เอกสารประกอบการอบรมคอมพิวเตอร์โอลิมปิกวิชาการ ค่าย 2
27 มีนาคม – 22 เมษายน 2563

ขั้นตอนวิธีการจับคู่สายอักขระ
(String matching algorithm)

ปัญหาการจับคู่สายอักขระ (String matching problem)

หลักการ

- ให้ T = T[0..n-1] : ข้อความ (text)

  P = P[0..m-1] : แบบ/แบบรูป/แบบอย่าง (pattern) ที่ต้องการค้นหาในข้อความ T
- เมื่อ
  - n = |T| : ความยาว/ขนาดของ T
  - m = |P| : ความยาว /ขนาดของ P

ปัญหาการจับคู่สายอักขระ (String matching problem)

0	1	…	s	…	m	…	n-m+1	…	n
T[0]	T[1]	…	T[s]	…	T[s+m]	…	T[n-m]	…	T[n]
			P[0]	…	P[m-1]				

- P ปรากฏที่ตำแหน่ง s ใน T เมื่อ $T[s...s+m]=P[1...m-1]$ สำหรับ $0<=s<=n-m$

  หรือ $T[s+j]=P[j]$ สำหรับทุก $j=0,...,m-1$

ตัวอย่าง

- T = AGCATGCTGCAGTCATGCTTAGGCTA (|T|=26)
- P = GCT (|P|=3)
- P เกิดขึ้นกี่ครั้งใน T ในตำแหน่งใด?
  - T = AGCATGCTGCAGTCATGCTTAGGCTA
  - P เกิดขึ้น 3 ครั้งใน T ในตำแหน่งที่ 6,17,23

สัญลักษณ์

- $\sum$: เซตของอักขระ
- $\sum^*$: เซตของสายอักขระจำกัดทั้งหมดที่มาจาก $\sum$
- การเชื่อมกันของสายอักขระ x และสายอักขระ y เขียนแทนด้วย xy
- z เป็นสายอักขระแบบเติมหน้า (prefix) ของ x ถ้า x=zy ,$\exists y \in \sum^*$
- z เป็นสายอักขระแบบเติมท้าย (suffix) ของ x ถ้า x=yz ,$\exists y \in \sum^*$

String matching algorithm

- ขั้นตอนวิธีแบบตรง (Naïve algorithm or Brute-Force algorithm)
- ขั้นตอนวิธีของ Rabi-Karp (Rabi-Karp algorithm)
- ขั้นตอนวิธีของ Knuth-Morris-Pratt (Knuth-Morris-Pratt, KMP, algorithm)

Naïve algorithm

- Input : ข้อความ T และ แบบรูป P
- Output : ตำแหน่ง s ของ T ถ้าพบ แต่ถ้าไม่พบให้ค่า -1

```
Naïve-String-Matcher(T,P)
    n=length(T)
    m=length(P)
    For s=0 to n-m
        if P[0..m-1]=T[s..s+m] then
            Return s
        Endif
    Endfor
    Return -1
```

ตัวอย่าง

T : ababbaabaaab

P : abaa

• ababbaabaaab	: T		
• ababbaabaaab	step 1	aba#	mismatch: 4th letter
• ababbaabaaab	step 2	#...	mismatch: 1st letter
• ababbaabaaab	step 3	ab#.	mismatch: 3rd letter
• ababbaabaaab	step 4	#...	mismatch: 1st letter
• ababbaabaaab	step 5	#...	mismatch: 1st letter
• ababbaabaaab	step 6	a#..	mismatch: 2nd letter
• ababbaabaaab	step 7	abaa	success
• ababbaabaaab	step 8	#...	mismatch: 1st letter
• ababbaabaaab	step 9	.#..	mismatch: 2nd letter

Quiz

- จงเขียนโปรแกรมภาษา C/C++ แก้ปัญหา String matching โดยใช้ Naïve algorithm