Хеширование

Минский ШАД. Осень

21 февраля 2015 г.

1 Обозначения

В данной домашней работе будет много задач на строки. Введём следующие обозначения:

- $|s|$ — длина строки s
- $s[i]$ — i-й символ строки s
- $s[i \dots j]$ — подстрока строки s, которая начинается в индексе i и заканчивается в индексе j
- $\overline{s}$ — «перевёрнутая» строка s
- $\mathrm{ord}(c)$ — произвольная инъективная функция из алфавита строки в целые числа. Тут будем считать, что символы пронумерованы по алфафиту, т.е. $\mathrm{ord}(a) = 1, \mathrm{ord}(b) = 2, \dots$

Например, если $s =$ «abacaba», то:

- $|s| = 7$
- $s[3] =$ 'c'
- $s[1 \dots 3] =$ «bac»
- $\overline{s[1 \dots 3]} =$ «cab»

2 Тематические задачи

1. Предложить, как решать с помощью полиномиального хеширования следующие задачи (везде вам дана строка s, причём $|s| = n$):
   (a) [½ балла] По данным парам (l_1, r_1) и (l_2, r_2) отвечать на запрос, правда ли, что равны две строки $s[l_1 \dots r_1]$ и $s[l_2 \dots r_2]$ за $\mathcal{O}(1)$. Разрешается делать препроцесс за $\mathcal{O}(n)$
   (b) [½ балла] По данным парам (l_1, r_1) и (l_2, r_2) отвечать на запрос, правда ли, что равны две строки $s[l_1 \dots r_1]$ и $\overline{s[l_2 \dots r_2]}$ за $\mathcal{O}(1)$. Разрешается делать препроцесс за $\mathcal{O}(n)$
   (c) [½ балла] По данной паре (l, r) отвечать на запрос, правда ли, что строка $s[l \dots r]$ является палиндромом за $\mathcal{O}(1)$. Разрешается делать препроцесс за $\mathcal{O}(n)$
   (d) [½ балла] Найти по данным (i, j) найти длину наибольшего общего префикса двух строк $s[i \dots |s|]$ и $s[j \dots |s|]$ за $\mathcal{O}(\log n)$. Разрешается препроцесс за $\mathcal{O}(n)$
   (e) [½ балла] Вычислить z-функцию строки за $\mathcal{O}(n \log n)$ (т.е. найти z_i для всех $i = \overline{1 \dots |s|}$). z_i — длина наидлиннейшей подстроки, которая начинается в символе с индексом i и совпадает с префиксом строки.
   (f) [½ балла] Для пары (i, j) выяснить, какой суффикс лексикографически меньше: который начинается в i или который начинается в j. Время работы $\mathcal{O}(\log n)$. Препроцесс за $\mathcal{O}(n)$

(g) [½ балла] Построить суффиксный массив для строки s за время $\mathcal{O}(n \log^2 n)$. i-суффиксом (suf_i) назовём подстроку $s[i \dots |s|]$. Суффиксный массив a_i — перестановка первых n чисел, такая, что $\mathrm{suf}_{a_i} < \mathrm{suf}_{a_{i+1}}$ для любого $i = \overline{1 \dots |s|-1}$. Сравнение проводится лексикографически.

2. Предложить функцию для хеширования мультимножеств. А именно, по мультимножеству A и числу m ваша функция $h(A, m)$ должна выдавать число в диапазоне $0 \dots .. 2^m - 1$, такое что (можно считать, что у вас есть хеш-функция для любого возможного элемента мультимножества, которая вычисляется за $\mathcal{O}(1)$):
   (a) [½ балла] $h(A, m) = h(B, m)$, если $A = B$
   (b) [½ балла] Функция должна быть легко обновляемая (т.е. при добавлении элемента в мультимножество должно быть можно пересчитать значение $h(A \cup \{x\}, m)$ за $\overline{o}(|A|^\varepsilon)$, для любого $\varepsilon > 0$)
   (c) [½ балла] Функция должна быть суръективна (можно считать, что функция хеширования элемента суръективна)
   (d) [3 балла] Функция должна быть стойкой. С целью упрощения будем считать, что функция стойкая, если выполняется хотя бы одно из двух:
   - Рассмотрим конечное множество элементов B и будем считать, что все элементы мультимножества лежат в B. Зададимся числом n и рассмотрим множество мультимножеств $S_n = \{A : |A| \leqslant n\}$. Функцию будем называть стойкой, если $\forall m$ и для любого k ($0 \leqslant k < 2^m$), $P\{h(A, m) = k | A \in S_n\} \to \frac{1}{2^m}$, при $n \to \infty$
   - Функцию будет называть стойкой, если для достаточного большого m и $|A| \nexists$ такая константа k, что $\forall |A| \ \exists C, D$, такие что $|C| \leqslant k$, $|D| \leqslant k$ и $h(A, m) = h((A \cup C) \setminus D, m)$

3 Задачи на повторение

3. [½ балла] Дан отсортированный массив различных целых чисел. Надо определить, существует ли такой индекс i, что $a_i = i$. Сложность алгоритма должна быть $\mathcal{O}(\log n)$, где n — длина массива.

4. [½ балла] Пусть мы имеем два положительные неубывающие функции $f(x)$ и $g(x)$, причём $f(n) = \mathcal{O}(g(n))$. Правда, что $2^{f(n)} = \mathcal{O}(2^{g(n)})$? Если это может как выполняться, так и не выполняться, привидете примеры обоих случаев. Иначе докажите утверждение.

5. [½ балла] Пусть у нас есть k отсортированных последовательностей из n чисел каждая. Предлагается такой алгоритм слияния их в одну: сначала сольём две первых последовательности, затем результат с третьей, и так далее. Какова сложность полученного алгоритма? Считаем, что слияние двух массивов происходит за их суммарную длину. Какова сложность полученного алгоритма.

4 Практические задачи

Ссылка на контест: `https://contest.yandex.ru/contest/1080/problems/`

6. [1 балл] **Задача A**. Дана строка S. Необходимо найти самую длинную подстроку, которая встречается в S хотя бы два раза. Вхождения могут перекрываться. Ожидаемая сложность $\mathcal{O}(|S| \log |S|)$.

7. [1 балл] **Задача B**. Реализуйте решение задачи про суффиксный массив через хеши.

Задание	1	2	3	4	5	6	7	Сумма
Баллы	3½	4½	½	½	½	1	1	11½