Поиск подстрок

Гусев Илья

Московский физико-технический институт

Москва, 2017

Содержание

Префикс-функция

Префикс-функция - массив чисел π, где $\pi[i]$ - такая наибольшая длина k наибольшего суффиксаs $s[i-k+1 \dots i]$ подстроки $s[0 \dots i]$, совпадающего с её префиксом $s[0 \dots k]$, но не совпадающего со всей строкой s.

$$\pi[s,i] = \max_{k=0 \dots i} \{k : (s[0 \dots k] = s[i-k+1 \dots i])\}$$

Пример: abcabcd

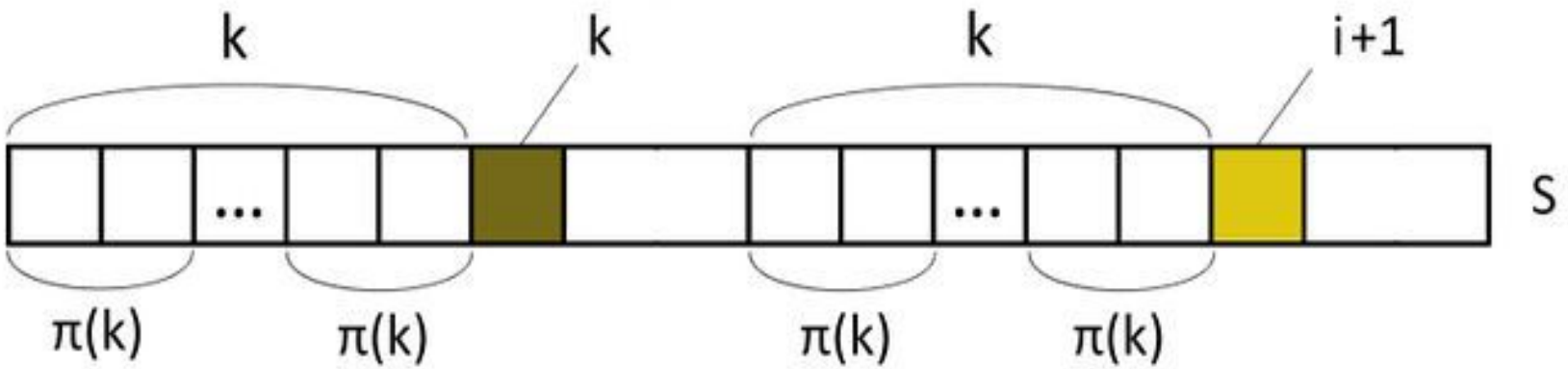
s	a	b	c	a	b	c	d
π	0	0	0	1	2	3	0

Вычисление префикс-функции

Утверждения:

1. $\forall i \to \pi[i+1] \leqslant \pi[i] + 1$
2. $\forall i : s[i+1] = s[\pi[i]] \to \pi[i+1] = \pi[i] + 1$
3. $\forall i : s[i+1] \neq s[\pi[i]] \to \pi[i+1] \leq \pi[i]$

На картинке $k = \pi[i]$

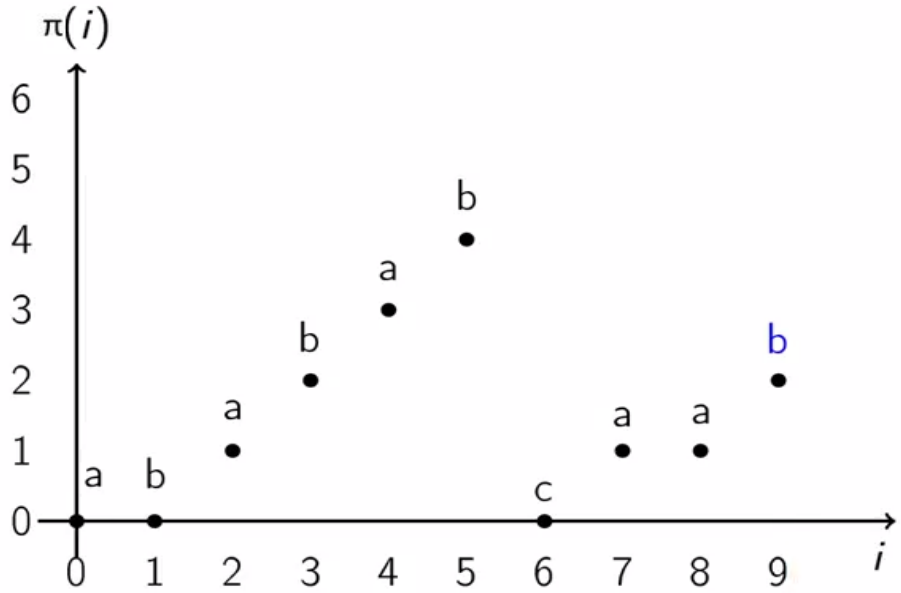


Для третьего случая итерируем $k = \pi[k]$, пока следующий символ не совпадёт.

Сложность вычисления префикс-функции

Утверждения:

1. $\forall i \to \pi[i] < n - 1$
2. π по 1 разу за шаг вычисления, если выполняются условия 2 случая.
3. π увеличивается не более, чем $n - 1$ раз за весь алгоритм.
4. $\forall i \to \pi[i] \geq 0$
5. π уменьшается не более, чем $n - 1$ раз.
6. Всего не более $2\dot{n}$ шагов $\Longrightarrow O(n)$



Задача

Найти **лексикографически-минимальную** строку, построенную по префикс-функции, в алфавите a-z. Примеры:

π	0	0	0	1	2	3	0
s	a	b	c	a	b	c	b

π	0	0	1	2	3	4	5	0
s	a	b	a	b	a	b	a	c

Решение

1. $\pi[i] \neq 0 \implies s[i] = s[\pi[i] - 1]$
2. $\pi[i] = 0 \implies s[i] = \max\{s[\pi[i-1]] + 1, s[\pi[\pi[i-1] - 1]] + 1, \ldots\}$

Первое очевидно и следует напрямую из определения префикс-функции.
Второе опирается на несколько фактов:

1. Нельзя допустить, чтобы новый символ сделал суффикс, совпадающий с **каким-либо** префиксом по префиксам префиксов).
2. $+1$ - всегда достаточно из-за того, что префикс всегда минимален. Делая $+1$ (например, из a в b) - гарантированно получаем 0 в префикс-функции.
3. Из всех возможных вариантов продолжения выбираем минимальный.

Z-функция

Z-функция - массив чисел z, где $z[i]$ - длина наибольшего префикса строки s, который равен префиксу i-ого суффикса $s[i \dots n-1]$.

$$z[s,i] = \max_{k=0 \dots n-1-i} \{k : (s[0 \dots k] = s[i \dots i+k])\}$$

Примеры:

s	a	a	a	a	a
z	0	4	3	2	1

s	a	a	a	b	a	a	b
z	0	2	1	0	2	1	0

s	a	b	a	c	a	b	a
z	0	0	1	0	3	0	1

Задача

Найти **лексикографически-минимальную** строку, построенную по z-функции, в алфавите a-z.

z	5	3	2	1	0
s	a	a	a	a	b

Решение

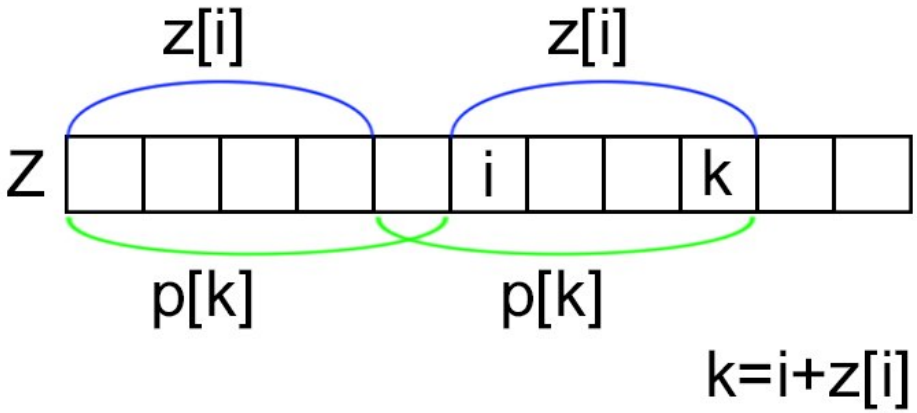
Один из возможных подходов:

1. z-функция $\Longrightarrow$ префикс-функция
2. Решаем предыдущую задачу

Z-функция $\Longrightarrow$ префикс-функция

Утверждения:

1. $\forall i \in [0, n), \forall j \in [0, z[i]) \to \pi[i+j] \geq j+1$
2. $\forall i, \forall i', \forall j \in [0, z[i]), \forall j' \in [0, z[i']) : i < i', i+j = i'+j' \to \pi[i+j] = \pi[i'+j'] \geq j+1 > j'+1$ - на следующих итерациях значение префикс функции не увеличится
3. Если наталкиваемся на уже заданное значение π, переходим к $i+1$



Сложность: $O(n)$, так как каждый элемент меняется ровно один раз и останавливаемся на каждом не более 1 раза.

Алгоритм Кнута—Морриса—Пратта

Дано: есть шаблон T, строка S, $len(T) < len(S)$.
Найти: все вхождения T в S.
Решение: $concat(T, \#, S)$, считаем префикс функцию. Где $\pi[i] = len(T)$, там и есть конец вхождения.
Сложность: $O(len(T) + len(S))$
Нюанс: при реализации запрещается хранить все значения префикс-функции! Ограничьтесь только нужными.

Полезные ссылки I

- Видео про префикс-функцию на Курсере
  https://ru.coursera.org/learn/algorithms-on-strings/lecture/5lDsK/computing-prefix-function
- E-maxx: префикс-функция
  http://e-maxx.ru/algo/prefix_function
- Викиконспекты: префикс-функция
  https://neerc.ifmo.ru/wiki/index.php?title=Префикс-функция
- E-maxx: z-функция
  http://e-maxx.ru/algo/z_function