Правительство Российской Федерации

Федеральное государственное автономное образовательное

учреждение высшего образования

«Национальный исследовательский университет

«Высшая школа экономики»

Факультет компьютерных наук

Департамент программной инженерии

Дисциплина **«Алгоритмы и структуры данных»** (часть 2)

Учебный год 2015-2016, модуль 4

**Домашнее задание** (4 части)

Отчёт о выполненной работе

**Студент:** Ярных Роман Вячеславович

**Группа:** БПИ141 (2)

**Статус представляемой работы:**

Часть 1: выполнено полностью

Часть 2: выполнено полностью

Часть 3: выполнено полностью

Часть 4: отсутствует

Москва 2016

**СОДЕРЖАНИЕ**

1. ЧАСТЬ №1 2

1.1 Постановка задачи 2

1.2 Краткий анализ задачи 2

1.3 Детализации реализации 3

1.4 Дополнительная функциональность 12

1.5 Заключение 12

2. ЧАСТЬ №2 13

2.1 Постановка задачи 13

2.2 Краткий анализ задачи 13

2.3 Детализации реализации 14

2.4 Дополнительная функциональность 21

2.5 Заключение 21

3. ЧАСТЬ №3 22

3.1 Постановка задачи 22

3.2 Краткий анализ задачи 22

3.3 Детализации реализации 23

3.4 Дополнительная функциональность 26

3.5 Заключение 26

# 1. ЧАСТЬ №1

## 1.1 Постановка задачи

Цель части №1 заключается в создании программы для поиска тех слов, прочитанных из файла, которые есть в исходном словаре, и тех, которые отсутствуют с использованием различных структур данных и алгоритмов и анализом эффективности их использования.

## 1.2 Краткий анализ задачи

Задача поиска схожих строк очень актуальна в области построения поисковых систем, анализа ДНК, нахождения грамматических ошибок и перевода текстов. Так как анализ текстов требует много времени и ресурсов для осуществления поиска строк, то необходимо изучить всевозможные алгоритмы и оценить их эффективность. В рамках данной части было реализовано четыре алгоритмы для поиска схожих строк: поиск на базе префиксных деревьев, поиск на базе PATRICIA деревьев, поиск на базе ассоциативного массива и поиск на базе хэш-таблиц.

Первый подход заключается в использовании хэш-таблиц. Хэш-таблицы представляют собой разновидность ассоциативного массива, в котором ключами являются значения хэш-функции, а в качестве значений может выступать любой тип, для которого возможно вычисление строки. Главное преимущество использования хэш-таблиц заключается в том, что поиск строки в словаре занимает в среднем O(1). Библиотека STL содержит реализацию контейнера на основе хэш-таблиц в виде класса unsorted\_set. Класс unsorted\_set использует закрытую адресацию для вставки, поиска и удаления значения из словаря.

Второй подход заключается в использовании ассоциативного массива, который реализуется через красно-черное дерево. Красно-черные деревья являются одним из видов двоичных деревьев, которые поддерживают самобалансировку. Поиск в среднем занимает O(log(n)), что хуже, чем при использовании хэш-таблиц, однако красно-черные деревья выигрывают в худшем случае, так сложность составляет O(log(n)) против O(n) у хэш-таблиц. Библиотека STL содержит реализацию контейнера на базе КЧ-деревьев в виде класса set.

Третий подход заключается в использовании префиксных деревьев. Префиксное дерево состоит из узлов, которые не содержат ключи, а являются метками в дереве. Поэтому любая строка по своей сути является набором меток, по котором можно пройти от корневого узла до листа. Каждый лист является пустым и служит только для того, чтобы сигнализировать о конце поиска. Сложность поиска, вставки и удаления в среднем, лучшем и худшем случае занимает O(k), где k – длина строки. Главное преимущество использования префиксных деревьев заключается в том, поиск производится достаточно быстро, причем скорость поиска не проседает на больших данных. Также важное отличие от других типов данных связано с особенностью потребления памяти. Так как префиксные деревья не хранят ключи, то по сравнению с хэш-таблицами и красно-черными деревьями они выигрывают по потреблению памяти у последних.

Запишем сложность каждого из методов в таблицу:

|  |  |  |  |
| --- | --- | --- | --- |
| Метод | В лучшем случае | В худшем случае | В среднем |
| Хэш-таблица | O(1) | O(n) | O(1) |
| Ассоциативный массива | O(log(n)) | O(log(n)) | O(log(n)) |
| Префиксные деревья | O(k), где k – длина слова | O(k), где k – длина слова | O(k), где k – длина слова |

## 1.3 Детализации реализации

Программа состоит из нескольких программных модулей:

1. main.cpp – чтение из файла, выполнение предложенного алгоритма поиска и вывод результата в консоль или файл;
2. AbstractDictionary.h – абстрактный класс, предоставляющий интерфейс для работы с словарем;
3. HashDictionary.h – класс, реализующий словарь на базе хэш-таблиц;
4. SetDictionary.h – класс, реализующий словарь на базе ассоциативного массива;
5. TrieDictionary.h – класс. реализующий словарь на базе префиксного дерева;
6. PTrieDictionary.h – класс, реализующий словарь на базе PATRICIA дерева.

Класс AbstractDictionary и его наследники отвечают за загрузку в память словаря, поиска слов, которые присутствуют в исходном словаре, из текста, прочитанного из файла. Опишем методы и поля классов:

|  |  |
| --- | --- |
| Dictionary | |
| Методы | void loadFromFile(istream&) – загружает словарь из входящего потока |
| void addWord(string word) – добавляет слово в словарь |
| pair<vector<string>, vector<string>> findOccurrences(istream&) – ищет все слова в файле, которые встречаются в словаре и те, которые нет |
| bool isIncluded(string word) – проверяет, есть ли слово в словаре |
| HashDictionary | |
| Поля | unsorted\_set<string> dict\_ – словарь на базе хэш-таблицы |
| Методы | void loadFromFile(istream&) – загружает словарь из входящего потока |
| void addWord(string word) – добавляет слово в словарь |
| pair<vector<string>, vector<string>> findOccurrences(istream&) – ищет все слова в файле, которые встречаются в словаре и те, которые нет |
| bool isIncluded(string word) – проверяет, есть ли слово в словаре |
| HashDictionary() – конструктор |
| ~HashDictionary() – дескруктор |
| SetDictionary | |
| Поля | set<string> dict\_ – словарь на базе ассоциативного массива |
| Методы | void loadFromFile(istream&) – загружает словарь из входящего потока |
| void addWord(string word) – добавляет слово в словарь |
| pair<vector<string>, vector<string>> findOccurrences(istream&) – ищет все слова в файле, которые встречаются в словаре и те, которые нет |
| bool isIncluded(string word) – проверяет, есть ли слово в словаре |
| SetDictionary() – конструктор |
| ~SetDictionary() – дескруктор |
| TrieDictionary | |
| Поля | Node head\_ – корневой узел префиксного дерева |
| Методы | void loadFromFile(istream&) – загружает словарь из входящего потока |
| void addWord(string word) – добавляет слово в словарь |
| pair<vector<string>, vector<string>> findOccurrences(istream&) – ищет все слова в файле, которые встречаются в словаре и те, которые нет |
| bool isIncluded(string word) – проверяет, есть ли слово в словаре |
| TrieDictionary() – конструктор |
| ~TrieDictionary() – дескруктор |
| PTrieDictionary | |
| Поля | Node head\_ – корневой узел PATRICIA дерева |
| Методы | void loadFromFile(istream&) – загружает словарь из входящего потока |
| void addWord(string word) – добавляет слово в словарь |
| pair<vector<string>, vector<string>> findOccurrences(istream&) – ищет все слова в файле, которые встречаются в словаре и те, которые нет |
| bool isIncluded(string word) – проверяет, есть ли слово в словаре |
| PTrieDictionary() – конструктор |
| ~PTrieDictionary() – дескруктор |

Покажем взаимосвязь классов в программе с помощью UML диаграммы:



Приведем исходные коды реализации. Исходный код класса Dictionary:

#pragma once  
#ifndef **CHW\_2\_ABSTRACTDICTIONARY\_H**#define **CHW\_2\_ABSTRACTDICTIONARY\_H**#include **<iostream>**#include **<vector>  
  
using namespace** std;  
  
**class** Dictionary {  
**public**:  
 **virtual void** loadDictFromFile(istream& input);  
 **virtual void** addWord(string word) = 0;  
 **virtual** pair<vector<string>, vector<string>> findOccurrences(istream& input);  
 **virtual bool** isIncluded(string word) = 0;  
 **virtual** ~Dictionary() = 0;  
};  
  
**void** Dictionary::loadDictFromFile(istream &input) {  
 string temp\_;  
 **while**(getline(input, temp\_)){  
 addWord(temp\_);  
 }  
}  
  
pair<vector<string>, vector<string>> Dictionary::findOccurrences(istream &input) {  
 string temp\_;  
 vector<string> occurrences\_;  
 vector<string> notIncluded\_;  
 **while**(input >> temp\_){  
 stringstream ss;  
 **bool** lastNonAlpha = **false**;  
 **for**(string::iterator it = temp\_.begin(); it != temp\_.end(); it++){  
 **if**((\*it < **'a'** || \*it > **'z'**) && !lastNonAlpha) {  
 string word\_ = ss.str();  
 **if** (isIncluded(word\_)) {  
 occurrences\_.push\_back(word\_);  
 }  
 **else**{  
 notIncluded\_.push\_back(word\_);  
 }  
 ss.str(**""**);  
 lastNonAlpha = **true**;  
 }  
 **else if**(isalpha(\*it)){  
 lastNonAlpha = **false**;  
 ss << (\*it);  
 }  
 }  
 **if**(ss.str().size() > 0){  
 string word\_ = ss.str();  
 **if** (isIncluded(word\_)) {  
 occurrences\_.push\_back(word\_);  
 }  
 **else**{  
 notIncluded\_.push\_back(word\_);  
 }  
 }  
 }  
 **return** pair<vector<string>, vector<string>>(occurrences\_, notIncluded\_);  
}  
  
Dictionary::~Dictionary(){  
  
}  
  
#endif *//CHW\_2\_ABSTRACTDICTIONARY\_H*

Исходный код класса HashDictionary:

#pragma once  
#ifndef **CHW\_2\_HASHDICTIONARY\_H**#define **CHW\_2\_HASHDICTIONARY\_H**#include **<unordered\_set>**#include **<sstream>**#include **"Dictionary.h"  
  
class** HashDictionary : **public** Dictionary {  
**public**:  
 **void** addWord(string word);  
 **bool** isIncluded(string word);  
 HashDictionary();  
 **virtual** ~HashDictionary();  
**private**:  
 unordered\_set<string> \*dict\_;  
};  
  
**void** HashDictionary::addWord(string word) {  
 dict\_->insert(word);  
}  
  
**bool** HashDictionary::isIncluded(string word){  
 **return** dict\_->find(word) != dict\_->end();  
}  
  
HashDictionary::HashDictionary() {  
 dict\_ = **new** unordered\_set<string>();  
}  
  
HashDictionary::~HashDictionary() {  
 **delete** dict\_;  
}  
  
#endif *//CHW\_2\_HASHDICTIONARY\_H*

Исходный код класса SetDictionary:

#ifndef **CHW\_2\_SETDICTIONARY\_H**#define **CHW\_2\_SETDICTIONARY\_H**#include **<set>**#include **"Dictionary.h"  
  
class** SetDictionary : **public** Dictionary {  
  
**public**:  
  
 **virtual void** addWord(string word);  
  
 **virtual bool** isIncluded(string word);  
  
 SetDictionary();  
  
 **virtual** ~SetDictionary();  
  
**private**:  
  
 set<string> \*dict\_;  
};  
  
**void** SetDictionary::addWord(string word) {  
 dict\_->insert(word);  
}  
  
**bool** SetDictionary::isIncluded(string word) {  
 **return** dict\_->find(word) != dict\_->end();  
}  
  
SetDictionary::SetDictionary() {  
 dict\_ = **new** set<string>;  
}  
  
SetDictionary::~SetDictionary() {  
 **delete** dict\_;  
}  
  
#endif *//CHW\_2\_SETDICTIONARY\_H*

Исходный код класса TrieDictionary:

#ifndef **CHW\_2\_TRIEDICTIONARY\_H**#define **CHW\_2\_TRIEDICTIONARY\_H**#include **"Dictionary.h"**#include **"StringTrie.h"  
  
class** TrieDictionary : **public** Dictionary {  
  
**public**:  
 **void** addWord(string word);  
  
 **bool** isIncluded(string word);  
  
 TrieDictionary();  
  
 **virtual** ~TrieDictionary();  
  
**private**:  
 StringTrie \*trie\_;  
};  
  
**void** TrieDictionary::addWord(string word) {  
 trie\_->add(word);  
}  
  
**bool** TrieDictionary::isIncluded(string word) {  
 **return** trie\_->contains(word);  
}  
  
TrieDictionary::TrieDictionary() {  
 trie\_ = **new** StringTrie;  
}  
TrieDictionary::~TrieDictionary() {  
 **delete** trie\_;  
}  
  
#endif *//CHW\_2\_TRIEDICTIONARY\_H*

Исходный код класса StringTrie:

#ifndef **CHW\_2\_STRINGTRIE\_H**#define **CHW\_2\_STRINGTRIE\_H**#define **ALPH\_N** 26  
#include **"Trie.h"  
  
class** StringTrie : **public** Trie{  
**private**:  
 **struct** Node{  
 **char** key;  
 Node\* children[**ALPH\_N**];  
 };  
 Node head\_;  
**public**:  
 **void** add(string key);  
 **bool** contains(string key);  
 **void** \_clear(Node\* node);  
 **virtual** ~StringTrie();  
};  
  
**void** StringTrie::add(string key) {  
 Node\* temp\_ = &head\_;  
 **for**(string::iterator it = key.begin(); it != key.end(); it++){  
 Node\* child\_ = temp\_->children[\*it - **'a'**];  
 **if**(!child\_){  
 child\_ = **new** Node;  
 child\_->key = \*it;  
 temp\_->children[\*it - **'a'**] = child\_;  
 }  
 temp\_ = child\_;  
 }  
}  
  
**bool** StringTrie::contains(string key) {  
 **bool** inc\_ = **true**;  
 Node \*temp\_ = &head\_;  
 **for**(string::iterator it = key.begin(); it != key.end(); it++){  
 **if**(!temp\_->children[\*it - **'a'**]){  
 inc\_ = **false**;  
 **break**;  
 }  
 **else**{  
 temp\_ = temp\_->children[\*it - **'a'**];  
 }  
 }  
 **return** inc\_;  
}  
  
**void** StringTrie::\_clear(Node\* node){  
 **if**(!node)  
 **return**;  
 **for**(**int** i = 0; i < **ALPH\_N**; i++){  
 **if**(!node->children[i]) {  
 \_clear(node->children[i]);  
 **delete** node->children[i];  
 }  
 }  
}  
  
StringTrie::~StringTrie() {  
 Node\* temp\_ = &head\_;  
 \_clear(temp\_);}  
  
#endif *//CHW\_2\_STRINGTRIE\_H*

## 1.4 Дополнительная функциональность

Дополнительная функциональность реализации отсутствует.

## 1.5 Заключение

# 2. ЧАСТЬ №2

## 2.1 Постановка задачи

Цель части №2 заключается в написании программного кода для оценки эффективности использования различных комбинаций контейнеров и алгоритмов для поиска слов, отличных на один символ и имеющих один и тот же размер.

## 2.2 Краткий анализ задачи

В соответствии с целью были поставлены следующие задачи:

1. разработать алгоритмы (методы) для поиска слов, отличающихся на один символ;
2. оценить эффективность использованных алгоритмов и контейнеров;
3. сделать выводы по поводу использования данных алгоритмов на практике.

Для решения поставленных задач было разработано три метода поиска слов:

1. наивный поиск с использованием std::map<string, vector<string>>;
2. улучшенный поиск с использованием std::map<string, vector<string>> и предварительной группировкой слов по их длине;
3. поиск с предварительной группировкой слов по длине, а также по общему префиксу и суффиксу.

Первый метод заключается в наивном поиске слов, путем прохода по всему массиву и сравниванию одного слова с другим. В случае, если одно слово может быть получено путем изменения буквы в другом, то для первого в список заносится второе. Данный метод имеет сложность в лучшем худшем и среднем случаях , что является, несомненно, самым плохим вариантом поиска.

Второй метод имеет асимптотически такую же сложность, что и первый, но при его использовании улучшается коэффициент при за счет уменьшения числа сравнений путем группировки слов по длине. При таком методе слова сравниваются только в своей группе и имеют одну и ту же длину, что позволяет не производить заведомо лишние сравнения. В общем, сложность второго подхода равна , где – время группировки слов по длине. Также важно отметить, что для второго метода требуется дополнительной памяти.

Третий вариант улучшает изначальный наивный поиск с помощью группировки слов не только по размеру, но и общему префиксу и суффиксу. Если для группировки слов с общим префиксом и суффиксом используется std::set<string>, то сложность по времени в худшем, лучшем и среднем случае составит , где – размер префиксного или суффиксного массива. Заметим, что если все слова имеют один префикс или суффикс, при условии , то , следовательно, сложность метода составит . При использовании std::unordered\_set<string> сложность метода в лучшем и среднем случае составит , в худшем – . Как и во втором методе, потребуется дополнительной памяти.

## 2.3 Детализации реализации

Было решено перенести функциональность поиска схожих строк в новые классы, не входящие в иерархию классов, ответственных за реализацию работы словарей. С этой целью были разработаны классы FuzzySearcher, NaiveSearcher, OptNaiveSearcher и QuickSearcher. Вершиной иерархии выступает FuzzySearcher, определяющий интерфейс методов для конечного клиента. Классы NaiveSearcher, OptNaiveSearcher и QuickSearcher реализует каждый из предложенных 3 методов. Опишем структуру классов:

|  |  |
| --- | --- |
| FuzzySearcher | |
| Поля | set<string> \*dict\_ – исходный словарь |
| Методы | FuzzySearcher() – конструктор по умолчанию |
| FuzzySearcher(set<string>) – конструктор, принимающий в качестве аргумента исходный словарь |
| ~FuzzySearcher() – деструктор |
| map<string, set<string>> fuzzySearch() – поиск строк, отличных на один символ |
| bool \_areDifferentByOneSym(string, string) – сравнение строк |
| string \_getPrefix(string) – получение префикса строки |
| string \_getSuffix(string) – получение суффикса строки |
| NaiveSearcher | |
| Поля | set<string> \*dict\_ – исходный словарь |
| Методы | NaiveSearcher() – конструктор по умолчанию |
| NaiveSearcher(set<string>) – конструктор, принимающий в качестве аргумента исходный словарь |
| ~NaiveSearcher() – деструктор |
| map<string, set<string>> fuzzySearch() – поиск строк, отличных на один символ |
| bool \_areDifferentByOneSym(string, string) – сравнение строк |
| string \_getPrefix(string) – получение префикса строки |
| string \_getSuffix(string) – получение суффикса строки |
| OptNaiveSearcher | |
| Поля | set<string> \*dict\_ – исходный словарь |
| Методы | OptNaiveSearcher() – конструктор по умолчанию |
| OptNaiveSearcher(set<string>) – конструктор, принимающий в качестве аргумента исходный словарь |
| ~OptNaiveSearcher() – деструктор |
| map<string, set<string>> fuzzySearch() – поиск строк, отличных на один символ |
| bool \_areDifferentByOneSym(string, string) – сравнение строк |
| string \_getPrefix(string) – получение префикса строки |
| string \_getSuffix(string) – получение суффикса строки |
| QuickSearcher | |
| Поля | set<string> \*dict\_ – исходный словарь |
| Методы | QuickSearcher() – конструктор по умолчанию |
| QuickSearcher(set<string>) – конструктор, принимающий в качестве аргумента исходный словарь |
| ~QuickSearcher() – деструктор |
| map<string, set<string>> fuzzySearch() – поиск строк, отличных на один символ |
| bool \_areDifferentByOneSym(string, string) – сравнение строк |
| string \_getPrefix(string) – получение префикса строки |
| string \_getSuffix(string) – получение суффикса строки |

Приведем исходные коды классов. Исходный код класса FuzzySearcher:

#pragma once  
#ifndef **CHW\_2\_FUZZYSEARCHER\_H**#define **CHW\_2\_FUZZYSEARCHER\_H**#include **<iostream>**#include **<string>**#include **<set>  
  
using namespace** std;  
  
**class** FuzzySearcher {  
**public**:  
 FuzzySearcher();  
 FuzzySearcher(set<string> ws);  
 **virtual** ~FuzzySearcher();  
 **virtual** map<string, set<string>> fuzzySearch() = 0;  
  
**protected**:  
 **int** \_hamming\_distance(string a, string b);  
 **bool** \_areDifferentByOneSym(string a, string b);  
 string \_getPrefix(string s);  
 string \_getSuffix(string s);  
 set<string> \*dict\_;  
};  
  
FuzzySearcher::FuzzySearcher() {  
  
}  
  
FuzzySearcher::FuzzySearcher(set<string> ws) {  
 dict\_ = **new** set<string>(move(ws));  
}  
  
FuzzySearcher::~FuzzySearcher() {  
 **delete** dict\_;  
}  
  
**int** FuzzySearcher::\_hamming\_distance(string a, string b) {  
 **const char**\* a\_ = a.c\_str();  
 **const char**\* b\_ = b.c\_str();  
 **int** d\_ = 0;  
 **for**(**int** i = 0; i < a.length(); i++){  
 d\_ += ((a\_[i] - b\_[i]) == 0 ? 0 : 1);  
 }  
 **return** d\_;  
}  
  
**bool** FuzzySearcher::\_areDifferentByOneSym(string a, string b) {  
 **bool** dif\_ = **false**;  
 **for**(**int** i = 0; i < a.length(); i++){  
 **if**(a[i] != b[i]){  
 **if**(dif\_) {  
 dif\_ = **false**;  
 **break**;  
 }  
 **else** dif\_ = **true**;  
 }  
 }  
 **return** dif\_;  
}  
  
string FuzzySearcher::\_getPrefix(string s) {  
 **int** prefixEnd\_ = (**int**)floor(s.length() / 2);  
 **return** s.substr(0, (**unsigned long**)prefixEnd\_ + 1);  
}  
  
string FuzzySearcher::\_getSuffix(string s) {  
 **int** prefixEnd\_ = (**int**)floor(s.length() / 2);  
 **return** s.substr((**unsigned long**)prefixEnd\_ + 1, s.length() - prefixEnd\_ - 1);  
}  
  
  
#endif *//CHW\_2\_FUZZYSEARCHER\_H*

Исходный код NaiveSearcher:

#ifndef **CHW\_2\_NAIVYSEARCHER\_H**#define **CHW\_2\_NAIVYSEARCHER\_H**#include **"FuzzySearcher.h"**#include **<map>  
  
class** NaiveSearcher : **public** FuzzySearcher{  
  
**public**:  
 NaiveSearcher(set<string> s) : FuzzySearcher(s){}  
 **virtual** map<string, set<string>> fuzzySearch();  
};  
  
  
#endif *//CHW\_2\_NAIVYSEARCHER\_H*map<string, set<string>> NaiveSearcher::fuzzySearch() {  
 map<string, set<string>> fsmw\_;  
 **for**(**auto** it = dict\_->begin(); it != dict\_->end(); it++){  
 set<string> mw\_;  
 string current\_ = \*it;  
 **for**(**auto** it1 = dict\_->begin(); it1 != dict\_->end(); it1++){  
 string other\_ = \*it1;  
 **if**(other\_.length() != current\_.length())  
 **continue**;  
 **if**(\_areDifferentByOneSym(current\_, other\_))  
 mw\_.insert(other\_);  
 }  
 fsmw\_.insert(pair<string, set<string>>(current\_, mw\_));  
 }  
 **return** fsmw\_;  
}

Исходный код OptNaiveSearcher:

#ifndef **CHW\_2\_OPTNAIVESEARCHER\_H**#define **CHW\_2\_OPTNAIVESEARCHER\_H**#include **"FuzzySearcher.h"**#include **<map>  
  
class** OptNaiveSearcher : **public** FuzzySearcher{  
  
**public**:  
 OptNaiveSearcher(set<string> s) : FuzzySearcher(s){}  
 **virtual** map<string, set<string>> fuzzySearch();  
};  
  
  
#endif *//CHW\_2\_OPTNAIVESEARCHER\_H*map<string, set<string>> OptNaiveSearcher::fuzzySearch() {  
 map<string, set<string>> fsmw\_;  
 map<**unsigned long**, set<string>> lg\_;  
 **for**(**auto** it = dict\_->begin(); it != dict\_->end(); it++){  
 **if**(lg\_.find(it->length()) != lg\_.end()){  
 lg\_[it->length()].insert(\*it);  
 }  
 **else**{  
 set<string> clws\_;  
 clws\_.insert(\*it);  
 lg\_.insert(pair<**unsigned long**, set<string>>(it->length(), clws\_));  
 }  
 }  
 **for**(**auto** lg\_it = lg\_.begin(); lg\_it != lg\_.end(); lg\_it++){  
 **for**(set<string>::iterator word\_ = lg\_it->second.begin(); word\_ != lg\_it->second.end(); word\_++){  
 set<string> mw\_;  
 **for**(set<string>::iterator oword\_ = lg\_it->second.begin(); oword\_ != lg\_it->second.end(); oword\_++){  
 **if**(\_areDifferentByOneSym(\*word\_, \*oword\_))  
 mw\_.insert(\*oword\_);  
 }  
 fsmw\_.insert(pair<string, set<string>>(\*word\_, mw\_));  
 }  
 }  
 **return** fsmw\_;  
}

Исходный код QuickSearcher:

#ifndef **CHW\_2\_QUICKSEARCHER\_H**#define **CHW\_2\_QUICKSEARCHER\_H**#include **"FuzzySearcher.h"**#include **<map>**#include **<unordered\_set>**#include **<unordered\_map>  
  
class** QuickSearcher : **public** FuzzySearcher{  
  
**public**:  
 QuickSearcher(set<string> s) : FuzzySearcher(s) {}  
 **virtual** map<string, set<string>> fuzzySearch();  
};  
  
  
#endif *//CHW\_2\_QUICKSEARCHER\_H*map<string, set<string>> QuickSearcher::fuzzySearch() {  
 map<string, set<string>> fsmw\_;  
 map<**unsigned long**, set<string>> lg\_;  
 **for**(**auto** it = dict\_->begin(); it != dict\_->end(); it++){  
 **if**(lg\_.find(it->length()) != lg\_.end()){  
 lg\_[it->length()].insert(\*it);  
 }  
 **else**{  
 set<string> clws\_;  
 clws\_.insert(\*it);  
 lg\_.insert(pair<**unsigned long**, set<string>>(it->length(), clws\_));  
 }  
 }  
 **for**(**auto** lg\_it = lg\_.begin(); lg\_it != lg\_.end(); lg\_it++){  
 unordered\_map<string, unordered\_set<string>> prefixMap\_;  
 unordered\_map<string, unordered\_set<string>> suffixMap\_;  
 **for**(**auto** word\_ = lg\_it->second.begin(); word\_ != lg\_it->second.end(); word\_++){  
 string prefix\_ = \_getPrefix(\*word\_);  
 string suffix\_ = \_getSuffix(\*word\_);  
 **if**(prefixMap\_.find(prefix\_) != prefixMap\_.end()){  
 prefixMap\_[prefix\_].insert(suffix\_);  
 }  
 **else**{  
 prefixMap\_[prefix\_] = unordered\_set<string>();  
 }  
  
 **if**(suffixMap\_.find(suffix\_) != suffixMap\_.end()){  
 suffixMap\_[suffix\_].insert(prefix\_);  
 }  
 **else**{  
 suffixMap\_[suffix\_] = unordered\_set<string>();  
 }  
 }  
 **for**(**auto** word\_ = lg\_it->second.begin(); word\_ != lg\_it->second.end(); word\_++){  
 set<string> mw\_;  
 string ownPrefix\_ = \_getPrefix(\*word\_);  
 string ownSuffix\_ = \_getSuffix(\*word\_);  
 unordered\_set<string> sufs\_ = prefixMap\_[ownPrefix\_];  
 unordered\_set<string> prefs\_ = suffixMap\_[ownSuffix\_];  
 **for**(**auto** suf\_ = sufs\_.begin(); suf\_ != sufs\_.end(); suf\_++){  
 **if**(\_areDifferentByOneSym(ownSuffix\_, \*suf\_))  
 mw\_.insert(ownPrefix\_ + (\*suf\_));  
 }  
 **for**(**auto** pref\_ = prefs\_.begin(); pref\_ != prefs\_.end(); pref\_++){  
 **if**(\_areDifferentByOneSym(ownPrefix\_, \*pref\_))  
 mw\_.insert((\*pref\_) + ownSuffix\_);  
 }  
 fsmw\_.insert(pair<string, set<string>>(\*word\_, mw\_));  
 }  
 }  
 **return** fsmw\_;  
}

## 2.4 Дополнительная функциональность

Дополнительная функциональность реализации отсутствует.

## 2.5 Заключение

# 3. ЧАСТЬ №3

## 3.1 Постановка задачи

Цель части №3 разработать методы для поиска ошибочных слов и вариантов их замены путем удаления одного символа, вставки одного символа и транспозиции букв в строке, разработать программу для реализации методов и оценить их эффективность.

## 3.2 Краткий анализ задачи

Поиск слов, написанных с ошибкой, и вариантов их исправления является неотъемлемой частью системы проверки орфография современных программ для редактирования документов. В соответствии с предложенной проблемой были поставлены следующие задачи:

1. разработать методы для поиска ошибочных слов в тексте и вариантов их исправления путем преобразований через вставку или удаления одного символа, либо транспозицию букв внутри строки;
2. написать программу, реализующие данные методы;
3. оценить эффективность использования данных методов;
4. сделать выводы по поводу использования данных методов на практике.

Перед тем, как находить ошибки в тексте, необходимо составить словарь, включающие эталонные слова для проверки орфографии. Словарь целесообразно хранить в тех структурах данных, которые позволяют сократить время поиска до максимума. В нашей реализации было решено использовать хэш-таблицы, так как они обеспечивают в большинстве случаев линейную сложность поиска. В редких случаях возможно возникновение коллизий, которые приводят к тому, что время поиска уже не линейно, а квадратично. Но в среднем время поиска в хэш-таблице лучшем, чем в красно-черных деревьях, чья сложность поиска составляет .

При сканировании текста, программа ищет слово в словаре и если его не находит, что фиксирует это слово как ошибочное, сохраняет его номер, а затем производит поиск вариантов замены. Замена может быть произведена либо через удаление одного символа из исходного слова или вставки в него символа, либо через транспозицию букв.

Первый метод производит поиск все тех слов в словаре, которые могут быть получены путем перестановки (транспозиции) букв в исходном слове. Наивный поиск даст нам сложность , что не приемлемо для работы с большим текстом и словарем, поэтому необходимо снизить сложность до приемлемого уровня. Снизить сложность можно путем предварительной группировки слов по длине и общему набору символов, входящих в эти слова. Транспозиция обладает таким свойством, что слова одинаковой длины и отличные лишь перестановкой символов имеют одинаковый набор символов, входящих в них. Распишем последовательность шагов выбранного метода:

1. загружаем словарь из исходного файла;
2. группируем слова по длине строки;
3. в каждой группе разбиваем слова на новые группы (ассоциативными массивами), ключами которых являются отсортированная последовательность символов исходных строк, а значения – сами строки;
4. для проверяемого слова вычисляем длину строки;
5. длина строки позволит нам найти нужную группу слов, в которых мы и будем выбирать варианты замены;
6. в найденной группе находим слова, которые имеют тот же набор символов, что и проверяемое слово;
7. эти слова и будут вариантами замены.

Оценим сложность данного метода. Группировка слов по длине занимает . Группировка по набору символов занимает также . Поиск вариантов замены имеет сложность в лучшем и среднем случае , если используются хэш-таблицы, в худшем случае – . Если же используются красно-черные деревья, то сложность составляет в лучшем, среднем и худшем случае .

## 3.3 Детализации реализации

Для реализации разработанных методов был написан класс SpellSearcher, который отвечает за проверку слова на предмет орфографической ошибки и вариантов замены. Опишем структуру класса:

|  |  |
| --- | --- |
| SpellSearcher | |
| Поля | set<string> \*dict\_ – исходный словарь |
| unordered\_map<unsigned long, unordered\_map<string, unordered\_ser<string>>> ltg\_ – группа слов с одинаковой длиной и набором букв |
| Методы | SpellChecker() – конструктор |
| ~SpellChecker() – деструктор |
| void loadDictionary(istream&) – загрузка словаря из входного потока |
| set<string> options\_T(string) – поиск вариантов замены через транспозицию |
| set<string> options\_D(string) – поиск вариантов замены через удаление символа |
| set<string> options\_I(string) – поиск замены через вставку символа |
| bool spellCheck(string) – проверка на предмет наличия орфографической ошибки |

Приведем исходный код класса SpellChecker:

#pragma once  
#ifndef **CHW\_2\_SPELLCHECKER\_H**#define **CHW\_2\_SPELLCHECKER\_H**#include **<string>**#include **<set>**#include **<unordered\_map>**#include **<unordered\_set>  
using namespace** std;  
  
**class** SpellChecker {  
**public**:  
 SpellChecker();  
 **virtual** ~SpellChecker();  
 **void** loadDictionary(istream& input);  
 set<string> options\_T(string s);  
 set<string> options\_D(string s);  
 set<string> options\_I(string s);  
 **bool** spellCheck(string s);  
**private**:  
 set<string> \*dict\_;  
 unordered\_map<**unsigned long**, unordered\_map<string, unordered\_set<string>>> ltg\_;  
 **int** \_transposeDistance(string a, string b);  
 **void** \_initLTG();  
};  
  
SpellChecker::SpellChecker(){  
 dict\_ = **new** set<string>();  
}  
  
SpellChecker::~SpellChecker() {  
 **delete** dict\_;  
}  
**void** SpellChecker::loadDictionary(istream &input) {  
 string temp\_;  
 **while**(getline(input, temp\_)){  
 dict\_->insert(temp\_);  
 }  
 \_initLTG();  
}  
  
**bool** SpellChecker::spellCheck(string s) {  
 **return** dict\_->find(s) != dict\_->end();  
}  
  
set<string> SpellChecker::options\_T(string s) {  
 sort(s.begin(), s.end());  
 **return** set<string>(ltg\_[s.length()][s].begin(), ltg\_[s.length()][s].end());  
}  
  
set<string> SpellChecker::options\_D(string s) {  
 **return** set<string>();  
}  
  
set<string> SpellChecker::options\_I(string s) {  
 **return** set<string>();  
}  
  
**int** SpellChecker::\_transposeDistance(string a, string b) {  
 **if**(a.length() != b.length())  
 **return** -1;  
 **int** tc\_ = 0;  
 **int** i = 1;  
 **while**(i < a.length()){  
 **if**((a[i] != b[i]) && (a[i] == b[i - 1]) && (b[i] == a[i - 1])){  
 tc\_++;  
 }  
 i++;  
 }  
 **return** tc\_;  
}  
  
**void** SpellChecker::\_initLTG(){  
 **for**(**auto** w\_ = dict\_->begin(); w\_ != dict\_->end(); w\_++){  
 **if**(ltg\_.find(w\_->length()) == ltg\_.end()){  
 ltg\_[w\_->length()] = unordered\_map<string, unordered\_set<string>>();  
 }  
 string temp\_(\*w\_);  
 sort(temp\_.begin(), temp\_.end());  
 **if**(ltg\_[w\_->length()].find(temp\_) == ltg\_[w\_->length()].end()){  
 ltg\_[w\_->length()][temp\_] = unordered\_set<string>();  
 }  
 ltg\_[w\_->length()][temp\_].insert(\*w\_);  
 }  
}  
  
#endif *//CHW\_2\_SPELLCHECKER\_H*

## 3.4 Дополнительная функциональность

Дополнительная функциональность реализации отсутствует.

## 3.5 Заключение

Я, Ярных Роман Вячеславович, подтверждаю, что предоставляемое на проверку Домашнее задание по дисциплине «Алгоритмы и структуры данных» (модуль 4, 2015-2016 учебного года) выполнено мною самостоятельно. Данный отчет не содержит объяснений (пояснительного текста), фрагментов псевдокода и/или кода программ, взятых из других работ (электронных или печатных), без указания в тексте явных ссылок на эти источники.

Я проинформирован и полностью осознаю, что при обнаружении в отчёте факта списывания, подлога, плагиата и пр., моей оценкой за домашнее задание будет 0 (нуль) и ко мне будут применены меры, изложенные в документе *«Порядок применения дисциплинарных взысканий при нарушениях академических норм в написании письменных учебных работ в НИУ-ВШЭ»* (<http://www.hse.ru/studyspravka/loc>).