字符串入门

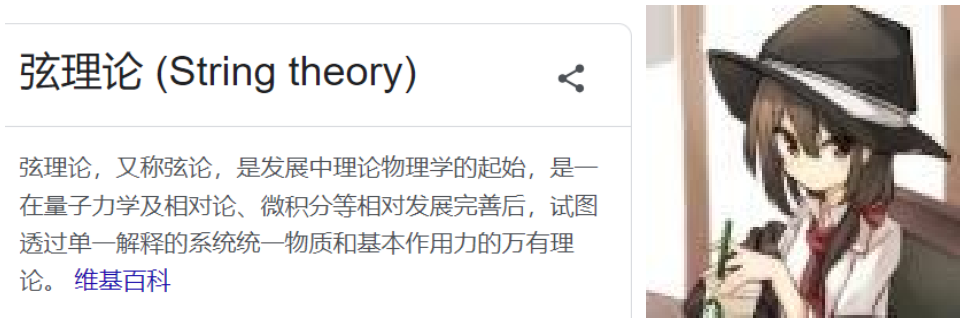
Cage

msannu

December 20th, 2023

First of all

啥是字符串?



(划掉)

$|\Sigma|$ 小? DS（自动机）? 均摊? 复用信息? 科技?

希望选择的几道典型例题能帮助大家找到做“串串题”的“感觉”。

kmp

相信大家都会。

kmp

相信大家都会。

kmp 为初学字符串的我们提供了一个强大的结构，border。

但同时，由于 kmp 是均摊的，在一些情况下我们希望避免均摊，该怎么改进?

对于 $|\Sigma|$ （字符集）比较小的情况，考虑 kmp 本身也是一个 ACAM。在 fail 树上维护额外信息（父向方向上，上一个字符 c 出现的位置），亦可作到 $\mathcal{O}(n|\Sigma|)$。

kmp

相信大家都会。

kmp 为初学字符串的我们提供了一个强大的结构，border。

但同时，由于 kmp 是均摊的，在一些情况下我们希望避免均摊，该怎么改进?

对于 $|\Sigma|$ （字符集）比较小的情况，考虑 kmp 本身也是一个 ACAM。在 fail 树上维护额外信息（父向方向上，上一个字符 c 出现的位置），亦可作到 $\mathcal{O}(n|\Sigma|)$。

当 Σ 比较大的时候，我们可以使用可持久化数组去维护额外信息，可以做到 $n \log |\Sigma|$。

kmp

相信大家都会。

kmp 为初学字符串的我们提供了一个强大的结构，border。

但同时，由于 kmp 是均摊的，在一些情况下我们希望避免均摊，该怎么改进？

对于 $|\Sigma|$ （字符集）比较小的情况，考虑 kmp 本身也是一个 ACAM。在 fail 树上维护额外信息（父向方向上，上一个字符 c 出现的位置），亦可作到 $\mathcal{O}(n|\Sigma|)$。

当 Σ 比较大的时候，我们可以使用可持久化数组去维护额外信息，可以做到 $n \log |\Sigma|$。

但这些都不得不去依赖所谓“字符集”，进行“数据结构”上的小把戏，我们有没有更加优美**不均摊**的字符串匹配做法呢？

notation

开始前我们还是需要来明确一些名词，记号。

border: 若 k 满足 $s[1,2,\cdots,k]=s[n-k+1,\cdots,n]$ 我们称 $s[1,2,\cdots,k]$ 是 s 的 border。

period: 若 k 满足 $\forall i\in[1,n-k], s_i=s_{i+k}$ 称 k 是 s 的一个周期。

如果 s 有一个长为 k 的 border，则必然存在长为 $n-k$ 的 period。

WPL (Weak Periodicity Lemma.)

Lemma

若 p, q 均为 s 的周期，且 $p + q \leq |s|$ 则 $\gcd(p, q)$ 亦为 s 的周期。

Proof.

弱周期引理证明是容易的，我们考虑在 $p + q \leq |s|$ 时，由贝祖定理产生的 $ap - bq$ 等价类间两两可到达。

$(a, b) \to (a, b - 1)\texttt{or}(a - 1, b)$。 □

PL (Periodicity Lemma.)

Lemma

若 p, q 均为 s 的周期，且 $p+q-\gcd(p,q) \le |s|$ 则 $\gcd(p,q)$ 亦为 s 的周期。

Proof.

周期引理作为 border 理论中一重要引理，有各种各样的证明方式，这里给出一种生成函数的证明。

记：$s_p(x) = \frac{A(x)}{1-x^p}, s_q(x) = \frac{B(x)}{1-x^q}$。

$$s_p(x) - s_q(x) = \frac{1-x^{\gcd(p,q)}}{(1-x^p)(1-x^q)}\left(\frac{A(x)(1-x^q)}{1-x^{\gcd(p,q)}} - \frac{B(x)(1-x^p)}{1-x^{\gcd(p,q)}}\right)$$

注意到后式 $P(x)$ 是度数不超过 $p+q-\gcd(p,q)$ 的多项式。

则若 $s_p(x) - s_q(x) \mod x^{p+q-\gcd(p,q)} = 0 \implies P(x) \equiv 0.$ □

PL (Periodicity Lemma.)

WPL/PL 给我们的启示是巨大的。PL 的证明的最后推出的 $P(x) \equiv 0$ 昭示着“周期”和“整周期”之间的区别似乎不大。

下面有一个众所周知的性质：一个字符串的 border 可以被**划分** $\mathcal{O}(\log n)$ 段等差数列。

证明考虑切换 period 时，串长减半。

那 fail 树呢?

JOJO

Source: HNOI2019 JOJO[1]

题意: 动态维护字符串每一个前缀的 Border 集合大小的和。

操作形如，往字符串后加入 k 个字符 c。以及时间回溯操作（不强在)。

$|\Sigma| = 26, q \leq 10^5$。

hint:

[1] https://loj.ac/p/3055

JOJO

Source: HNOI2019 JOJO[1]

题意: 动态维护字符串每一个前缀的 Border 集合大小的和。

操作形如，往字符串后加入 k 个字符 c。以及时间回溯操作（不强在)。

$|\Sigma| = 26, q \leq 10^5$。

hint: 题目保证两次插入的字符不同。

[1] https://loj.ac/p/3055

JOJO

Prefixuffix

Source:「POI2012」Prefixuffix[2]

题意：定义 s_1, s_2 “循环相同”当且仅当，存在 s_1 的一个前缀 t_0 (可空)，$s_1 = t_0 + t_1$，使得 $s_2 = t_1 + t_0$。

给定一个长为 n 的串 t，求出满足条件最大的长度 $L \leq \frac{n}{2}$：$s[1, L]$ 与 $s[n - L + 1, n]$ 循环相同。

$n \leq 10^6$。

[2]https://loj.ac/p/2704

Prefixuffix

Source:「POI2012」Prefixuffix[2]

题意：定义 s_1, s_2 “循环相同”当且仅当，存在 s_1 的一个前缀 t_0 (可空)，$s_1 = t_0 + t_1$，使得 $s_2 = t_1 + t_0$。

给定一个长为 n 的串 t，求出满足条件最大的长度 $L \leq \frac{n}{2}$：$s[1, L]$ 与 $s[n - L + 1, n]$ 循环相同。

$n \leq 10^6$。

被期待的做法是线性的。

[2]https://loj.ac/p/2704

Prefixuffix

Manacher

如何求出来一个串的所有“本质不同的”回文子串呢?

我们有结论，忽略平凡回文子串（单个字符），一个串本质不同的回文子串有至多 n 个。

考虑从长度为 n 的串加入一个字符新增的回文串个数至多为 1。

基于此我们可以得到 manacher 算法。

PAM

将 $\mathcal{O}(n)$ 个回文串，以回文后缀关系连边，形成的树被称为“回文树”。

考虑类似 manacher 的处理方式，增量构造。

需要特别注意的是，回文串的回文后缀亦是该串的 border。

Z 函数

如果我们希望快速求出来一个串的所有后缀与该串的 lcp 该怎么办?

这个算法的感觉和 manacher 是非常像的，所以放到了一起。
具体的，我们从 $i:1\to n$ 的顺序增量计算 $\texttt{lcp}(s,s[i,n])$。
维护类似“右端点最靠右”的 lcp。

pkusc23 d1t1

题目大意：给定 S, T，对于每个 i 求 S_i 换为 T_i 后 S 的最长 Border。

$n \leq 10^6$，$n \leq 10^7$？可能暂时没得提交位置。

pkusc23 d1t1

题目大意：给定 S, T，对于每个 i 求 S_i 换为 T_i 后 S 的最长 Border。

$n \le 10^6$，$n \le 10^7$？可能暂时没得提交位置。

应该就是模板题吧。

pkusc23 d1t1

题目大意：给定 S, T，对于每个 i 求 S_i 换为 T_i 后 S 的最长 Border。

$n \le 10^6$，$n \le 10^7$？可能暂时没得提交位置。

应该就是模板题吧。

切割

来源：【数据删除】。

题目大意：给你个字符串，每次询问其子串有多少种切割方式，使得其可以被切分成两个回文串。

$n, q \leq 2 \times 10^5$。

切割

ACAM

值得注意的是：对于长度和为 $l = \Sigma_{i=1}^{k}|S_i|$ 的串集，我们可以用 $O(l)$ 的时间暴力跳 fail 来建立 fail 树。

而对于 n 个结点的 Trie 我们经典建立 $ACAM$ 的时间复杂度是 $\mathcal{O}(n\Sigma)$，而瓶颈在于“节点复制”，这也提示这我们可以通过“可持久化”数据结构来优化这一部分的复杂度。

>.<

Source: 集训队互测 Round 6[3]

题目大意：给定一个 n, m 的图，ban 了 k 条路经 ($\Sigma_{i=1}^{k}|l_i| \leq 2 \times 10^5$)，求 $1 \rightsquigarrow n$ 最短路（不必是简单路径)。

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

喂，你不是讲字符串的吗?

[3]https://qoj.ac/contest/1037/problem/5034

>.<

喂，你不是讲字符串的吗?

Boring Problem

Source:The 2020 ICPC Macau B[4]

题目大意：你有一个串 S 和一个长度均为 m 的串集 $\mathcal{T}$，在 S 后面随机加字符，什么时候 S 子串中出现 $\mathcal{T}$ 中的串就停止。

给一个串 R 对 R 每一个前缀求期望停止步数。

$n \leq 100, n \times m \leq 10^4, |\Sigma| \leq 26$?

$n \leq 100, n \times m \leq 10^5, |\Sigma| \leq 10^5$?

[4]https://codeforces.com/gym/103119/problem/B

解释

Boring Problem

SAM

Suffix Automata 后缀自动机。后缀自动机**并不是**最小化的子串自动机。

对于复杂度，抛去内存拷贝，我们感性地说明构造 SAM 是线性的。

画图.jpg

拜神

Source: P7361 「JZOI-1」拜神[5]
先来道开胃小菜。
给定字符串 s,q 次询问 $s[l,r]$ 出现大于两次的最长子串长度。
$|s| \leq 5 \times 10^4, q \leq 10^5$

[5]https://www.luogu.com.cn/problem/P7361

拜神

String Journey

Source: CF1063F[6]

大致题意：求最大的 k，使得字符串 s 存在一个 k 划分，k 划分规则是：$s = u_1 + t_1 + u_2 + t_2 + \cdots + t_k + t_{k+1}$。

需要保证，$t_i \subset t_{i+1}$，u_i 无特殊限制。

$n \leq 5 \times 10^5, |\Sigma| \leq 26$?

$n \leq 2 \times 10^6$?

[6] https://codeforces.com/contest/1063/problem/F

String Journey

String Journey

Invincible Hotwheels

Source: 2022ccpc guilin I [7]

大致题意：给你一个字符串集 $\mathcal{S}$，求所有二元组 (i,k) 满足，存在且只存在一个 j，$\mathcal{S}_k \subset \mathcal{S}_j \subset \mathcal{S}_i$。

$\Sigma l_i \leq 2 \times 10^6, |\Sigma| \leq 26$。

~~好像是炒冷饭？不管了，炒的就是冷饭。~~

[7] https://codeforces.com/gym/104008/problem/I

Invincible Hotwheels

Everybody Lost Somebody

Source: 15th HL Provincial CP contest E[8]

题目大意：给你一个串 s 的全部 sa 数组和部分 ht 数组。让你求出满足条件字典序最小的串。

$n \leq 5000, |\Sigma| \leq 26$? $n \leq 10^6$? $n \leq 5 \times 10^6$?

[8] https://codeforces.com/gym/102803/problem/E

Everybody Lost Somebody

Everybody Lost Somebody

玄题咋讲

省流：~~没活了，给大家咬（几）个打火机。~~

differences

Source: CERC2022 F[9]

题目大意：给一个大小为 m 的字符串集 $\mathcal{S}$ 其中恰好存在一个串满足该串与其余串的编辑距离是 k。

请你找出这个串。

$n, m \leq 10^5, nm \leq 10^7, |\sigma| \leq 4$。

[9] https://qoj.ac/contest/1070/problem/5254

Border 的四种求法

Source: [BJWC2018] Border 的四种求法[10]
题目大意：给定一个串 s，q 次询问区间 $[l, r]$ 的 border。
$n, q \leq 2 \times 10^5$。
题如其名，这道题有众多做法，只介绍 dag 链剖分做法。
~~你自己尝了吗~~

[10] https://www.luogu.com.cn/problem/P4482

Border 的四种求法

节日庆典

Source: [JSOI2019] 节日庆典[11]
题目大意：给你一个串 s 求出对于 s 所有前缀的最小表示法。

[11] https://www.luogu.com.cn/problem/P5334