

12 Строки: Z, π, Бор

12.1 Практика

Хешами пока пользоваться нельзя.

Строки обычно сравнивают *лексикографически* (например, так слова упорядочены в словаре). Строка s лексикографически меньше строки t , если:

- либо s является собственным (то есть не совпадающим со всей строкой) префиксом строки t ;
- либо для некоторого числа $k \geq 0$ выполнено $s[0..k) == t[0..k)$ и $s[k] < t[k]$.

Пусть дана строка $s[0..n)$.

i -м суффиксом s называют $s[i..n)$, i -м префиксом s называют $s[0..i]$.

Выпишем все n суффиксов строки s и отсортируем их в лексикографическом порядке. *Суффиксный массив* — это номера суффиксов в этом отсортированном массиве. Например, для $s = \text{"acaba"}$ отсортированный массив суффиксов будет $[\text{"a"}, \text{"aba"}, \text{"acaba"}, \text{"ba"}, \text{"caba"}]$, а суффиксный массив будет $[4, 2, 0, 3, 1]$.

1. Периоды строки

Период строки s — число $T > 0$ такое, что для любого индекса $0 \leq i < (|s| - T)$ выполнено $s[i] = s[i + T]$.

Найти все периоды строки за $\mathcal{O}(n)$. Решить с помощью Z-функции и с помощью префикс-функции.

2. Позиция в суффиксном массиве

Найдите позицию строки (то есть нулевого суффикса) в её суффиксном массиве. $\mathcal{O}(n)$.

3. Поиск словарных слов

Даны словарь из слов s_1, \dots, s_k и текст t . Суммарная длина слов из словаря — L , максимальная — l_{max} . Найти все вхождения всех словарных слов в t за время:

- $\mathcal{O}(L + k|t|)$.
- $\mathcal{O}(|\Sigma|L + l_{max}|t|)$.
- (*) $\mathcal{O}(L + l_{max}|t|)$.

4. Разбиение текста

Дан словарь слов суммарной длины L и текст T . Длины слов в словаре не более l . Представьте текст в виде конкатенации минимального количества словарных слов за время $\mathcal{O}(L + l|T|)$. SK: время работы не пишем, чтобы думать было проще Слова можно использовать более одного раза.

5. Строки в дереве

Дано подвешенное дерево, на ребрах которого написаны непустые строки суммарной длины n , и дан образец p (все над алфавитом Σ). Найдите все соответствующие вхождениям p отрезки вертикальных путей за $\mathcal{O}(n + |p| \cdot |\Sigma|)$.

6. Одна ошибка и ты ошибся

Научиться искать образец в строке, если допустимо различие в один символ между образцом и найденной подстрокой. $\mathcal{O}(n + m)$.

7. Различные подстроки

Найдите число различных подстрок строки за $\mathcal{O}(n^2)$.

8. (*) Восстановление строки

За $\mathcal{O}(n)$ восстановить строку, если дана её

- Z-функция;
- префикс-функция.

Если таких строк несколько, восстановить любую.

12.2 Домашнее задание

Хешами пока пользоваться нельзя.

1. (1) Закрепление

Дана строка s . Её циклические сдвиги – это строки вида $s[i..n) + s[0..i)$ для $0 \leq i < n$ (сама строка s тоже является циклическим сдвигом s).

По строке s определите, верно ли, что она (лексикографически) не больше любого своего циклического сдвига. Время $\mathcal{O}(|s|)$.

2. (1) Префиксы префиксов

Дана строка s длины n . Для каждого i от 1 до n найти количество непустых префиксов строки $s[0..i)$, равных суффиксу той же строки $s[0..i)$. Время $\mathcal{O}(n)$.

3. (1) Т9

Есть словарь. У каждого слова есть «частота использования» изначально равная 1.

Придумайте структуру данных, которая поможет обрабатывать два запроса

- Пользователь ввёл словарное слово w , увеличить его частоту на 1.
- Пользователь ввёл строку s и хочет увидеть 5 самых частых слов, начинающихся на s .

Можно считать, что алфавит имеет константный размер, на каждый запрос отвечать за линейное время от длины запроса и длины ответа на запрос.

4. (1) Различные подстроки

Найдите число различных подстрок строки за $\mathcal{O}(n^2)$.

Подсказка: любая подстрока – это префикс какого-то суффикса.

5. (+1) (*) Дана матрица A размера $a \times b$ и матрица B размера $c \times d$, состоящие из букв некоторого алфавита Σ . Необходимо найти все вхождения матрицы B в матрицу A в качестве подматрицы. При этом $c \leq a$, $d \leq b$. Сложность по времени и памяти – $\mathcal{O}(a \cdot b)$.

6. (+1) (*) Породите две строки

Даны строки s и t . Найдите кратчайшую строку p такую, что строка p^∞ (строка p , повторённая бесконечно много раз) содержит s и t как подстроки. Если таких p несколько, найдите любую. Время $\mathcal{O}(|s| + |t|)$.