

Лабораторная работа № 5 по курсу дискретного анализа: Суффиксные деревья

Выполнил студент группы 08-307 МАИ *Путилин Дмитрий*.

Условие

Найти в заранее известном тексте поступающие на вход образцы с использованием суффиксного массива.

Метод решения

Суффиксным массивом строки $s[1..n]$ называется массив suf целых чисел от 1 до n , такой, что суффикс $s[\text{suf}[i]..n]$ — i -й в лексикографическом порядке среди всех непустых суффиксов строки s .

Для построения используется алгоритм Цифровой сортировки со сложностью $O(n)$. Его идея заключается в том, что происходит сортировка строк длины 2^i ; ($i = 0, \dots, \log n$).

Для начала происходит нулевой шаг: сортируются буквы строки, затем каждой букве присваивается определенный класс. Потом происходит k -следующих шагов пока $2^k \leq n$.

Соответственно на каждом i шаге происходит работа со строками длины 2^i , их аналогичным образом сортируют по первым $2^i - 1$ буквам, так как последующие уже отсортированы, дальше присваивают им новые классы эквивалентности. Алгоритм заканчивается, когда в массиве классов классы строкам не будут от 1 до n .

После происходит работа с полученными на вход паттернами бинарным поиском. Для каждого паттерна находятся соответствующие индексы в строке и добавляются в вектор. В момент вывода выполняется проверка на пустой вектор, если она выполнена, то выводится результат.

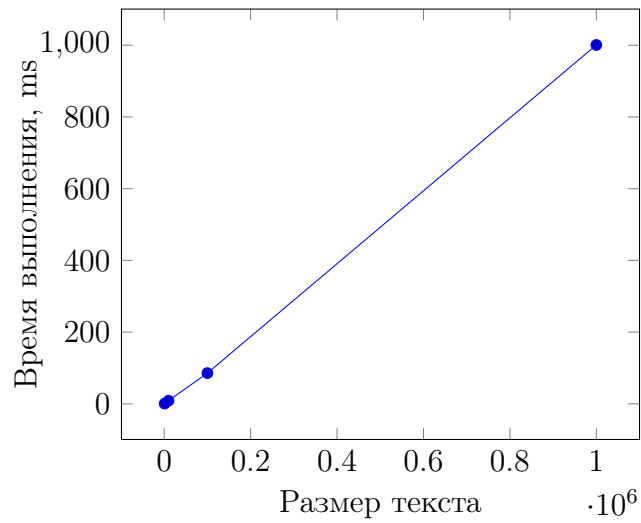
Описание программы

Функция suf принимает на вход строку, а возвращает vector индексов. В main происходит считывание строк, после происходит поиск их вхождений с помощью суффиксного массива. После этого происходит вывод результата.

Дневник отладки

1. Первая посылка - WA, так как не проверял на пустой вектор индексов
2. Вторая посылка - TL, отключая синхронизацию C и C++ ввода/вывода
3. Третья посылка - ОК

Тест производительности



Заметим, что график линейен, следовательно алгоритмическая сложность вычисления суффиксного массива $O(n)$.

Выводы

Был написан алгоритм с вычислением суффиксного массива методом Цифровой сортировки, который использовался для нахождения индексов вхождений паттернов в текст.