

# 立命合宿 2018 Day 3

## G：検閲により置換

原案，問題文：杉江

解説：鈴木

解答：栗田，鈴木，杉江

# 問題概要

- テキスト  $T$  とパターン集合  $\mathcal{P}$  が与えられる
- $T$  の一部を '\*' に置き換えて、任意のパターン  $P \in \mathcal{P}$  が出現しないようにしたい
- '\*' に置き換えるべき最小文字数はいくつ？

$T = \text{brainfuck}$



$\text{brai} * \text{fuc} *$

$\mathcal{P} = \{\text{rain}, \text{inf}, \text{fuck}\}$

- 制約
  - $|T| \leq 100,000$
  - $|\mathcal{P}| \leq 100,000, \sum_{P \in \mathcal{P}} |P| \leq 100,000$

# 解法：文字列マッチング+貪欲法

- $T$ の前方から各文字列 $P \in \mathcal{P}$ をマッチングしていく
- マッチングした順に区間を並べる
  - $[s_1, t_1], [s_2, t_2], \dots, [s_k, t_k]$  ( $t_1 \leq t_2 \leq \dots \leq t_k$ )

$T = \text{brainfuck}$

$\mathcal{P} = \{\text{rain}, \text{inf}, \text{fuck}\}$



$[2, 5], [4, 6], [6, 9]$

- for  $i = 1, \dots, k$ :
  - if 区間 $[s_i, t_i]$ に '\*' で置き換えられた文字がない
    - then  $t_i$  を隠す

$\text{brainfuck} \rightarrow \text{brai} * \text{fuck} \rightarrow \text{brai} * \text{fuc} *$

# 等価な問題

- 区間スケジューリング問題

$n$  個の区間  $[s_1, t_1], \dots, [s_n, t_n]$  が与えられる.  
区間が被らないように最大何個選べるか.

- 解法

- $t_i$  の昇順にソート
- 先頭から順に貪欲に選ぶ

$[2,5], [4,6], [6,9]$



# 高速な複数文字列マッチング

- Aho-Corrasick
  - 圧倒的ライブラリ力
- ローリングハッシュを使う方法
  - 「短い文字列集合」と「長い文字列集合」にわけるとよい
    - 大体  $\sqrt{\text{長さの総和}}$  未満と以上にわけるとよい
    - 短い文字列集合：短いが多い
      - 長さごと ( $\sqrt{\text{長さの総和}}$  種類未満) にハッシュ値を分類し、まとめてマッチング
    - 長い文字列集合：長いが少ない ( $\sqrt{\text{長さの総和}}$  個以下)
      - 普通に1個ずつハッシュ値を計算してマッチング
- 他にも色々できると思います

# Memory Limit Exceeded の罣

- 最悪で $10^7$ 個程度の区間が出現しうる
- すべて陽に管理してしまうとメモリが足りない

# 省メモリ化

- すべての区間が  $[1, N]$  に収まることを利用する
- 各  $t \in \{1, \dots, N\}$  について、終点が  $t$  であるような区間の中で始点が最大のものだけ覚えておく
  - $maximum[t] := \max\{s \mid \text{区間}[s, t] \text{ が存在する}\}$
  - 終点  $t$  の区間を使うとき  $[maximum[t], t]$  を選んだことにする
- これで区間管理部分の空間計算量は  $O(N)$

# Writer解

	言語	行数	Byte	備考
栗田	C++	149	3709	Trie
鈴木	C++	86	1845	口リハ
杉江	C++	105	3210	Aho



# 提出状況

- Acceptance Rate
  - 28% (23 / 80)
- First Accept
  - オンサイト
    - RUPC\_ENERGY\_STAR (1:18)
  - オンライン
    - pekempey (0:41)