字符串

2024 年 12 月

黄焖鸡

本次课主要内容

1. 基本例题选讲
2. 有关“log段等差数列”的题目选讲
3. 后缀数据结构题目选讲

知识回顾

1. Trie 树：合并相同的前缀。
2. 自动机：给定一个串，从前往后依次在自动机上遍历就完成了匹配。
3. KMP：设 f_i 表示 $S[1 \dots i]$ 最长 border 长度，可以利用其来建出自动机。KMP 也可以用哈希实现（怎么实现？）。$i - f_i$ 就是周期。
4. AC 自动机：多个模式串建 Trie 树，求出每个结点最长的后缀使得是 Trie 的前缀（称为 fail），再由此建自动机。
5. Manacher：求出以每个位置为中心的最长回文串长度，从左往右扫，记录当前最靠右的回文串，在已知信息基础上扩展，每次扩展必定推进右端点。
6. Z 函数：求 LCP($s[i \dots n], s$)，同 Manacher，从左往右扫，记录当前匹配区间的右端点。

例题

Teasers

划分 ARC060D

给定一个字符串，求至少能将其划分为几个非周期串（不是 $A^k\ (k \geq 2)$ 的形式的串），并求方案数。

$$|S| \leq 10^6$$

如果不是只有一种字符，则至多划分成两段（弱周期引理）。

例题

Teasers

印章 P3426

给出一个字符串 S，你有一个印章，不能抹去，同一个位置不能印两个字符，想要印出 S，求印章长度最小值。

$$|S| \leq 500000$$

提示：设 $f(i)$ 表示 i 这个前缀的答案，则 $f(i)$ 一定是 $1 \sim i$ 的 border。

设 $F[i]$ 表示前缀最长 border，$f(i)$ 要么是 $f(F[i])$ 要么是 i。

设 $g(i)$ 表示 $[1,i]$ 最多能用来覆盖哪个前缀，则 $g\big(f(F[i])\big) \geq i - F[i]$ 等价于 $f(i) = f(F[i])$。

例题

Teasers

括号序列 CF1610G

给定一个括号串，每次你可以删去一个长度为 2 的等于 () 的子序列，求若干次操作后可以得到字典序最小的字符串。

$$|S| \leq 300000$$

提示：如果移除了 S_i, S_j，何时能调整使这次操作更优？

总可以认为 $j = i + 1$，故只会删去连续一段匹配的括号。

对这 $O(n)$ 段匹配的括号 dp，可以用哈希快速比较字符串。

例题

Modest Substrings CF1110H

给定两个正整数 l, r，说一个字符串是好的当且仅当看作数（不能有前导 0）时在 l, r 之间。
求一个长为 n 的数字串好子串数最大，并输出方案。

$$l, r \leq 10^{800}, n \leq 2000$$

提示：如果 $r - l$ 很小直接建 AC 自动机就可以了。

注意到可以把一整个子树全存在的结点压缩成一个点，带上“过了 k 的长度会有 1 的贡献”的标记。在这棵压缩的树上 dp。

例题

Prefix of Suffixes QOJ9372

记 $z_i = LCP(S[i,\ldots],S)$。记 $f(S)=\sum_i \sum_{i\le j\le i+z_i-1} A_j B_i$。
每次在 S 后面在线添加一个字符，求 $f(S)$。

$|S|\le 10^6$，字符集 10^6。

提示：考虑 S 变长时 $f(S)$ 变化量。

若 i 有长为 x 的 border，会有 $A_i\times B_{i-x+1}$ 的贡献。然而，$\sum B_{i-x+1}$ 不好直接计算。

考虑 $\{i-x+1\}$：该集合除了增加 i，只会减少。变化量就是下一个位置不是 s_i 的位置。精细实现可以做到 $O(n)$。

例题

Matching Problem QOJ7748

给定两个字符串 S,T，记 $occ(S,T)$ 表示 S 在 T 中出现次数。

q 次询问 l,r，求 $\sum_i w_i \times occ(S[1 \dots i],T[l \dots r])$。

$$|S|,|T|,q \leq 5 \times 10^5$$

提示：每个 T 的后缀和 S 的 LCP 可以 Z 函数求，但如果直接算，不好用数据结构维护。

找出第一个超过的分界点，后面的只和 S 有关，可以 KMP 预处理。

例题

Teasers

Matching Problem II [OpenJudge - 2021C:Matching]

给定两个排列，长度为 $n \geq m$，求第二个排列在第一个排列中出现几次，其中出现定义为子串相对顺序一致。

$$n, m \leq 10^6$$

直接使用 KMP 算法，判断需不需要跳 fail 就是二维数点。

例题

Teasers

AC Automation Chicken QOJ9318

有一棵 Trie 树，树边从根往下；根据 Trie 建立 AC 自动机的 fail 树，树边从子结点往根。将这 $2n-2$ 条有向边告诉你，试还原出根和 Trie。

$$n \le 5 \times 10^5$$

提示：先找根，再还原 Trie。根有什么性质？知道根了 Trie 有什么性质？

根：周围全是重边。

如果重边构成一条链，根据 fail 也可以锁定 $O(1)$ 个候选的根。

例题

Teasers

AC Automation Chicken QOJ9318

有一棵 Trie 树，树边从根往下；根据 Trie 建立 AC 自动机的 fail 树，树边从子结点往根。将这 $2n-2$ 条有向边告诉你，试还原出根和 Trie。

$$n \leq 5 \times 10^5$$

知道根了，Trie 就是根出发的 BFS 树，而字符可以根据 fail 填：fail 不是根填的字符已经确定，否则可以随便填一种新字符。

可以用栈 $O(n)$ 判断合法性，当然也可以可持久化线段树。

例题

Teasers

合法缩写

ARC141F

给定一些字符串 $S_1, \dots S_n$，若存在字符串 S，使得不停从 S 中删去某个等于某个 S_i 的子串，直到删不了为止，最终能得到至少两种结果，则称字符串集合 $\{S_i\}$ 是坏的，否则是好的。

判断给定的字符串集是否好。

字符集 ABCD，$\sum |S_i| \leq 2000000$

提示：先想一些简单的不好的情况，找出不好的“症结”。

例题

合法缩写 **ARC141F**

判断给定的字符串集是否好。

字符集 ABCD，$\sum|S_i| \le 2000000$

问题的症结是消去的方式存在 overlap。

换句话说，若存在 A, B, C $(A \ne C)$，使得字符串集中两个字符串分别形如 AB, BC，则 ABC 会有两种消去方式——且这是唯一可能产生分歧之处，也即这是必要条件。这和充分条件之间还差了什么？

消去之后剩余的部分可能还会再被消去，对分析带来困扰。但可以适当预处理（从短到长贪心匹配子串），提前使得 S_i 互不为子串。充分了！

例题 Teasers

合法缩写 **ARC141F**

判断给定的字符串集是否好。

字符集 ABCD，$\sum|S_i| \le 2000000$

预处理使得 S_i 互不为子串可以借助 AC 自动机匹配。判断是否存在两个字符串形如 AB, BC $(A \neq C)$，可以枚举 AB，合法的 B 就是 fail 树上 AB 的到根链。注意在这里可以暴力跳到根链，因为总之是 $O(\sum|S_i|)$ 的。

Border 等差数列

一个字符串的所有 Border 和所有回文前缀均构成 $O(\log n)$ 个等差数列。

可以在 KMP 自动机上 / PAM 上预处理出每个等差数列的首末点。（回忆：PAM 是什么？）

每个 Border / 回文后缀等差数列均可写成 $A^k B, \ldots, B$（以及 $BA^k, \ldots, B$）的形式。

不讲的题：P5287，P4156

例题

字符串计数 **P1393**

给定 S，计算有多少个长为 n 字符集为 m 的字符串包含至少一个 S，要求支持任意模数。

$$|S|, n \le 10^6$$

二维 dp 显然无法优化，试着设计一维 dp。

设 $f(i)$ 表示 S 第一次出现在 $[i-|S|+1, i]$ 的前 i 个字符方案数，容斥转移：枚举上次出现在哪里。

转移需要枚举所有 Border 长度 x 并令 $f(i)$ 减去 $f(i-x)$，可以用 log 个前缀和维护。

例题

Palindrome Partition CF932G

给定一个字符串，求划分为偶数 $2k$ 段，第 i 段和第 $2k-i+1$ 段相同的方案数。

$$|S| \le 10^6$$

Reverses CF906E

给定两个字符串 S,T，问有几种方案，反转若干个 S 的不交区间，变成 T。

$$|S| = |T| \le 10^6$$

两个问题本质上都是 $f_i = \sum_{j<i} f_j \times [S[i \dots j]\text{回文}]$ 的问题。

例题

Palindrome Partition	CF932G

Reverses	CF906E

$f_i = \sum_{j<i} f_j \times [S[i \dots j]回文]$。将回文后缀划分为等差数列，考虑每个等差数列对应的后缀左端点位置：除了最短的一个，都和上次是重合的。

只需维护一个数组 g 描述这个过程，可以做到 $O(n \log n)$。

例题

区间本质不同回文子串　　**你们应该知道题号**

给定一个字符串，q 次询问区间本质不同子串个数。

$$n, q \leq 2 \times 10^5$$

枚举 r，把每个本质不同回文串最后一次出现的左端点处打上标记。

同前面的题，可以发现每段等差数列标记的变化量是 $O(1)$ 的。

例题

回文 **QOJ5037**

给定一个字符串，支持单点修改、询问区间回文后缀数量。

$$n, q \leq 200000, 8s$$

提示：直接线段树维护回文后缀构成的 log 段等差数列。如何合并？

当合并 S, T 时，讨论新回文后缀的中线处于何处。

1. 处于 T 且完全包含于 T：继承 T 的信息。
2. 处于 T 且包含 S 的一个后缀：注意此时 critical 的是 T 前缀的等差数列。
3. 处于 S, T 之间：只有 1 个串。
4. 处于 S：critical 的是 S 后缀的等差数列。

例题

回文 **QOJ5037**

给定一个字符串，支持单点修改、询问区间回文后缀数量。

$$n, q \le 200000, 8s$$

考虑情况 2，设等差数列的一段为 BA^k，$T = BA^kU$，显然 U 的前缀不是 A，需判断 U 是否是 A 的前缀。若不是，新回文串只有可能以 BA^k 为中心。否则，新回文串可以以一个等差数列的后缀为中心（只取决于 S）。后缀长度可以二分 + 哈希得到。
情况 4 也类似：设等差数列的一段为 A^kB，只需讨论 T 是否形如 A^kA'（其中 A' 为 A 前缀）。不是的话只能以 A^kB 为中心，否则可以以一段前缀为中心。

采用分块维护哈希，可以做到 $O(\log^2 n + \sqrt{n})$ 单次询问。

后缀数据结构例题选讲

$$\text{lcp}(sa_i, sa_j) = \min_{k=i\sim j-1} \text{lcp}(sa_k, sa_{k+1})$$

后缀树可以理解成反串的 parent 树，也可以理解成后缀 Trie 的压缩。
后面我们会看到，两者都有用。

下面给了几个经典题，感觉大家大部分都做过，可以自己课后看一眼。
1. [NOI2018] 你的名字
2. CF666E
3. CF700E
4. CF1608G
5. CF1801G
6. CF1098F
7. [NOI2023] 字符串

例题

打击复读

UOJ577

给定字符串 S，每个位置还给出权值 vl_i, vr_i，支持单点修改 vl，以及计算

$$\sum_{1 \le i \le j \le n} \sum_{S[p,q]=S[i,j]} vl_p \sum_{S[p,q]=S[i,j]} vr_q$$

$|S| \le 500000$，要求线性。

只需对每个左端点求出以其为左端点的串的出现位置 vr 之和之和即可。

提示：在后缀树上 dfs，每个结点代表一个子串 $[x,y]$，走一条转移边就是 $[x,y] \to [x,z]$ $(z > y)$，维护增量，走到叶子就算出了所求值。

例题

打击复读 UOJ577

$$\sum_{1\le i\le j\le n}\sum_{S[p,q]=S[i,j]} vl_p \sum_{S[p,q]=S[i,j]} vr_q$$

后缀树是压缩！如果存在 $[i,x]\to[i,y]$ 的边，要么 $y=n$，要么 $[i,y]$ 有两种出边。换句话说，只需计算“走到下一个有两种出边的点”的 vr 贡献，而这一贡献又可以在正串 SAM 上预处理。

总结：本题是利用正串 SAM 和正串后缀树“对称性”的典范。

例题

字符串问题 P5284

现有一个字符串 S。

Tiffany 将从中划出 n_a 个子串作为 A 类串，第 i 个（$1 \leqslant i \leqslant n_a$）为 $A_i = S(la_i, ra_i)$。

类似地，Yazid 将划出 n_b 个子串作为 B 类串，第 i 个（$1 \leqslant i \leqslant n_b$）为 $B_i = S(lb_i, rb_i)$。

现额外给定 m 组支配关系，每组支配关系 (x, y) 描述了第 x 个 A 类串**支配**第 y 个 B 类串。

求一个**长度最大**的目标串 T，使得存在一个串 T 的分割 $T = t_1 + t_2 + \cdots + t_k$（$k \geqslant 0$）满足：

- 分割中的每个串 t_i 均为 A 类串：即存在一个与其相等的 A 类串，不妨假设其为 $t_i = A_{id_i}$。
- 对于分割中所有相邻的串 t_i, t_{i+1}（$1 \leqslant i < k$），都有存在一个A_{id_i} 支配的 B 类串，使得该 B 类串为 t_{i+1} 的前缀。

方便起见，你只需要输出这个最大的长度即可。

特别地，如果存在无限长的目标串（即对于任意一个正整数 n，都存在一个满足限制的长度超过 n 的串），请输出 -1。

本题无非是一个 DAG 最长路径问题，只需快速建出 DAG。

例题

字符串问题 P5284

现有一个字符串 S。

Tiffany 将从中划出 n_a 个子串作为 A 类串，第 i 个（$1 \leqslant i \leqslant n_a$）为 $A_i = S(la_i, ra_i)$。

类似地，Yazid 将划出 n_b 个子串作为 B 类串，第 i 个（$1 \leqslant i \leqslant n_b$）为 $B_i = S(lb_i, rb_i)$。

现额外给定 m 组支配关系，每组支配关系 (x, y) 描述了第 x 个 A 类串**支配**第 y 个 B 类串。

求一个**长度最大**的目标串 T，使得存在一个串 T 的分割 $T = t_1 + t_2 + \cdots + t_k$（$k \geqslant 0$）满足：

- 分割中的每个串 t_i 均为 A 类串：即存在一个与其相等的 A 类串，不妨假设其为 $t_i = A_{id_i}$。
- 对于分割中所有相邻的串 t_i, t_{i+1}（$1 \leqslant i < k$），都有存在一个A_{id_i} 支配的 B 类串，使得该 B 类串为 t_{i+1} 的前缀。

方便起见，你只需要输出这个最大的长度即可。

特别地，如果存在无限长的目标串（即对于任意一个正整数 n，都存在一个满足限制的长度超过 n 的串），请输出 -1。

每种连边都可以看作在后缀树（反串 SAM）上向一个子树连，但有可能不是恰好向子树，是向一个结点中间。此时只需把该结点分裂。

例题

区间最长 Border **P4482**

给定一个字符串，q 次询问区间最长 Border。

$$n, q \leq 2 \times 10^5$$

直接写成 LCP 的形式硬做。最小的 i 使得 $LCP(l, i) \geq r - i + 1$ $(l < i \leq r)$。LCP 可以用后缀数组笛卡尔树处理。

枚举 LCP 取到区间最小值的位置，还可以枚举小的子树。

1. 若 l 在小子树内，就是在大子树内求个后继。
2. 若 i 在小子树内，考虑 i 对 (l, r) 询问的贡献，要求 $r \in [i, i + t - 1]$（t 定值），$l < i$ 且在另一侧。把所有 i 和所有询问一起做，从小到大加入 i，按照 i 加入 l，每次需要删掉一个关于树和 r 的矩形内所有点。

例题

优秀的拆分 **P1117**

给定一个字符串，求有几个形如 AABB 的子串。

$$n \le 2 \times 10^5$$

只需求出每个位置是几个 AA 的左/右端点（记为 f_j）。枚举 AA 的 A 长度 i，取所有 ki 位置为关键位置。

每个 AA 恰好经过两个关键位置 u, v，且“是 AA”可以被写为 $LCP(u,v), LCS(u,v)$ 的限制。

对于固定的 u, v，AA 是连续的一段，因此只需对 f 区间加。（这也使我们可以完全掌握 AA 出现位置）

例题

区间分段 **CF1043G**

q 次询问，每次询问：把 $S[l \dots r]$ 划分为若干段，使得划分出来的段至少有两段相同。至少有几种段。

$n, q \leq 200000$，$7s$

只要有重复的字符 c，就可以分为 $AcBcD$，答案 ≤ 4。

1. 答案为 1：判断是否周期，可以哈希。
2. 答案为 2：可以规约 AAB，ABA，BAA。前后用优秀的拆分，中间用最长 Border。
3. 答案为 3：可以规约 ABAC，BACA，BAAC。只有后者 non-trivial，是优秀的拆分的偏序问题。

例题

区间本质不同子串个数 **P6292**

给定一个字符串，q 次询问区间本质不同子串个数。

$$n, q \leq 2 \times 10^5$$

类似回文的版本，离线，r 变大，维护 l 的答案。但问题是，r 的一次移动对 l 的影响可能很大。

注意到在 SAM 上，跳后缀链接过程中，一段 endpos 相同连续段的答案变化量还是 $O(1)$，故实际上的影响是 $O($连续段数$)$。用 LCT Access 的均摊分析即知总变化量是 $O(n \log n)$ 级别的。

例题

机器蚤分组 **UOJ608**

定义 $f(S)$ 为 S 所有子串在包含关系的 DAG 上的最小链覆盖。给定一个字符串 S，q 次询问 $f(S[l \dots r])$。

$$n, q \le 2 \times 10^5$$

提示：找到 f 的一个下界并尝试证明。

$f(S)$ 等于最大的某个长度的本质不同子串个数。显然至少取这么多个子串，如何证明这么多也够？提示：Dilworth 定理。

可以取长度最小的反链中子串，尝试调整（把长度加一）。如果长度不全相同，一定存在合法的调整。

例题

机器蚤分组 **UOJ608**

定义 $f(S)$ 为 S 所有子串在包含关系的 DAG 上的最小链覆盖。给定一个字符串 S，q 次询问 $f(S[l \dots r])$。

$$n, q \leq 2 \times 10^5$$

提示：最大的某个长度的本质不同子串个数仍然不好算，有没有使其好算的转化？试着思考“什么时候子串会本质相同”。

事实上，$f(S)$ 的计算只需算最短的 len 使得长为 len 的子串两两不同。

考虑反面，只需算最长的出现两次的子串长度，这可以用区间本质不同子串数的 SAM+LCT 方法。

例题

NIT 学 KMP

InfOJ #53

给出字符串 S。还有一个字符串 T。初始时，T 为空串。

接下来有 q 次对 T 的修改，每次修改为以下两种之一：

- `1 U`，U 是一个字符串：表示在 T 的末尾接上 U。
- `2 l k`，$1 \le l \le |T|$，$k \ge 1$，表示重复 k 次，第 i 次的操作为：设 $pos = l + i - 1$，将 T_{pos} 接在 T 的末尾。

请在每次修改后求出 S 在 T 中的出现次数。

所有数据满足：$1 \le m \le 2 \times 10^5$，$1 \le q \le 10^4$，$1 \le |T| \le 10^{15}$，$\sum |U| \le 2 \times 10^5$。保证题目中的字符串仅包含小写字母。

提示：先考虑 $m = 1$ 和 $m \le 5$ 怎么做。

例题

NIT 学 KMP

InfOJ #53

当 m 很大时，考虑需要维护什么信息才能计算两个区间拼在一起时新增的贡献。

对于每个区间，维护它最长的前缀，使得该前缀在 S 中作为子串 $S[l_1 \dots r_1]$ 出现了；维护它最长的后缀，使得该后缀在 S 中作为子串 $S[l_2 \dots r_2]$ 出现了。

合并信息

假设已知平衡树上左儿子的前缀信息为 $[l_1, r_1]$，右儿子的前缀信息为 $[l_2, r_2]$，如何合并出整个串的前缀信息？

若 $[l_1, r_1]$ 的长度不等于左儿子的长度，显然整个串的前缀信息就是 $[l_1, r_1]$。否则，预处理出后缀数组，可以二分出 $S[l_1, r_1] + S[l_2, r_2]$ 在原串中的排名，进而得到新信息。

后缀信息是一样的，用反串后缀数组即可。

算新贡献

我们的问题是：S 在 $S[l_1, r_1] + S[l_2, r_2]$ 中出现了多少次。那，S 的一个前缀就是 $S[l_1, r_1]$ 的后缀，S 的一个后缀就是 $S[l_2, r_2]$ 的前缀。在正串 kmp fail 树上找到 $[l_1, r_1]$ 对应的节点，反串 fail 树上找到 $[l_2, r_2]$ 对应的节点，就是求两个点到根路径上有多少对结点对应 border 长度之和恰好为 n。

dfs 其中一棵树，在 dfs 到 x 点的子树之前，把另一棵树上 $n-x$ 对应的子树权值加一，询问就是查询 dfs 到 x 时，另一棵树上 y 的权值。这容易离线后（也就是先用平衡树得出结构，得到 fail 树上的具体询问后再离线）用树状数组处理。

例题

字符串游戏 **P10215**

给定一个字符串 S，A 和 B 玩游戏，每次当前玩家选一个后缀 $S[i \dots n]$，获得 $occ(S[i \dots n], S)$ 分，并将 $S[i \dots n]$ 删去（之后算 occ 的时候也在删去后的 S 上）。

两人都希望自己的得分 – 对方的得分最大，问最终先手得分 – 后手得分。

$$|S| \le 10^6$$

容易写出转移方程 $f(i) = \max\limits_{j<i} occ(S[j+1 \dots i], S[1 \dots i]) - f(j)$。首先，occ 的计算是关于 parent 树和位置的二维数点。

提示：利用 occ 单调性，你能得出什么结论？

例题

字符串游戏 **P10215**

两人都希望自己的得分 – 对方的得分最大，问最终先手得分 – 后手得分。

$$|S| \le 10^6$$

$f(i) = \max_{j<i} occ(S[j+1 \dots i], S[1 \dots i]) - f(j)$。注意到 occ 随着 j 减少依次减少，故若 $j < k, f(j) \ge f(k)$，则 j 就没用了。因此合法转移点构成 f 增加的单调栈。（事实上，occ 满足反向的四边形不等式，但由此导出的做法需要两个 log，无法通过本题）

提示：本来是根据 $f(i)$ 的值试图把 i 入栈，但现在我们无法定位转移点。能否一边将 i 入栈一边确定转移点？

例题

字符串游戏 **P10215**

两人都希望自己的得分 - 对方的得分最大，问最终先手得分 - 后手得分。

$|S| \le 10^6$

$f(i) = \max_{j<i} occ(S[j+1\dots i], S[1\dots i]) - f(j)$，合法转移点构成 f 增加的单调栈。如果用栈顶 p 转移 $f(i)$ 已经有 $f(i) \le f(p)$ 了，则可直接弹栈！这并不影响复杂度。但如果得到的 $f(i) > f(p)$，而实际转移点 $q < p$ 怎么办？

$$
\begin{aligned}
f(i) &= occ(S[q+1\dots i], S[1\dots i]) - f(q) \\
&\le occ(S[q+1\dots p], S[1\dots p]) - f(q) \\
&\le f(p)
\end{aligned}
$$

例题

Border 的第五种求法 **UOJ752**

给定一个字符串，q 次询问区间 Border 的 f_{occ} 之和。

$$n,q \le 5\times 10^5, 6s$$

Occ 和该字符串具体在 SAM 哪个点上有关，因此本题需要定位所有 Border 在 SAM 上的出现位置。

如果在 SAM 上依次匹配 $[l,r]$，走到的就是 $[l,r]$ 的所有前缀，因此“是 border”可以写为：能在上述过程中走到，且属于 $[1,r]$ 子树。

前者可以 DAG 剖分转为偏序问题，后者已经是偏序问题，其实就是总共 $O(q\log n)$ 个二维偏序。

Thanks!

拓展学习内容：

- 基本子串结构
- 基本子串字典（似乎可以被前者包含）
- Lyndon 理论
- Runs 理论