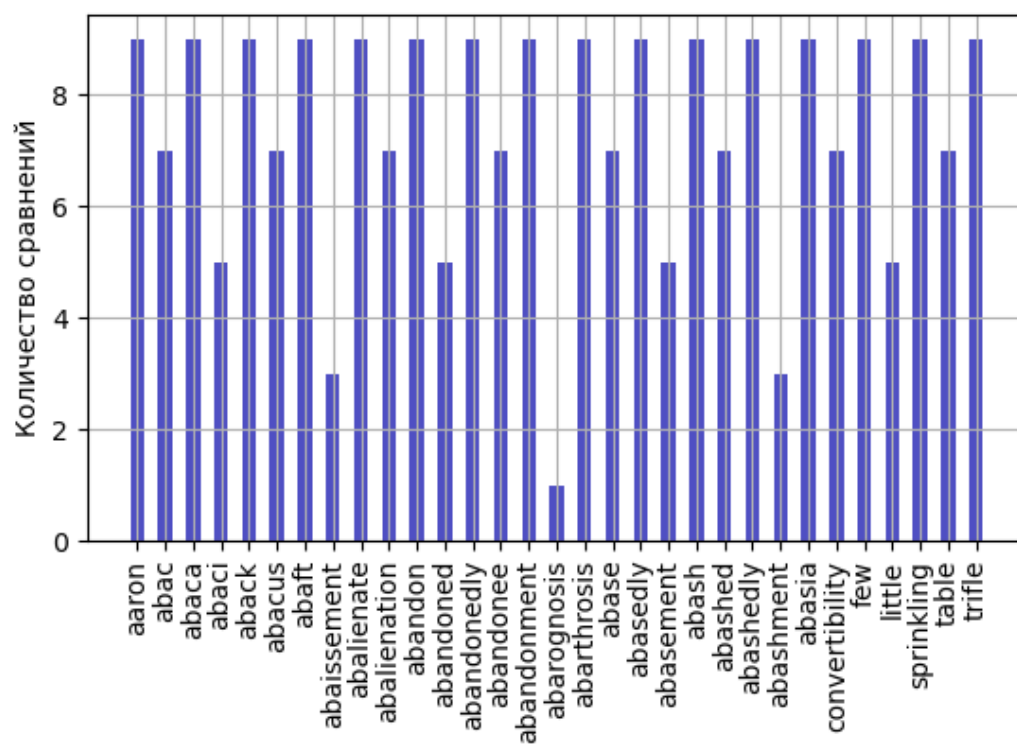


Рисунок 4.4 – Количество сравнений: бинарный поиск, сортировка по расположению

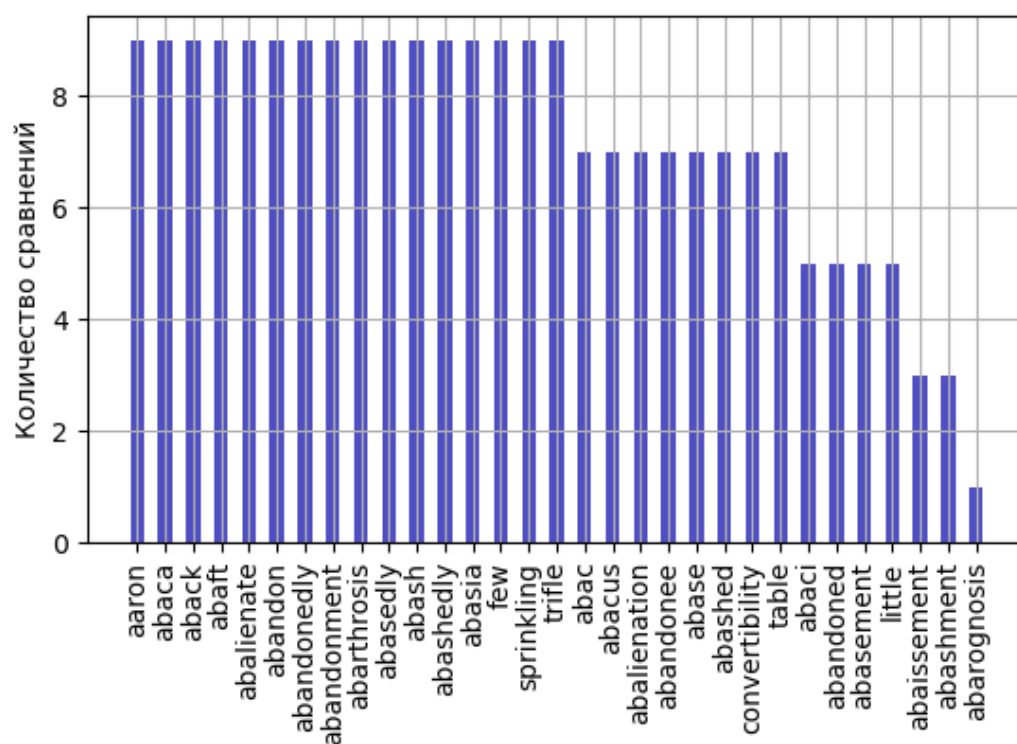


Рисунок 4.5 – Количество сравнений: бинарный поиск, сортировка по сравнениям

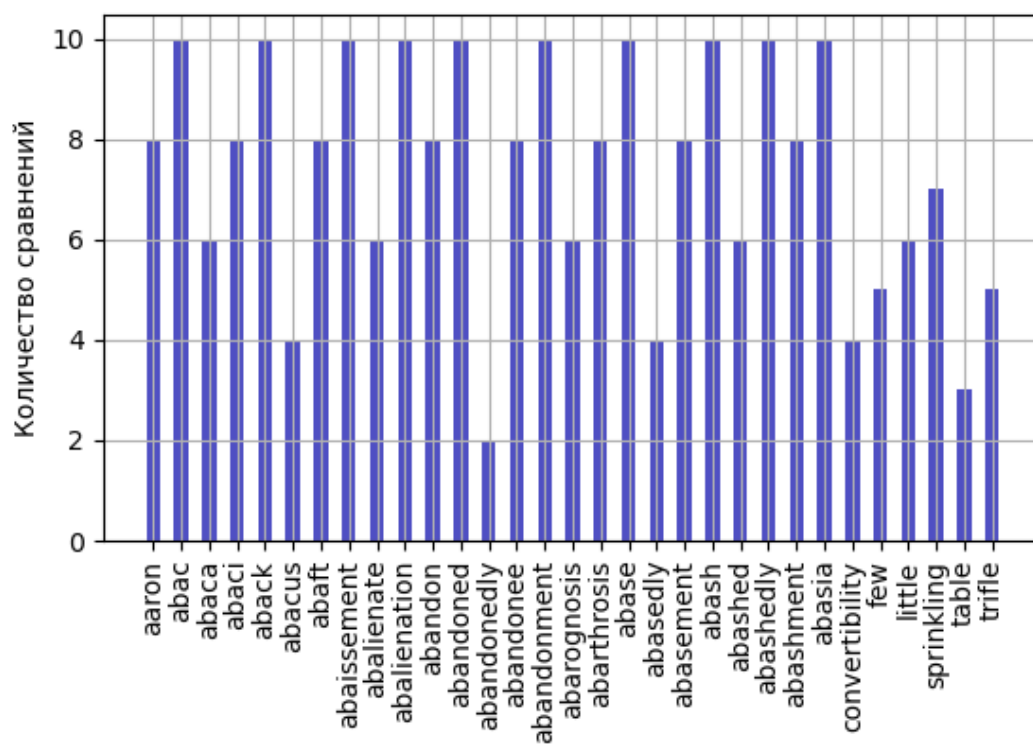


Рисунок 4.6 – Количество сравнений: поиск в сегментированном словаре, сортировка по расположению

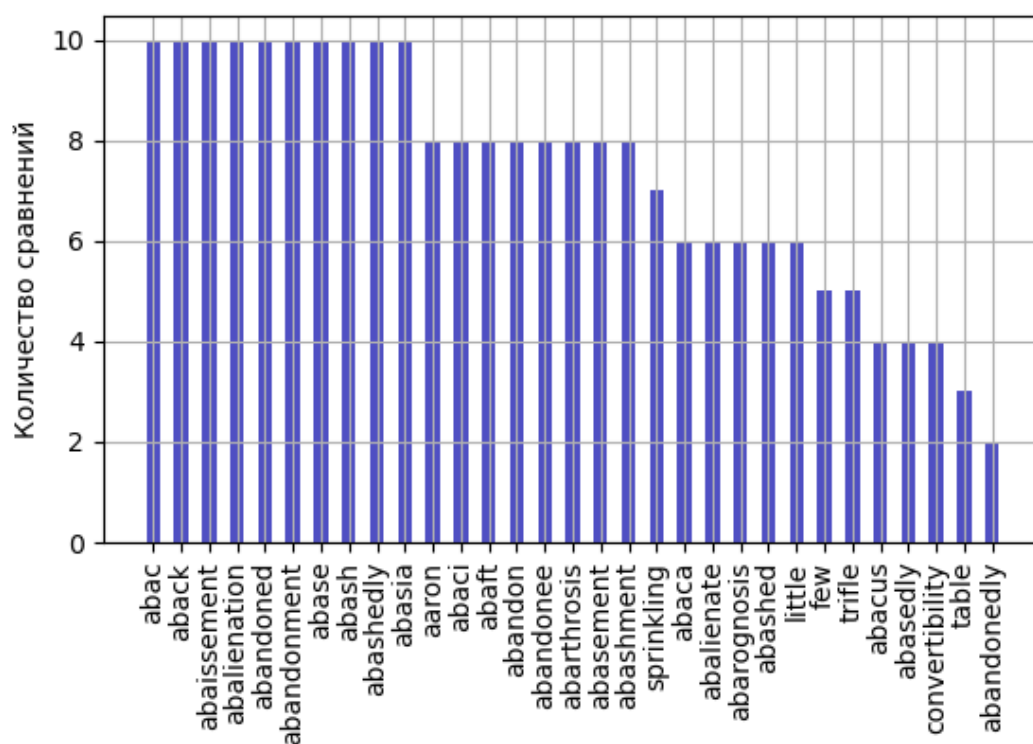


Рисунок 4.7 – Количество сравнений: поиск в сегментированном словаре, сортировка по сравнениям

Из рисунков 4.2 и 4.3 видно, что количество сравнений, необходимых для поиска данного ключа полным перебором определяется только позицией ключа в словаре и равно этой позиции. Наименьшее количество сравнений (1) потребуется для самого первого ключа, наибольшее (N -количество элементов в словаре) - для последнего. Для определения того, что ключа нет в словаре, потребуется столько же сравнений, сколько и для худшего случая.

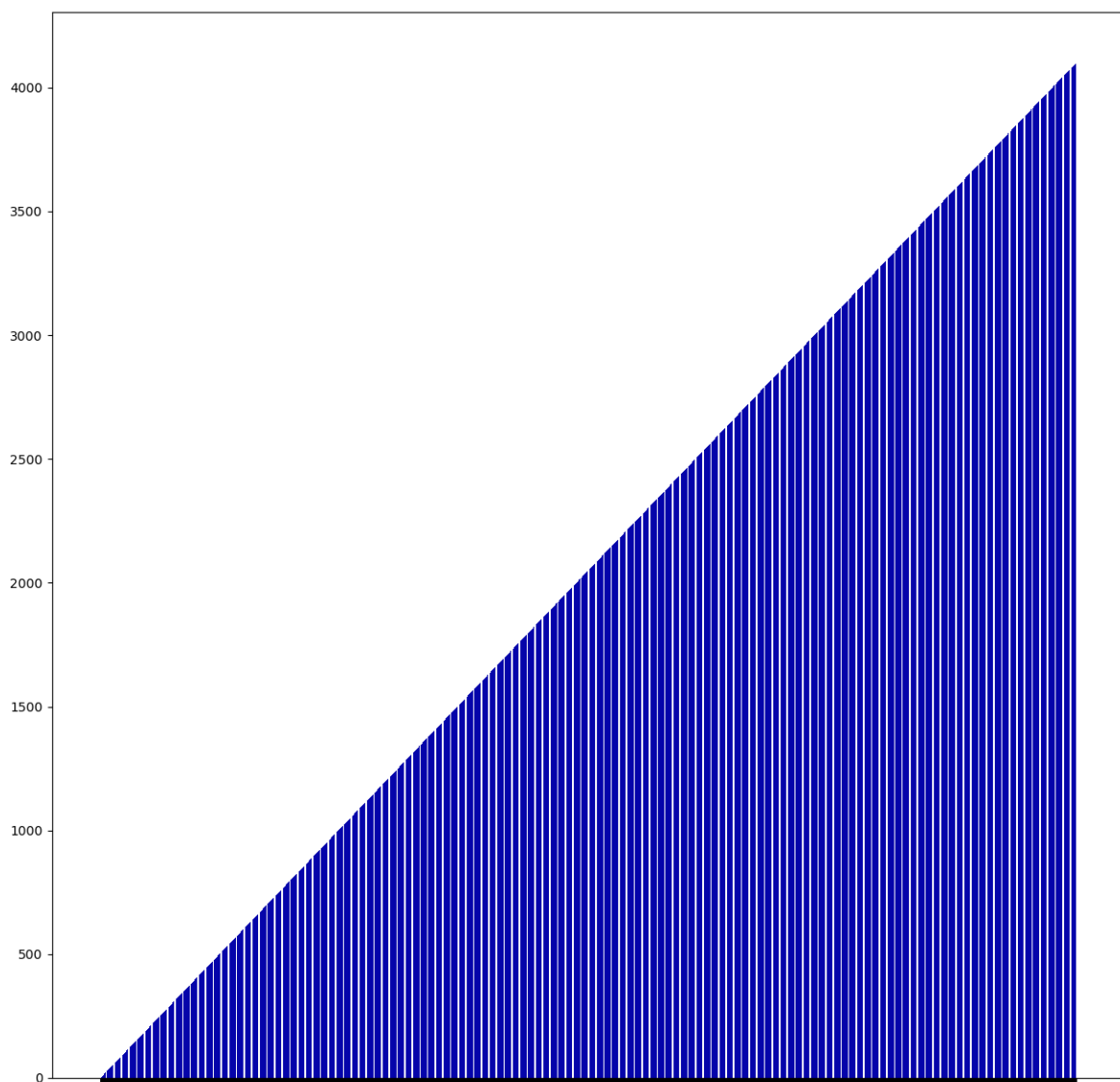
Из рисунков 4.4 и 4.5 видно, что количество сравнений, необходимых для поиска данного ключа бинарным поиском также определяется позицией ключа в словаре, но здесь зависимость более сложная, так как надо учитывать еще и ход разбиения исходного словаря. Наименьшее количество сравнений (1) потребуется для серединного ключа, далее 2 сравнения для элементов на границе $1/4$ и $3/4$, а наибольшее - $(2 * \log(N + 1)) - 1$ (в данном примере для нечетных). Для определения того, что ключа нет в словаре, потребуется столько же сравнений, сколько и для худшего случая.

Из рисунков 4.6 и 4.7 видно, что количество сравнений, необходимых для поиска данного ключа в сегментированном словаре определяется не только позицией ключа в своем сегменте, но и размером этого сегмента (так как наибольшие сегменты размещены ближе к началу). Наименьшее количество сравнений (2) потребуется для поиска самого первого ключа в самом большом сегменте, а наибольшее в общем случае определить нельзя: таковым может оказаться, например, ключ на «неудачной» позиции в самом большом сегменте или единственный ключ из самого маленького сегмента.

Для определения того, что заданный ключ отсутствует в словаре, в случае поиска в сегментированном словаре возможны два варианта. Если в словаре нет ключей, которые начинаются на ту же букву, что и искомый ключ, то понадобится столько сравнений, сколько сегментов в словаре. Если же такие ключи есть, то к количеству сравнений, которое потребовалось для поиска нужного сегмента, нужно добавить количество сравнений, которое потребуется бинарному поиску для определения отсутствия ключа в пределах сегмента.

Также для полноты картины на рисунках 4.8-4.13 приведены гистограммы, построенные для тех же алгоритмов по тому же принципу, но при

словаре из 4100 элементов.



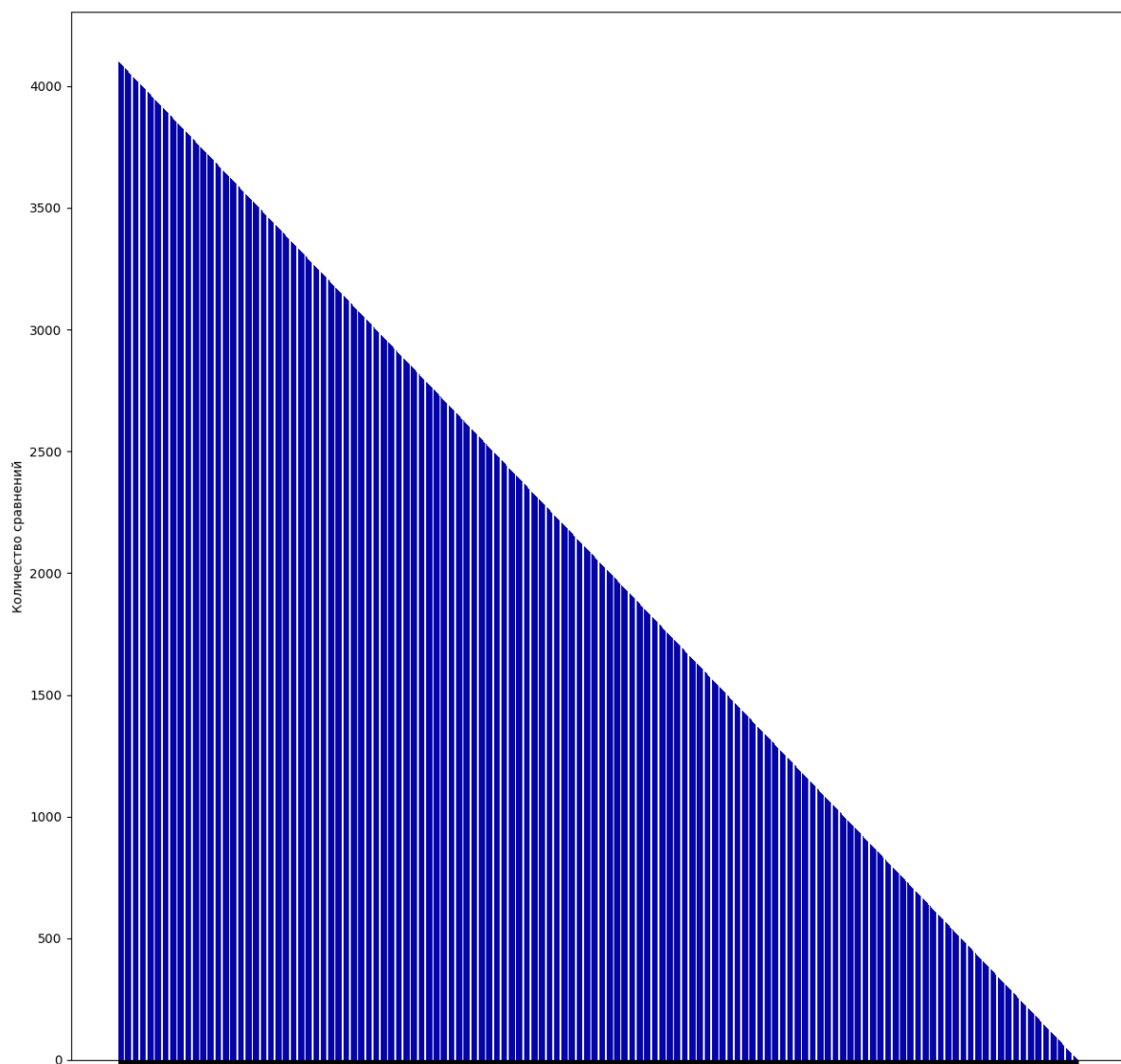


Рисунок 4.9 – Количество сравнений: полный перебор, сортировка по сравнениям, большой словарь

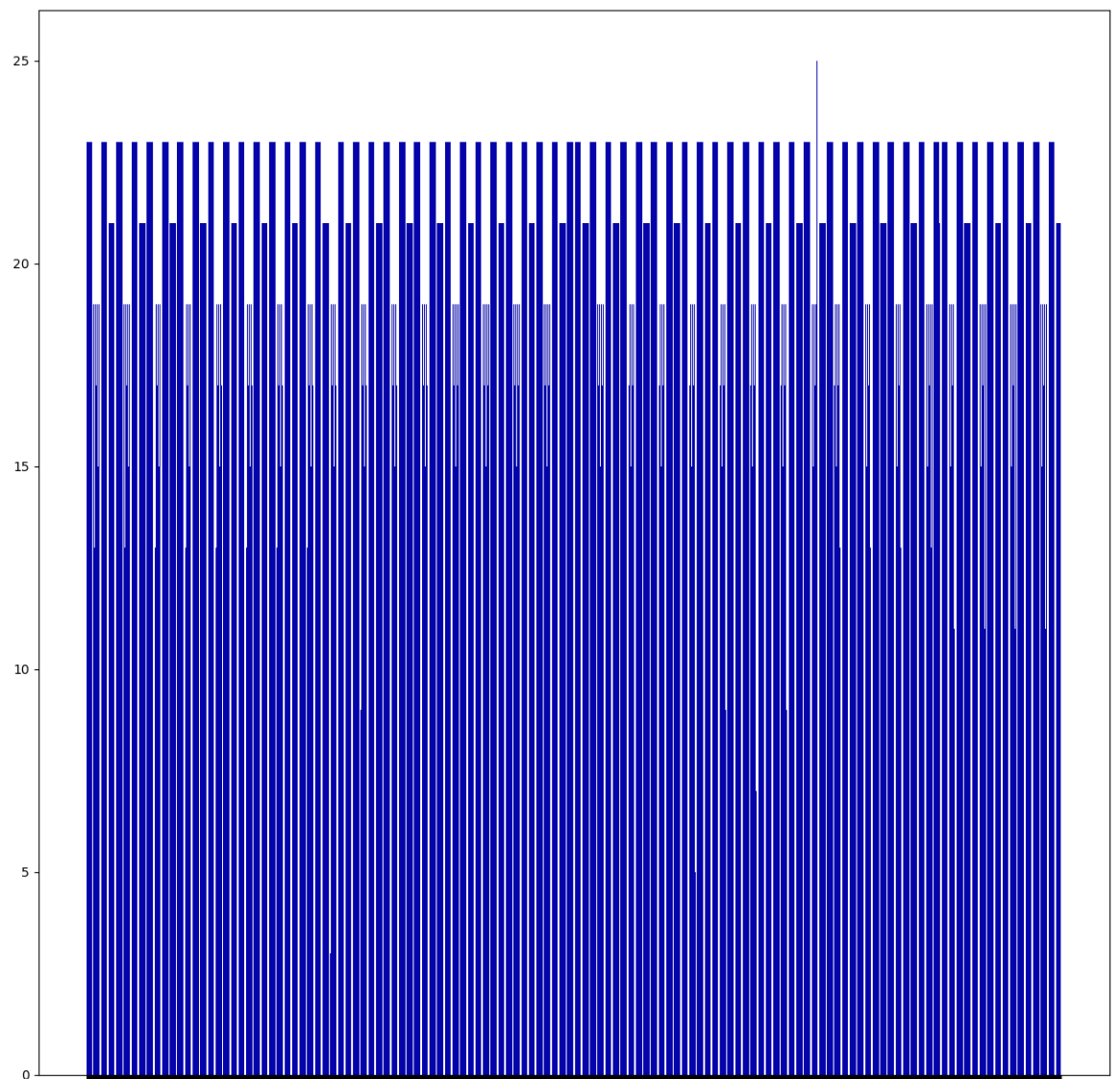
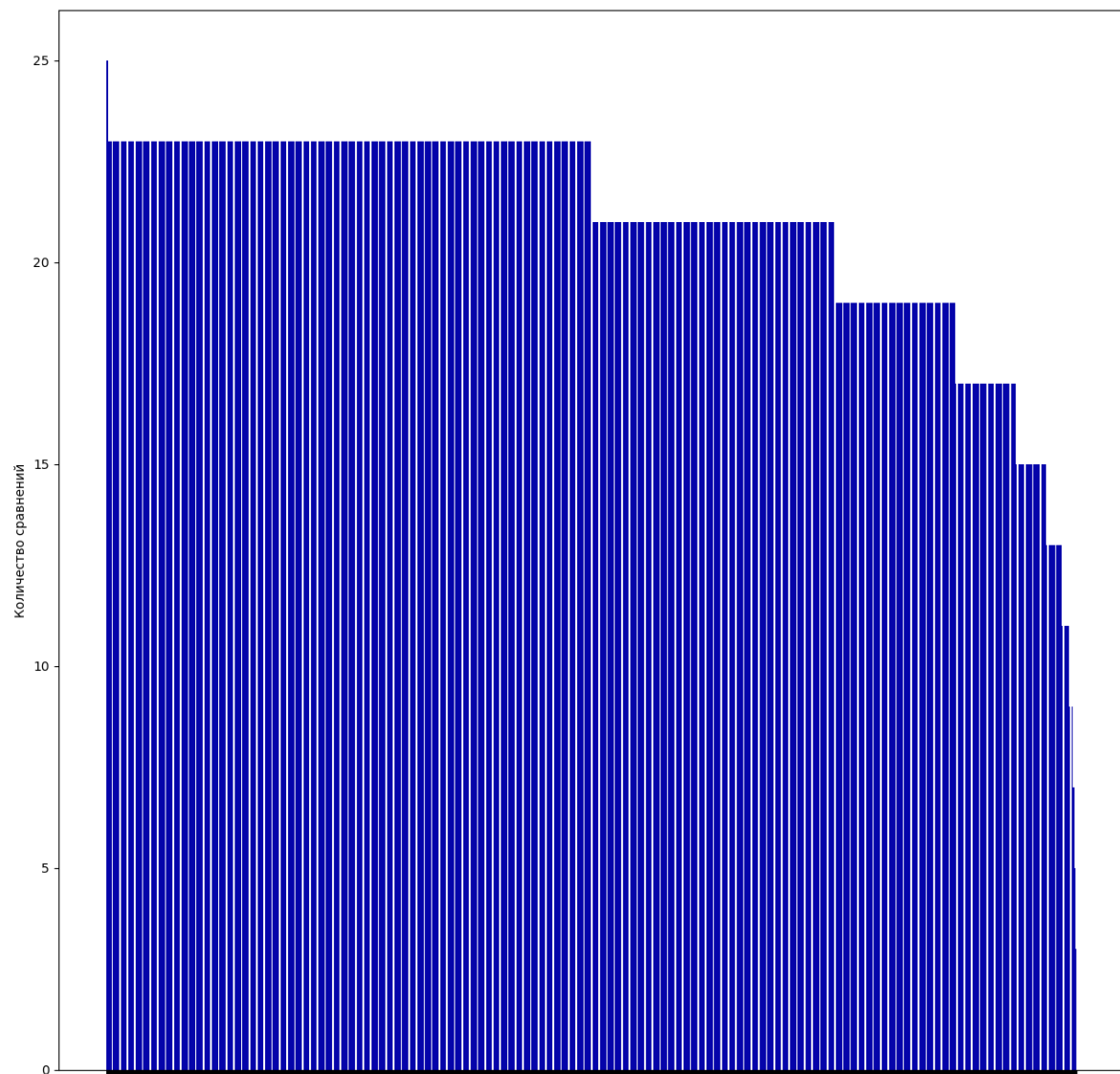


Рисунок 4.10 – Количество сравнений: бинарный поиск, сортировка по расположению, большой словарь



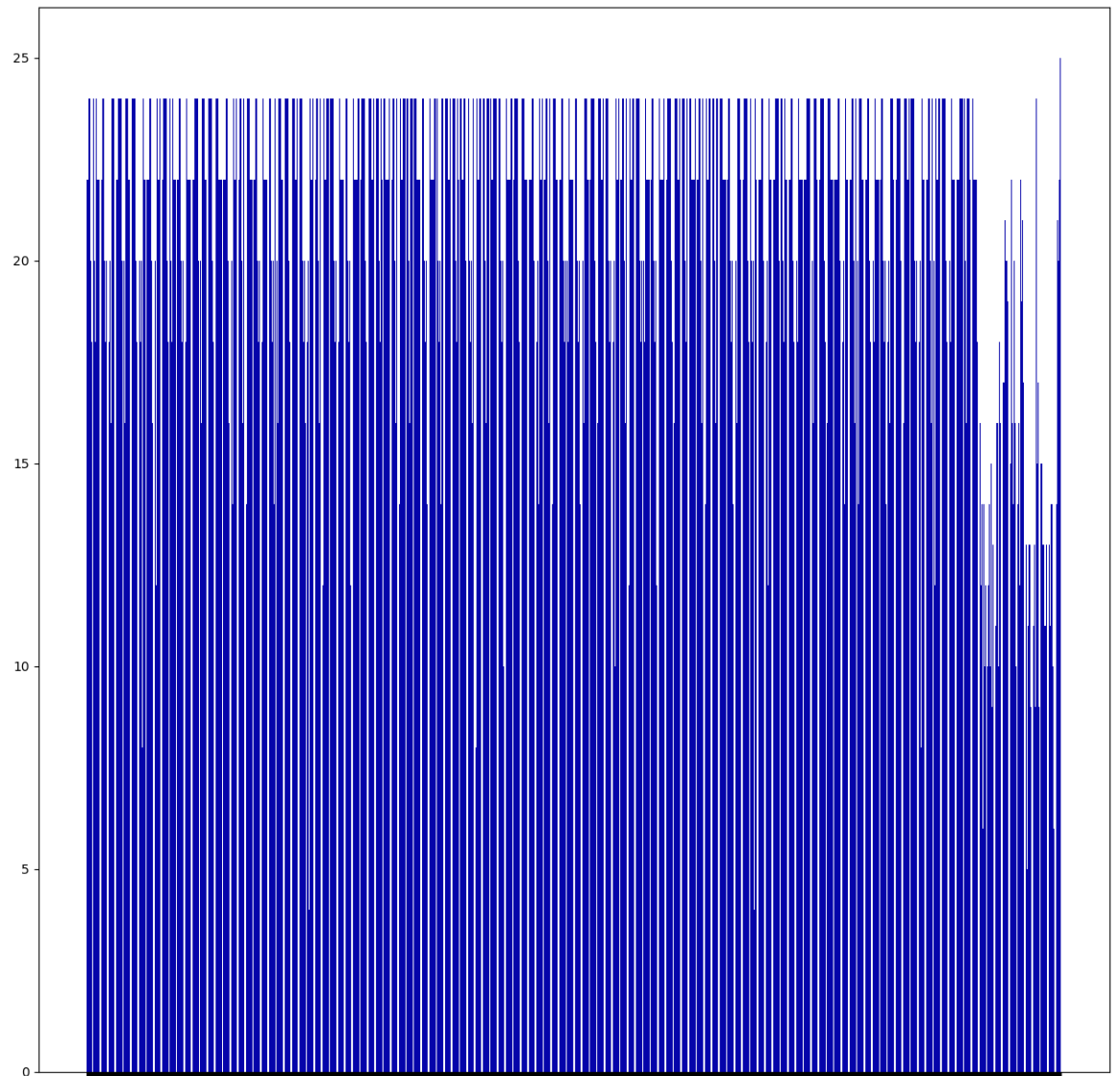


Рисунок 4.12 – Количество сравнений: поиск в сегментированном словаре, сортировка по расположению, большой словарь

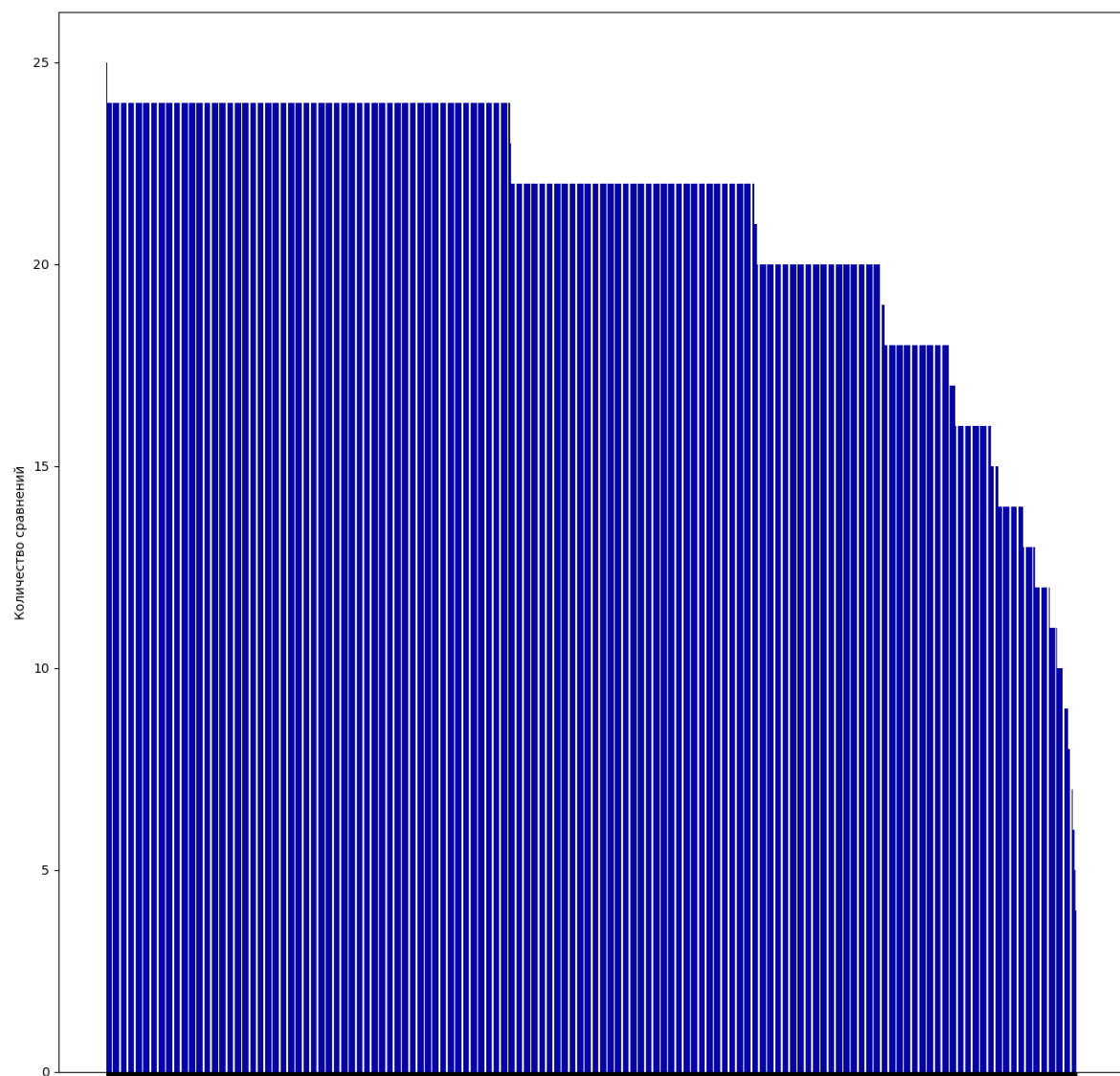


Рисунок 4.13 – Количество сравнений: поиск в сегментированном словаре, сортировка по сравнениям, большой словарь

В таблице 4.1 приведены значения минимального (min), среднего (mean), медианного (median) и максимального (max) количества сравнений, которые потребовались рассматриваемым реализациям при поиске всех ключей в словаре из 4100 элементов. В таблице введены следующие обозначения алгоритмов: full - поиск полным перебором, binary - бинарный поиск, segment - поиск в сегментированном словаре.

	full	binary	segment
min	1	1	2
mean	2050.5	21.2	21.02
median	2050	23	22
max	4100	25	25

Таблица 4.1 – Количество сравнений.

Таким образом, в общем случае наибольшее количество сравнений требуется в алгоритме полного перебора. Соотношение среднего, медианного и максимального количества сравнений в бинарном поиске и поиске в сегментированном словаре определяется размером словаря и количеством и размерами сегментов в сегментированном словаре.

4.4 Сравнение трудоемкостей

Рассмотрим трудоемкости реализаций при словаре с N ключами. Трудоемкость дается по количеству сравнений, которое потребовалось алгоритму для решения задачи.

- Поиск полным перебором

- Трудоемкость в среднем может быть рассчитана как математическое ожидание по формуле (4.1), где Ω – множество всех возможных случаев.

$$\begin{aligned}\sum_{i \in \Omega} p_i \cdot f_i &= \frac{1}{N+1} \cdot \left(\left(\sum_{i \in [1, N]} i \right) + N \right) = \\ &= \frac{1}{N+1} \cdot \left(\frac{1+N}{2} \cdot N + N \right) \sim \frac{N}{2}\end{aligned}\tag{4.1}$$

- Лучший случай - когда ключ стоит на первой позиции в словаре или отсутствует в нем. Тогда придется сравнить искомый ключ только один раз, а трудоемкость составит $O(1)$.
- Худший случай - когда ключ стоит на последней позиции в словаре, и тогда придется сравнить искомый ключ с каждым ключом словаря, а трудоемкость составит $O(n)$.
- Добавлять новый элемент в словарь можно на любую позицию, трудоемкость добавления $O(1)$.

- Бинарный поиск

- Трудоемкость в среднем составляет $O(\log N)$ [3].
- Лучший случай - когда ключ стоит на серединной позиции в словаре, и тогда придется сравнить искомый ключ только один раз, а трудоемкость составит $O(1)$.
- Первый худший случай - когда ключ стоит на такой позиции, что исходный словарь придется разбивать на части вплоть до того момента, когда левая и правая границы поиска совпадут. Тогда придется сравнить искомый ключ с каждым ключом словаря

$2 \cdot \log(N + 1) - 1$ раз и трудоемкость составит $O(\log N)$. Вторым худшим случаем с той же трудоемкостью является ситуация, когда ключ отсутствует в словаре.

- Добавлять новый элемент в словарь нужно так, чтобы он остался отсортированным, трудоемкость добавления в худшем случае $O(N)$.

- Поиск в сегментированном словаре

- Трудоемкость в среднем при множестве всех возможных случаев Ω может быть рассчитана по формуле (4.2).

$$\sum_{i \in \Omega} (f_{\text{выбор сегмента } i\text{-ого эл.}} + f_{\text{бинарный поиск } i\text{-ого эл.}}) \cdot p_i \sim \sum_{i \in \Omega} (i + \log n_i), \quad (4.2)$$

где n_i - количество элементов в i -ом сегменте. Например, для словаря из 17000 элементов, структура и процесс расчетов в котором приведены на рисунке 4.14, средняя трудоемкость в количестве сравнений составляет 17.5.

```
key: p, i: 1, ni: 1763, log(ni): 11, for each: 12, for all: 21156
key: m, i: 2, ni: 1690, log(ni): 11, for each: 13, for all: 21970
key: e, i: 3, ni: 1416, log(ni): 11, for each: 14, for all: 19824
key: i, i: 4, ni: 1401, log(ni): 11, for each: 15, for all: 21015
key: f, i: 5, ni: 1344, log(ni): 11, for each: 16, for all: 21504
key: h, i: 6, ni: 1102, log(ni): 11, for each: 17, for all: 18734
key: c, i: 7, ni: 1088, log(ni): 11, for each: 18, for all: 19584
key: a, i: 8, ni: 1065, log(ni): 11, for each: 19, for all: 20235
key: l, i: 9, ni: 971, log(ni): 10, for each: 19, for all: 18449
key: d, i: 10, ni: 958, log(ni): 10, for each: 20, for all: 19160
key: g, i: 11, ni: 888, log(ni): 10, for each: 21, for all: 18648
key: b, i: 12, ni: 845, log(ni): 10, for each: 22, for all: 18590
key: o, i: 13, ni: 841, log(ni): 10, for each: 23, for all: 19343
key: n, i: 14, ni: 704, log(ni): 10, for each: 24, for all: 16896
key: j, i: 15, ni: 222, log(ni): 8, for each: 23, for all: 5106
key: k, i: 16, ni: 222, log(ni): 8, for each: 24, for all: 5328
key: s, i: 17, ni: 203, log(ni): 8, for each: 25, for all: 5075
key: r, i: 18, ni: 102, log(ni): 7, for each: 25, for all: 2550
key: t, i: 19, ni: 80, log(ni): 7, for each: 26, for all: 2080
key: w, i: 20, ni: 38, log(ni): 6, for each: 26, for all: 988
key: v, i: 21, ni: 21, log(ni): 5, for each: 26, for all: 546
key: u, i: 22, ni: 17, log(ni): 5, for each: 27, for all: 459
key: q, i: 23, ni: 11, log(ni): 4, for each: 27, for all: 297
key: y, i: 24, ni: 5, log(ni): 3, for each: 27, for all: 135
key: z, i: 25, ni: 3, log(ni): 2, for each: 27, for all: 81

Mean: 17.514882352941175
```

Рисунок 4.14 – Пример расчета трудоемкости поиска в сегментированном словаре

- Лучший случай - когда ключ стоит на первой позиции в самом большом сегменте словаря, и тогда придется сравнить искомый ключ только 2 раза, а трудоемкость составит $O(1)$.
- Худший случай сильно зависит от того, как распределятся ключи по сегментам словаря – насколько далеко от начала окажется нужный сегмент и насколько «неудачной» для бинарного поиска окажется позиция ключа в этом сегменте. Таковым может оказаться, например, ключ на «неудачной» позиции в самом боль-

шом сегменте или единственный ключ из самого маленького сегмента.

- Для определения того, что заданный ключ отсутствует в словаре, в случае поиска в сегментированном словаре возможны два варианта. Если в словаре нет ключей, которые начинаются на ту же букву, что и искомый ключ, то понадобится столько сравнений, сколько сегментов в словаре и трудоемкость составит N . Если же такие ключи есть, то к количеству сравнений, которое потребовалось для поиска нужного сегмента, нужно добавить количество сравнений, которое потребуется бинарному поиску для определения отсутствия ключа в пределах сегмента и трудоемкость составит $i + \log n_i$, где i - номер сегмента, начинающегося на ту же букву, n_i - количество элементов в этом сегменте.
- Добавлять новый элемент в словарь нужно так, чтобы его сегменты остались отсортированы по размеру, а сегмент, в который попадет добавляемый ключ – по лексикографическому порядку.

4.5 Сравнение времени выполнения реализаций алгоритмов

Сравнивалось процессорное время работы реализаций алгоритмов поиска в словаре по ключу: алгоритма полного перебора, алгоритма бинарного поиска и алгоритма поиска в сегментированном словаре. Эти реализации сравнивались по времени работы при количестве ключей в словаре 100, 1000, 5000, 10000 и 40000.

Для каждого алгоритма и каждого количества ключей проводилось 2 замера времени. Первый был связан с поиском ключей, которые были в словаре, причем осуществлялся поиск каждого ключа, что позволило симитировать равновозможность выбора ключа (время поиска всех ключей суммировалось). Второй замер был связан с поиском ключей, которых не было в словаре. Таких «несуществующих» ключей генерировалось столько, сколько всего ключей было в словаре (время поиска всех ключей также суммировалось).

Так как некоторые задачи выполняются достаточно быстро, а замеры времени имеют некоторую погрешность, они для каждой реализации, каждого объема словаря и каждого набора ключей для поиска выполнялись 10 раз, а затем вычислялось среднее время работы.

На рисунке 4.15 приведены результаты сравнения времени выполнения реализаций алгоритмов. На графике введены следующие обозначения: full_x - алгоритм полного перебора, binary_x - алгоритм бинарного поиска, segment_x - алгоритм поиска в сегментированном словаре, x=e (exists) - время работы соответствующего алгоритма при поиске существующих ключей, x=n (not exist) – при поиске несуществующих ключей.

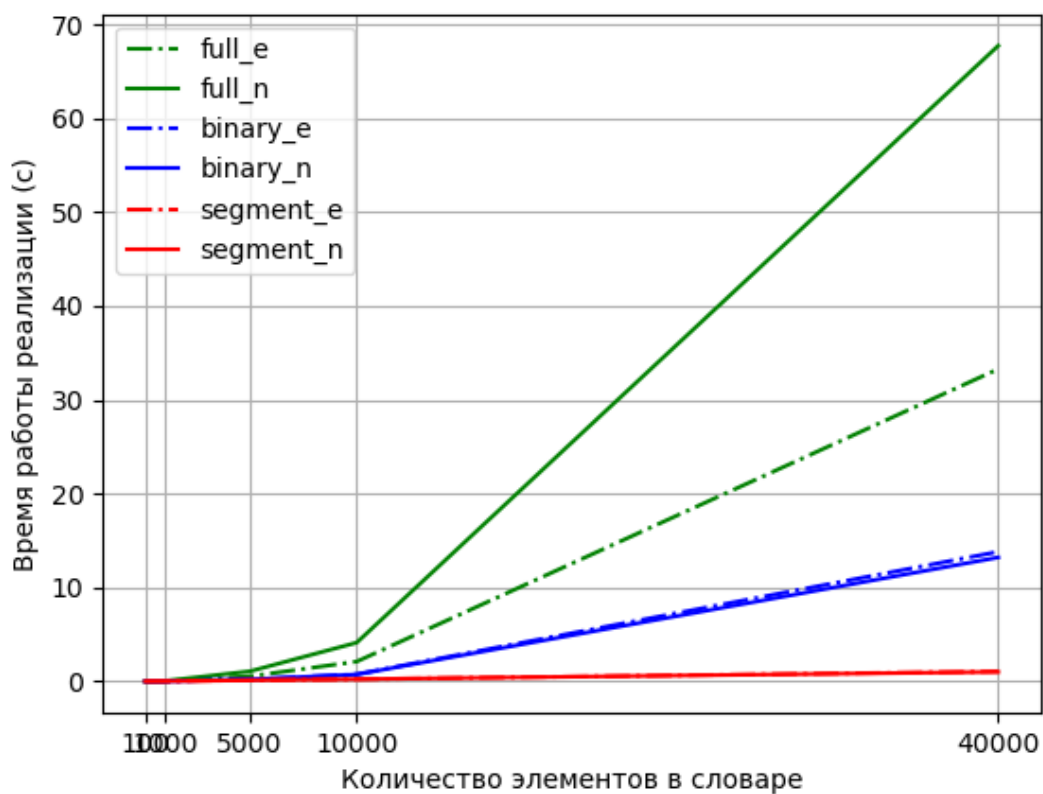


Рисунок 4.15 – Сравнение времени работы реализаций в зависимости от объема словаря

Как видно из графиков, и для поиска существующего ключа, и для определения того, что ключа в словаре нет, наибольшее количество процессорного времени требуется алгоритму полного перебора, наименьшее - поиску в сегментированном словаре. При рассмотрении каждого алгоритма отдельно можно отметить, что для поиска существующих ключей требуется меньше времени, чем для определения того, что ключа в словаре нет.

Однако необходимо учитывать, что в случае бинарного поиска может понадобиться время на предварительную сортировку словаря, а в случае поиска в сегментированном словаре - на саму сегментацию словаря и сортировку значений внутри сегментов.

С увеличением размеров словаря разница во времени работы алгоритмов увеличивается.

4.6 Вывод из исследовательской части

Таким образом, алгоритм полного перебора, как метод «грубой силы» в среднем требует больше процессорного времени для поиска ключа в словаре, но зато он не требует особого способа хранения словаря и дополнительной памяти, а новые элементы можно добавлять на любую позицию. Такой метод подходит для словаря небольших размеров.

С увеличением количества элементов в словаре стоит обращаться к менее простым методам поиска – алгоритму бинарного поиска или к сегментации словаря с сортировкой сегментов по частотности обращения и последующим бинарным поиском в пределах сегмента. Но они потребуют дополнительных затрат: первый дополнительно потребует хранения словаря в отсортированном виде, а второй – отсортированности среди сегментов (подсегментов) и в пределах самих сегментов (подсегментов), а также дополнительной памяти.

Также при использовании сегментированного словаря необходимо провести анализ для определения наиболее подходящей структуры, количества сегментов и подсегментов, признаков для их сортировки.

Заключение

В результате выполнения лабораторной работы была достигнута поставленная цель: был реализован эффективный алгоритм поиска по словаре.

В рамках выполнения работы были решены следующие задачи:

- 1) изучены три алгоритма поиска в словаре: алгоритм полного перебора, алгоритм бинарного поиска и алгоритм поиска в сегментированном словаре;
- 2) разработаны и реализованы изученные алгоритмы;
- 3) проведен сравнительный анализ трудоёмкости реализаций алгоритмов на основе теоретических расчетов (в среднем; в лучшем и худшем случаях);
- 4) проведен сравнительный анализ процессорного времени выполнения реализаций алгоритмов на основе экспериментальных данных;
- 5) проведен сравнительный анализ количества сравнений с заданным ключом, необходимых для поиска каждым алгоритмом значения по ключу и для определения отсутствия заданного ключа в словаре.

Литература

- [1] Множество и словарь [Электронный ресурс]. Режим доступа: https://acm.khpnets.info/w/index.php?title=\T2A\CYRM\T2A\cyrn\T2A\cyro\T2A\cyrrzh\T2A\cyre\T2A\cyrs\T2A\cyrt\T2A\cyrv\T2A\cyro_\T2A\cyri_\T2A\cyrs\T2A\cyr1\T2A\cyro\T2A\cyrv\T2A\cyra\T2A\cyrr\T2A\cyrsftsn._\T2A\CYRR\T2A\cyre\T2A\cyra\T2A\cyr1\T2A\cyri\T2A\cyrz\T2A\cyra\T2A\cyrc\T2A\cyri\T2A\cyrya_\T2A\cyrn\T2A\cyra_\T2A\cyrd\T2A\cyre\T2A\cyrr\T2A\cyre\T2A\cyrv\T2A\cyrsftsn\T2A\cyrya\T2A\cyrh_\T2A\cyrp\T2A\cyro\T2A\cyri\T2A\cyrs\T2A\cyrk\T2A\cyra (дата обращения: 28.11.2021).
- [2] АЛГОРИТМИЧЕСКИЕ СТРАТЕГИИ [Электронный ресурс]. Режим доступа: https://portal.tpu.ru/f_ic/files/school/materials/ppt/2.pdf (дата обращения: 28.11.2021).
- [3] Целочисленный двоичный поиск [Электронный ресурс]. Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=\T2A\CYRC\T2A\cyre\T2A\cyr1\T2A\cyro\T2A\cyrrch\T2A\cyri\T2A\cyrs\T2A\cyr1\T2A\cyre\T2A\cyrn\T2A\cyrn\T2A\cyrrery\T2A\cyrrishrt_\T2A\cyrd\T2A\cyrv\T2A\cyro\T2A\cyri\T2A\cyrrch\T2A\cyrn\T2A\cyrrery\T2A\cyrrishrt_\T2A\cyrp\T2A\cyro\T2A\cyri\T2A\cyrs\T2A\cyrk (дата обращения: 28.11.2021).
- [4] Вставка элемента в отсортированный массив [Электронный ресурс]. Режим доступа: <https://infopedia.su/17x10dd5.html> (дата обращения: 28.11.2021).
- [5] Лутц Марк. Изучаем Python, том 1, 5-е изд. Пер. с англ. — СПб.: ООО “Диалектика”, 2019. Т. 832.
- [6] time — Time access and conversions [Электронный ресурс]. Режим доступа: <https://docs.python.org/3/library/time.html> (дата обращения: 05.09.2021).
- [7] Python и Pycharm [Электронный ресурс]. Режим доступа: <https://py-charm.blogspot.com/2017/09/pycharm.html> (дата обращения: 05.09.2021).