CPCFI

Lecture: SUFFIX ARRAY AND SUFFIX TREE

Unit: 9 - STRING PROCESSING + COMPUTATIONAL GEOMETRY

Instructor: WILLSON

9.2.1 SUFFIX TRIE AND APPLICATIONS

↳ INFORMATION RETRIEVAL → PRONOUNCED "TRY"

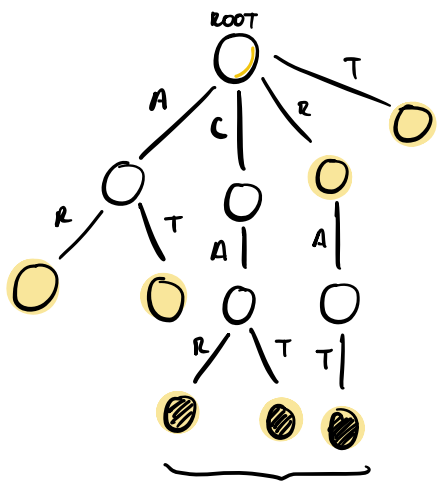
TRIE OF ALL POSSIBLE SUFFIXES OF A SET OF STRINGS S →
- EDGE: REPRESENTS A CHARACTER
- VERTEX: " " SUFFIX REPRESENTED BY ITS PATH LABEL

$S = \{$ "CAR", "CAT", "RAT" $\}$ → DICTIONARY

SUFFIXES OF $S = \{$ 'R', 'AR', 'CAR', 'T', 'AT', 'CAT', 'RAT' $\}$

SORTED SUFFIXES OF $S = \{$ 'AR', 'AT', 'CAR', 'CAT', 'R', 'RAT', 'T' $\}$

SUFFIX TRIE:



THIS PATHS APPEAR IN THE DICTIONARY

EACH VERTEX v CONTAINS TWO BOOLEAN FLAGS:
1. A SUFFIX ENDS AT VERTEX v
2. A WORD " " " "

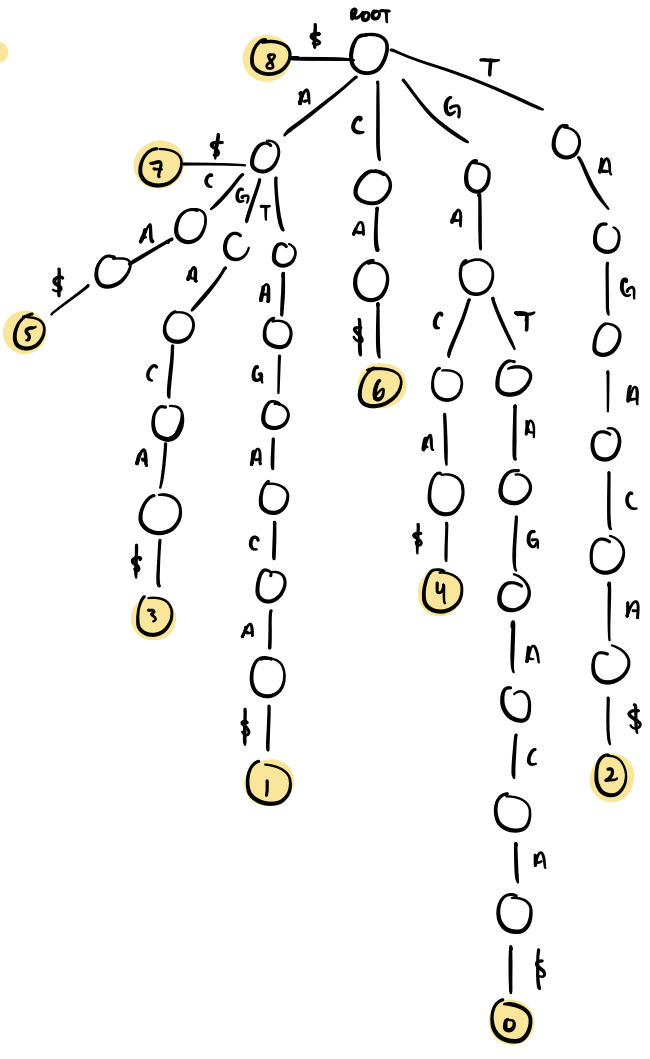
APPLICATIONS:

- EFFICIENT DATA STRUCTURE FOR DICTIONARY
- DETERMINE IF A QUERY/PATTERN P EXISTS IN THE DICTIONARY IN $O(|P|)$
  - ↳ TRAVERSE FROM THE ROOT DOWNWARDS

9.2.2 SUFFIX TRIE

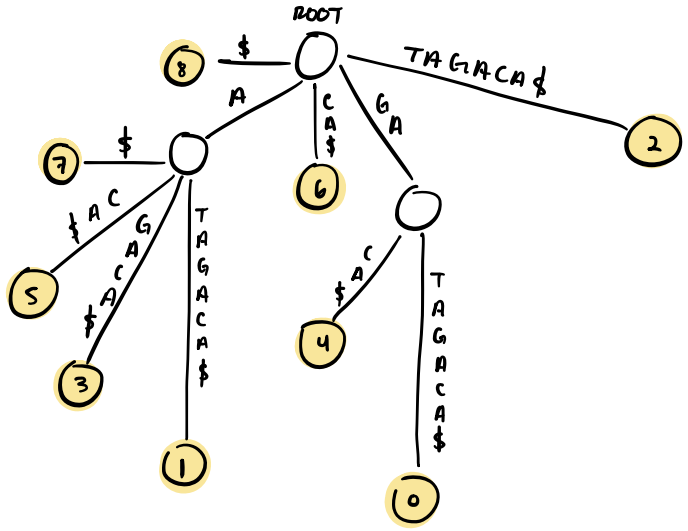
- ONE OF THE PROBLEMS OF A SUFFIX TRIE IS THE NUMBER OF REPEATED VERTICES FOR A GIVEN STRING → SUFFIX TREE SOLVES THIS PROBLEM
  - → T = "GATAGACA $"
    - → SPECIAL TERMINATING CHARACTER

i	SUFFIX
0	GATAGACA$
1	ATAGACA$
2	TAGACA$
3	AGACA$
4	GACA$
5	ACA$
6	CA$
7	A$
8	$



SUFFIX TRIE OF T

- SUFFIX TREE → SUFFIX TRIE WHERE WE MERGE VERTICES WITH ONLY ONE CHILD
  - → EDGE LABEL CAN HAVE MORE THAN ONE CHARACTER
  - → AT MOST 2^n VERTICES



SUFFIX TREE OF T

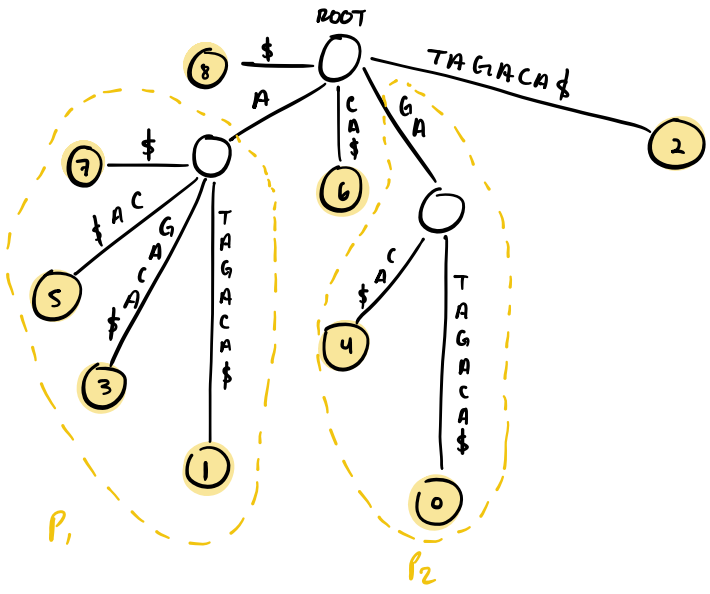
9.2.3 APPLICATIONS OF SUFFIX TREES (ASSUMING SUFFIX TREE IS ALREADY BUILT)

- [STRING MATCHING] IN $O(m + occ)$

  m: LENGTH OF PATTERN P
  occ: # OF OCCURRENCES

- A MATCH IS A COMMON PREFIX BETWEEN PATTERN STRING P AND SOME SUFFIXES OF STRING T

- FIND ALL OCCURRENCES OF STRING P IN STRING T



T = "GATAGACA$"
0 1 2 3 4 5 6 7 8

P_1 = 'A' → OCCURRENCES : {7, 5, 3, 1}

P_2 = "GA" → " : {4, 0}

P_3 = 'T' → " : {2}

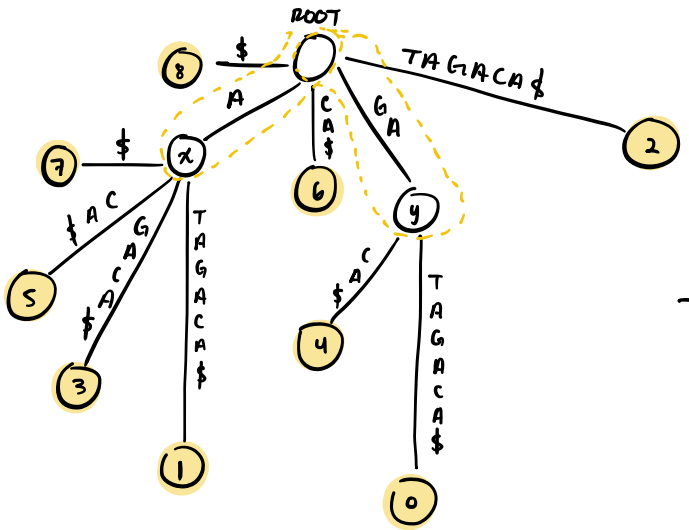
P_4 = 'Z' → " : { }

- [FINDING THE LONGEST REPEATED SUBSTRING] (LRS) IN $O(n)$

  - ANSWER: LENGTH OF PATH LABEL OF THE DEEPEST INTERNAL VERTEX x IN THE SUFFIX TREE

    ↳ CAN BE FOUND IN $O(n)$ USING TREE TRAVERSAL

  - INTERNAL VERTEX IN SUFFIX TREE IMPLIES THAT IT REPRESENTS MORE THAN ONE SUFFIXES OF T



→ INTERNAL VERTICES:

1) x: PATH LABEL LENGTH = 1 : A
2) y: " " " = 2 : GA

"GA" IS THE LRS

- [FINDING THE LONGEST COMMON SUBSTRINGS] IN $O(n)$

  - LCS OF TWO OR MORE STRINGS

  1) BUILD A SUFFIX TREE THAT COMBINES THE SUFFIX TREES OF TWO STRINGS T_1 AND T_2

     ↳ WE USE A TERMINATING SYMBOL FOR EACH T_1 AND T_2 ($, #)

  2) IDENTIFY INTERNAL VERTICES THAT HAVE VERTICES IN THEIR SUBTREES WITH DIFFERENT TERMINATING SYMBOLS

     ↳ COMMON PREFIXES BY STRINGS T_1 AND T_2

  3) ANSWER IS THE PATH LABEL WITH THE DEEPEST INTERNAL VERTEX

T_1 = "GATAGACA$" T_2 = "CATA#"

i	SUFFIX
0	GATAGACA$
1	ATAGACA$
2	TAGACA$
3	AGACA$
4	GACA$
5	ACA$
6	CA$
7	A$
8	$

i	SUFFIX
9	CATA#
10	ATA#
11	TA#
12	A#
13	#

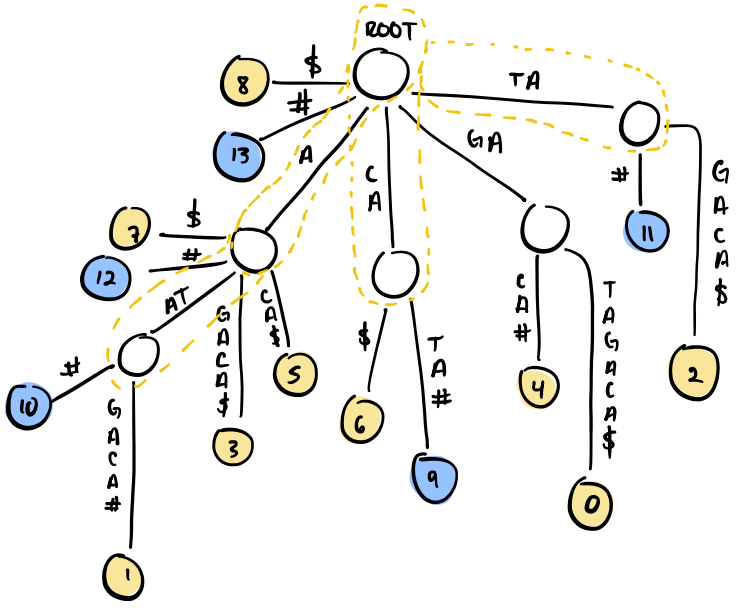
{A, C, G, T}

A: 1, 3, 5, 7, 10, 12

C: 6, 9

G: 0, 4

T: 2, 11



INTERNAL VERTICES TO CONSIDER:

1) PATH LABEL |A| = 1
2) " " |CA| = 2
3) " " |TA| = 2
4) " " |ATA| = 3

↓

ANSWER IS ATA

9.2.4 SUFFIX ARRAY

↳ MUCH SIMPLER TO BUILD THAN SUFFIX TREE $O(n \log n)$ CONST. TIME

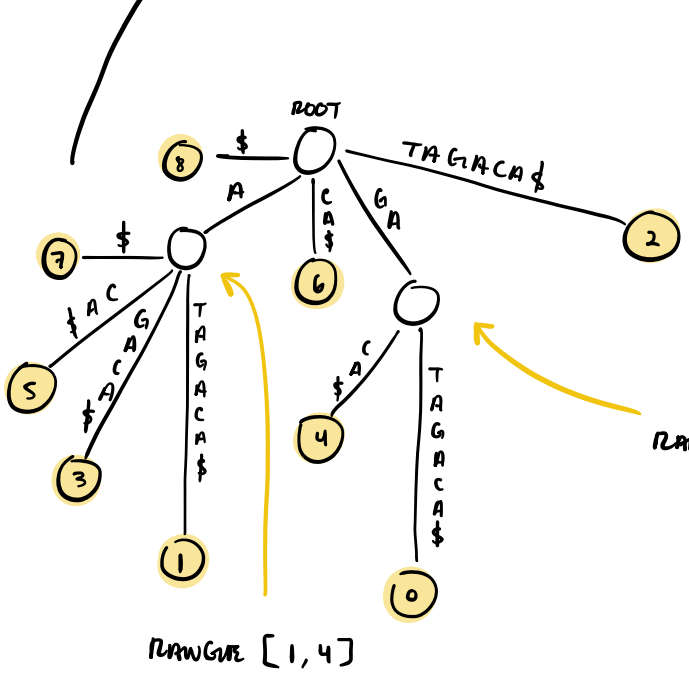
→ SUFFIX ARRAY: INTEGER ARRAY THAT STORES A PERMUTATION OF n INDICES OF SORTED SUFFIXES

T = "GATAGACA$"
n = 9

SA = {8, 7, 5, 3, 1, 6, 4, 0, 2}

i	SA[i]	SUFFIX
0	8	$
1	7	A$
2	5	ACA$
3	3	AGACA$
4	1	ATAGACA$
5	6	CA$
6	4	GACA$
7	0	GATAGACA$
8	2	TAGACA$

(1) THE TREE TRAVERSAL OF THE SUFFIX TREE VISITS THE LEAVES IN SUFFIX ARRAY ORDER



(2)

[AN INTERNAL VERTEX IN ST CORRESPONDS TO A RANGE IN SA]

↑

COLLECTION OF SORTED SUFFIXES THAT SHARE A COMMON PREFIX

(3)

[A TERMINATING VERTEX (LEAF) IN ST CORRESPONDS TO AN INDIVIDUAL INDEX IN THE SA]

9.2.5 APPLICATIONS OF SUFFIX ARRAY

- [STRING MATCHING] IN $O(m \log n)$

  m : LENGTH OF PATTERN P
  n : " " STRING T

  - PERFORM TWO $O(\log n)$ BINARY SEARCHES ON SORTED SUFFIXES AND UP TO $O(m)$ SUFFIX COMPARISONS

    ↳ TO FIND SMALLEST AND LARGEST i SUCH THAT THE PREFIX OF SA[i] MATCHES PATTERN P

  - ALL OF THE SUFFIXES BETWEEN LOWER AND UPPER BOUND ARE OCCURRENCES OF P IN T