
读作字符串
写作数据结构

温馨提醒

- 本专题仅适合下列人群（之一）：

- 会使用AC自动机的同学
- 会使用后缀自动机的同学
- 能熟练运用数据结构的同学
- 打算浪费一个晚上的同学

- ~~现在走还来得及~~

温馨提醒

- 许多字符串题搭配了一da定liang的数据结构与其相关的处理技巧，复杂度分析等。

- 包括但不限于:
- 线段树，平衡树，LCT（及其可持久化）

- 对于这些实现细节将不会提及。

Content

- 课件内容：字符串姿势普及
- 中间会夹杂一些题目

最简单基础的算法

- kmp
- 马拉车
- O(n) 求以某个位置为中心的回文串长度

- 本质不同的回文子串总数是O(n)
- 因为一个结尾最多只有一个。

怎么能忘了hash

- O(1)判断字符串是否相同
- 据说unsigned long long的哈希无论取什么位权都可以被**刻意**卡掉

- 998244353与1e9+7这两个模数
- 取5000长度字符串的所有子串就有冲突了

加大模数或者多模数

- 换更大的模数
- 但是需要龟速乘

- 搞多几个模数

Trie

- 字典树
- 位运算有时候也用这个搞

- 对于相当有限的字符集大小（比如2），需要注意的是，空间复杂度并不是O(总长度)，而是串数 + 总长 – 串数 * log 串数，其中的log以sigma为底。
- 原因是前log层会重复。这个姿势有时可以卡一下空间。

- 可以压缩，把度数=2的点缩起来，总点数就变成了O(串数)，但是也会丧失一些功能。

AC自动机

- Kmp是特殊的ac自动机
- 支持多串匹配
- 由于一个位置只对应一个串，异常好理解。
- Fail是当前串在trie上存在的的最长后缀。
- 沿着fail走是以当前位置为结尾的所有匹配串

- Ac自动机上跑dp是常见的姿势

简单题 3s

- **题目大意**

  一开始给出 N 个字符串，分别为 s_i。

  并有一个字符串集合 T，T 一开始是空的。

  共有 Q 个操作，有两种操作：

  1. 往 T 中加入一个字符串 p。

  2. 选择一个串 s_x，问集合 T 中有多少个串，包含 s_x。

- **数据范围**

  $N, Q \leq 10^5$

  $\sum |p|, \sum |s_i| \leq 2 \times 10^6$

CF590E 1.5s

- **题目大意**

  给定 N 个字符串 s_i，问最多能从中选择多少个串，使得其中不存在一个串是另一个串的子串。

- **数据范围**

  $N \leq 750$

  $L = \sum |s_i| \leq 10^7$

  $s_i \in \{a, b\}$

后缀三连

- 后缀数组
- 后缀自动机
- 后缀平衡树

- 还有上trie

后缀数组

- DC3不需要掌握，据说常数太大又不好打。

- O(n log n)给后缀排序
- O(n)求height（两个串同时去掉一个字符后的lcp长度-1）
- 配合height数组与rmq可以O(1)求子串lcp

- 容易理解，实现难度适中，使用频率中等

Trie上的后缀数组

- 这里排序排的是**从点x到根**的字符串
- 构造方法同样是倍增，区别并不大

- Height就直接倍增求，n log n

后缀自动机

- 接受所有子串的自动机
- 把所有子串分进了不同的right集合
- 理解不易，实现简单，使用频率高。

关于sam的几个有趣性质

- Fail是这个right集合的父亲。（所有合法right集合呈树形结构）
- 其包含的子串**更短**，出现位置**更多**（right集合更大）。

- **反串**的fail树是正串的后缀树
- 包含两个子串的最长公共**后缀**的right集合是包含他们的节点在fail树上的lca

- 沿着转移边走，right集合大小越来越小

Trie上的后缀自动机

- 也叫广义后缀自动机。
- bfs建法：按bfs序建，其中last取父亲
- dfs建法：按dfs序建。
- 复杂度据说bfs是O(n * sigma)，dfs是深度和。
- 证明emmmm（连sam都不证怎么会证这个

往事

- 给一颗n个点，字符集为300的trie
- 定义一个点对的价值是其所代表字符串的lcp+lcs.
- 求最大点对价值。
- N<=2e5

NOI2018 你的名字

- 给定一个母串s，有Q个询问：
  给出询问串T和区间$\mathbf{L},\mathbf{R}$，问T中有多少个本质不同的子串在s[L,R]中没有出现。

- |S|<=5e5, Q <= 1e5, $\sum T \leq 1e6$
- 68%的数据保证L=1,R=|S|

回文树 / 回文自动机

- 翁文涛的论文。
- 具体细节不赘述了，和ac自动机很类似。

Border

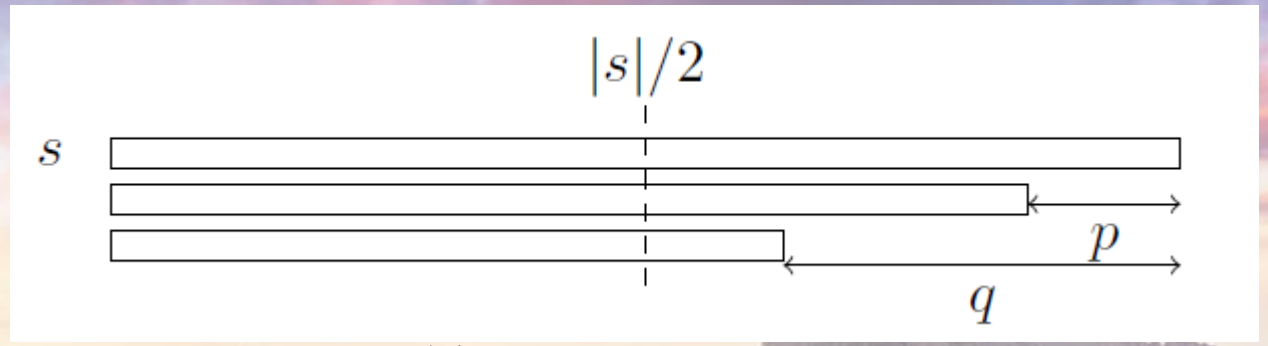
- 长度为x的Border就是长度为x的公共前后缀
- 就是你kmp出来那个fail
- 有x的border就是有n-x的周期
- **结论：所有border最多表示为log个等差数列。**
- **而且每一个等差数列都是排序后连续的一段。**
- 虽然这个结论太好记了，我们还是要证明一下。

证

- **弱周期引理：若a,b是周期，且a+b<=n，那么gcd(a,b)也是周期。**

- 证明：证辗转相减就可以了。不妨假设a<b
- 任意一个x都满足
- s[x]=s[x-a]=s[x+b-a]
- 或者s[x]=s[x+b]=s[x+b-a]，就能推出b-a是一个周期
- X+b-a不存在不要紧，这些x不影响
- 所以x-a>=1和x+b<=n有一条合法就可以了。
- 只要[a+1,n]交[1,n-b]是全集
- 所以就是a+b<=n.

用一下弱周期引理



- 先证明比s/2大的border可以被表示成一个等差数列
- 假设此时**最短**的周期是p
- 任意取一个比一半短的周期是q.
- 那么gcd(p,q)是一个周期
- 但是gcd(p,q)要>=p
- 所以q一定是p的倍数

- 当p,q比s/2要短的时候这都成立
- 周期是等差的，border也是等差的

那长度更小的border呢

- 小border是大border的border
- 找到第一个没有被安排到的border，长度又可以/2了
- 于是就是log n段了。

WC2016 论战捆竹竿

- 你有一个长度为n的母串，你每次可以把一个新的母串叠在旧的上面，但是重合的部分要相同。
- 例如:
- abcabc
- abcabc
- 现在问你，[1,w]中你能叠出多少种长度
- N<=5e5,w<=1e18
- 0.2s
- 256mb