字符串哈希

宁华

例 Acwing138 兔子与兔子

- 很久很久以前，森林里住着一群兔子。
- 有一天，兔子们想要研究自己的 DNA 序列。
- 我们首先选取一个好长好长的 DNA 序列（小兔子是外星生物，DNA 序列可能包含 26 个小写英文字母）。
- 然后我们每次选择两个区间，询问如果用两个区间里的 DNA 序列分别生产出来两只兔子，这两个兔子是否一模一样。
- 注意两个兔子一模一样只可能是他们的 DNA 序列一模一样。

例 Acwing138 兔子与兔子

- 输入格式
- 第一行输入一个 DNA 字符串 S。
- 第二行一个数字 m，表示 m 次询问。
- 接下来 m 行，每行四个数字 l1,r1,l2,r2，分别表示此次询问的两个区间，注意字符串的位置从 1 开始编号。

- 输出格式
- 对于每次询问，输出一行表示结果。
- 如果两只兔子完全相同输出 Yes，否则输出 No（注意大小写）。

- 数据范围
- $1 \le length(S), m \le 1000000$

例 Acwing138 兔子与兔子

- 输入样例:
- aabbaabb
- 3
- 1 3 5 7
- 1 3 6 8
- 1 2 1 2

- 输出样例:
- Yes
- No
- Yes

方法1：朴素算法

方法2：字符串哈希

- 取一固定值 P，把字符串看做 P 进制数，并分配一个大于 0 的数值，代表每种字符。
- 假如只有26个英文小写字母，则可以把字符串看作是 27 进制数（不使用数码0 （思考：为什么不使用数码0？））

- a —— 1
- b —— 2
- c —— 3
- ……
- z ——26

- S = “abc”
- T = “xyz”

- S 映射为 27 进制数：$(123)_{27}$
- T 映射为 27 进制数：$(24\ 25\ 26)_{27}$

- 比较 S 与 T 是否相等 转化为 判断 两个 27 进制数是否相等

- 一般来说，我们选择的 P 值通常远大于字符的种数，比如 P = 131 进制。

- P=131
- “abcdefghijk” ===> 1*131^10+2*131^9+...

- 随之带来的问题是?

问题：

- 映射得到的数值太大，标准数据类型存不下。

问题解决方法：

- 取一固定值 M，求出该P进制数对M的余数，作为该字符串的Hash值。
- 通常取 unsigned long long M = 2^64
- 溢出 相当于 自动对 2^64 取模，避免低效的取模(mod)运算

新的问题：

- 是否会出现两个不同字符串映射得到的数值相同？——冲突

问题解决方法：

- 一般来说，取 P=131 或 P=13331 ， 此时Hash值产生冲突的概率极低。

- 或者多取一些恰当的P和M，多进行几组Hash运算，当结果都相同时才认为两个字符串相同。

总结

- P=131 或 P=13331
- M=2^64

方法2：字符串哈希

- 10进制：
- 求 123456789 左起第 3 位到第 7 位这一段的值

- 解：
- 1234567 - 12*10^(7-3+1)
- = 1234567 - 1200000
- = 34567

- 求字符串 “abcdefghijklmn” 左起第 3 位到第 7 位这一段字符串的Hash值

例 Acwing138 兔子与兔子

- 完成代码并提交