МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МОЭВМ

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Построение и анализ алгоритмов»

Тема: «Алгоритм Ахо-Корасик»

Студент гр. 7304 _____Кошманов Н.А.

Преподаватель _____Филатов А.Ю.

Санкт-Петербург
2019

Цель работы

1) Разработать программу, решающую задачу точного поиска набора образцов с помощью алгоритма Ахо-Корасик.
2) Используя реализацию точного множественного поиска, решить задачу точного поиска для одного образца с *джокером*.

Пояснение задания

1)

Вход:

Первая строка содержит текст (T,$1 \leq |T| \leq 100000$).

Вторая - число n ($1 \leq n \leq 3000$), каждая следующая из n строк содержит шаблон из набора $P=\{p1,\ldots,pn\}$ $1 \leq |pi| \leq 75$

Все строки содержат символы из алфавита {A,C,G,T,N}

Выход:

Все вхождения образцов из P в T.

Каждое вхождение образца в текст представить в виде двух чисел - i p

Где i - позиция в тексте (нумерация начинается с 1), с которой начинается вхождение образца с номером p

(нумерация образцов начинается с 1).

Строки выхода должны быть отсортированы по возрастанию, сначала номера позиции, затем номера шаблона.

2)

Вход:

Текст (T,$1 \leq |T| \leq 100000$)

Шаблон (P,$1 \leq |P| \leq 40$)

Символ джокера

Выход:

Строки с номерами позиций вхождений шаблона (каждая строка содержит только один номер).

Номера должны выводиться в порядке возрастания.

Описание алгоритма Ахо-Корасик

Суть алгоритма заключена в использование структуры данных — бора и построения по нему конечного детерминированного автомата. Строится бор последовательным добавлением исходных строк. Изначально есть 1 вершина, корень - пустая строка. Добавление строки происходит так: начиная в корне, двигаемся по дереву, выбирая каждый раз ребро, соответствующее очередной букве строки. Если такого ребра нет, то мы создаем его вместе с вершиной. Так как процесс добавления строки может остановиться во внутренней вершине, то для каждой строки будем дополнительно хранить признак того является она строкой из условия или нет. Далее, строим конечный детерминированный автомат. Состояние автомата — это какая-то вершина бора. Переход из состояний осуществляется по 2 параметрам — текущей вершине v и символу ch по которому нам надо сдвинуться из этой вершины. Назовем суффиксной ссылкой вершины v указатель на вершину u, такую что строка u — наибольший собственный суффикс строки v, или, если такой вершины нет в боре, то указатель на корень. В частности, ссылка из корня ведет в него же. Если из текущей вершины есть ребро с символом v, то пройдем по нему, в обратном случае пройдем по суффиксной ссылке и запустимся рекурсивно от новой вершины. Алгоритм завершится, когда мы дойдем до конца строки.

Описание алгоритма точного поиска образца с джокером

Алгоритм точного поиска образца с джокером является аналогом алгоритма Ахо-Корасик и использует его в своей работе. Однако здесь строка-шаблон разделяется по символу-джокеру на строки, каждая из которых добавляется в бор.

Описание основных функций

int GetSuffFLink(int v) - функция находит суффиксную ссылку, если она не была найдена до этого, и возвращет ее.

int GetAutoMove(int v, int ch) - функция находит и возвращает элемент в боре, в который следует переместиться из вершины *v* по символу *ch*.

void FindAllPos(const string& s) и **void Check(int v, int i)** - функции находят позиции вхождения заданных шаблонов в строку.

Тестирование

1. Алгоритм Ахо-Корасик

   1)

```
E:\Study\4 sem\PIAA\Labs\Lab5>AK1
CCAGTCC
2
CC
TC
1 1
5 2
6 1
```

   2)

```
E:\Study\4 sem\PIAA\Labs\Lab5>AK1
GGGTNNAAC
3
AC
NA
GG
1 3
2 3
6 2
8 1
```

2. Алгоритма точного поиска образца с джокером

   1)

```
E:\Study\4 sem\PIAA\Labs\Lab5>AK2
ACACGTNNNA
AC&T&
&
3
```

   2)

```
E:\Study\4 sem\PIAA\Labs\Lab5>AK2
ACACCACA
A*
*
1
3
6
```

Вывод

В ходе выполнения лабораторной работы был изучен и реализован на языке программирования C++ алгоритм Ахо-Корасик для поиска множества подстрок в строке. Данный алгоритм выполняет точный поиск набора образцов в строке. В нем используются такие понятия, как бор, конечный детерминированный автомат, суффиксальные ссылки. Бор — это дерево, в котором каждая вершина обозначает какую-то строку. На ребрах между вершинами написана 1 буква, таким образом, добираясь по ребрам из корня в какую-нибудь вершину, мы получим строку, соответствующую этой вершине. Детерминированным конечным автоматом называется такой автомат, в котором нет дуг с меткое e, и из любого состояния по любому символу возможен переход не более, чем в одно состояние. Суффиксная ссылка вершины v- указатель на вершину u, такую, что строка u — наибольший собственный суффикс строки v, или, если такой вершины нет в боре, то указатель на корень.