

字符串哈希原理、代码、题目详解

前置知识

讲解003 - 二进制基础

讲解041 - 同余原理

本节课讲述：

哈希函数特征

字符串哈希：如何得到整个字符串的哈希值

字符串哈希：如何快速得到字符串中任意子串的哈希值

字符串哈希实战题目

哈希函数非常重要，下期视频会讲述哈希函数的更多内容

字符串哈希原理、代码、题目详解

哈希函数的用处

把复杂样本变成数字，以后复杂样本之间的对比，就变成了数字之间的对比

哈希函数的基本性质

- 1) 输入参数的可能性是无限的，输出的值范围相对有限
- 2) 输入同样的样本一定得到同样的输出值，也就是哈希函数没有任何随机机制
- 3) 输入不同的样本也可能得到同样的输出值，此时叫哈希碰撞
- 4) 输入大量不同的样本，得到的大量输出值，会几乎均匀的分布在整個输出域上，课上重点图解

哈希函数的种类很多，但都符合上述性质

性质4是最重要的，哈希碰撞理论上无法避免，但是好的哈希函数会让碰撞几率变的很低

可以把性质4直观理解为：

不管有没有规律，也不管多么相似，总之一堆不同的输入，得到的输出结果从分布上看，熵最大！

字符串哈希原理、代码、题目详解

哈希函数的算法有很多，字符串哈希是最常用的，也是唯一需要掌握代码实现的哈希函数

字符串哈希：如何得到整个字符串的哈希值

- 1) 理解long类型自然溢出，计算加、减、乘时，自然溢出后的状态等同于对2的64次方取模的值状态
- 2) 字符串转化成base进制的数字并让其自然溢出
- 3) base可以选择一些质数比如：433、499、599、1000000007
也可以选择已经被证实了很好用的值：31、131、1313、13131、131313等
建议选择质数，不要选经典值，因为会被出题人刻意构造碰撞
- 4) 转化时让每一位的值从1开始，不从0开始，这样就得到了一个long类型的数字代表字符串
- 5) 利用数字的比较去替代字符串的比较，可以大大减少复杂度

字符串哈希从理论上说会有碰撞导致出错，但现实中的算法考察样本量太少了，出错概率非常低
即便是出错了，也可以更换进制数base，再去赌，一定能赌赢
没错！是玄学！但是好用！堪称赌狗的胜利

字符串哈希原理、代码、题目详解

字符串哈希：如何快速得到字符串中任意子串的哈希值

1) 选择一个质数做进制数, base

2) 得到base的各种次方, 在自然溢出下的结果, 用pow数组记录

3) 得到每个位置的hash[i], $\text{hash}[i] = \text{hash}[i-1] * \text{base} + s[i] - 'a' + 1$

4) 子串s[l...r]的哈希值 = $\text{hash}[r] - \text{hash}[l-1] * \text{base}^{\text{的}(r-l+1)\text{次方}}$, 课上会重点讲解

字符串中子串对比变成哈希值对比非常好用的！大量节省时间

很多较难的算法都可以被字符串哈希替代，都是因为子串对比的代价变为 $O(1)$

字符串哈希易于理解且使用灵活，因为非常方便的子串对比，很多难题变得非常好想

字符串哈希原理、代码、题目详解

题目1

统计有多少个不同的字符串

测试链接：<https://www.luogu.com.cn/problem/P3370>

字符串哈希原理、代码、题目详解

题目2

独特子串的数量

给你一个由数字组成的字符串s，返回s中独特子字符串数量

独特子串定义：每一个数字出现的频率都相同

测试链接：<https://leetcode.cn/problems/unique-substrings-with-equal-digit-frequency/>

字符串哈希原理、代码、题目详解

题目3

字符串哈希得到子串哈希

利用字符串哈希的便利性替代KMP算法

测试链接：<https://leetcode.cn/problems/find-the-index-of-the-first-occurrence-in-a-string/>

字符串哈希也能替代Manacher算法

不过时间复杂度没有Manacher算法解决回文类的问题好

Manacher算法生成回文半径数组，时间复杂度 $O(n)$

字符串哈希替代Manacher算法生成回文半径数组，时间复杂度 $O(n * \log n)$

这里不再详述

字符串哈希原理、代码、题目详解

题目4

重复叠加字符串匹配

给定两个字符串a和b，寻找重复叠加字符串a的最小次数，使得字符串b成为叠加后的字符串a的子串
如果不存在则返回 -1。

字符串"abc"重复叠加0次是""

重复叠加1次是"abc"

重复叠加2次是"abccabc"

测试链接：<https://leetcode.cn/problems/repeated-string-match/>

字符串哈希原理、代码、题目详解

题目5

串联所有单词的子串

给定一个字符串s和一个字符串数组words

words中所有字符串长度相同

s中的串联子串是指一个包含 words中所有字符串以任意顺序排列连接起来的子串

例如words = { "ab", "cd", "ef" }

那么"abcdef"、"abefcd"、"cdabef"、"cdefab"、"efabcd"、"efcdab"都是串联子串。

"acdbef"不是串联子串，因为他不是任何words排列的连接

返回所有串联子串在s中的开始索引

你可以以任意顺序返回答案

测试链接：<https://leetcode.cn/problems/substring-with-concatenation-of-all-words/>

字符串哈希原理、代码、题目详解

题目6

根据匹配定义求匹配子串的数量

给定长为 n 的源串 s ，以及长度为 m 的模式串 p ，还有一个正数 k

s' 与 s 匹配的定义为， s' 与 s 长度相同，且最多有 k 个位置字符不同

要求查找源串中有多少子串与模式串匹配

测试链接：<https://www.luogu.com.cn/problem/P3763>