ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ

Υβριδικές Δομές Δεδομένων

Κεφάλαιο 6

ΥΒΡΙΔΙΚΕΣ ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ

Συνδυάζουν τη χρήση **δεικτών** και **πινάκων**

- Ψηφιακά Δένδρα – TRIES
- Interpolation Search Tree

TRIE

Το ζητούμενο:

Αποθήκευση και ανάκτηση πληροφορίας κειμένου εύκολα.

- Λέξεις
- Συμβολοσειρές
- Επιθέματα κλπ.

TRIE

Το ζητούμενο:

Αποθήκευση και ανάκτηση πληροφορίας κειμένου εύκολα.

- Λέξεις
- Συμβολοσειρές
- Επιθέματα κλπ.

Ταίριασμα Προτύπου ή Συμβολοσειράς

Εφαρμογές σε αναζήτηση/επεξεργασία κειμένου, data mining, βιοπληροφορική κλπ.

TRIE

Το ζητούμενο:

Αποθήκευση και ανάκτηση πληροφορίας κειμένου εύκολα.

- Λέξεις
- Συμβολοσειρές
- Επιθέματα κλπ.

Ταίριασμα Προτύπου ή Συμβολοσειράς

Εφαρμογές σε αναζήτηση/επεξεργασία κειμένου, data mining, βιοπληροφορική κλπ.

Μια απλή λύση είναι τα ψηφιακά δέντρα (TRIEs)

ΛΟΓΙΚΗ ΤΟΥ TRIE

$S = \{x_1, \ldots, x_n\}$

- Θέλουμε να αναπαραστήσω το S σε μια δομή
- Δεν στηριζόμαστε στις τιμές x_i
- Χρησιμοποιούμε **αναπαράσταση των στοιχείων** σαν μια ακολουθία χαρακτήρων

ΛΟΓΙΚΗ ΤΟΥ TRIE

$S = \{x_1, \ldots, x_n\}$

- Θέλουμε να αναπαραστήσω το S σε μια δομή
- Δεν στηριζόμαστε στις τιμές x_i
- Χρησιμοποιούμε **αναπαράσταση των στοιχείων** σαν μια ακολουθία χαρακτήρων

Μοιάζει με ΛΕΞΙΚΟ!!

- Οι λέξεις βρίσκονται ανάλογα με το γράμμα με το οποίο αρχίζουν
- Το ίδιο για το 2ο, 3ο, 4ο κλπ χαρακτήρα
- Είναι **φυλλοπροσανατολισμένο**!

ΟΡΙΣΜΟΣ

“Έστω σύμπαν U του οποίου τα στοιχεία είναι συμβολοσειρές μήκους λ πάνω σε ένα αλφάβητο K με |K| = k. Ένα σύνολο $S \subseteq U$ αναπαρίσταται σαν ένα k-δικο δένδρο που περιέχει όλα τα προθέματα των στοιχείων του S”

ΟΡΙΣΜΟΣ

“Έστω σύμπαν U του οποίου τα στοιχεία είναι συμβολοσειρές μήκους λ πάνω σε ένα αλφάβητο K με |K| = k. Ένα σύνολο $S \subseteq U$ αναπαρίσταται σαν ένα k-δικο δένδρο που περιέχει όλα τα προθέματα των στοιχείων του S”

Τι είναι πρόθεμα;

ΥΛΟΠΟΙΗΣΗ ΤΟΥ TRIE

1. Κάθε εσωτερικός κόμβος του δέντρου είναι ένας πίνακας μήκους k από δείκτες.

ΥΛΟΠΟΙΗΣΗ ΤΟΥ TRIE

1. Κάθε εσωτερικός κόμβος του δέντρου είναι ένας πίνακας μήκους k από δείκτες.
2. Κάθε θέση του πίνακα αντιστοιχίζεται σε ένα γράμμα του αλφαβήτου.

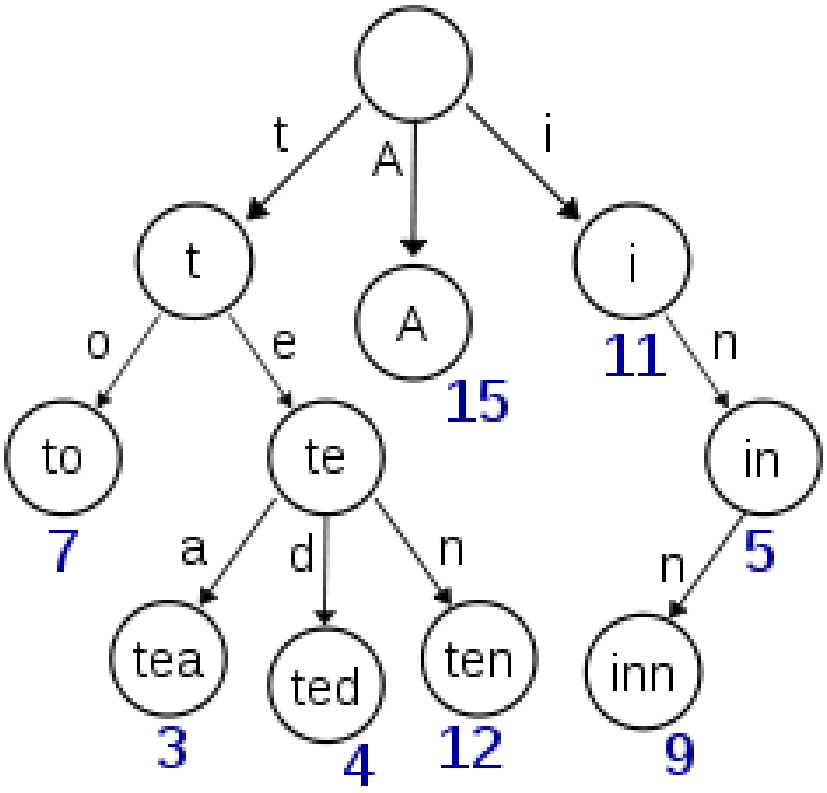
ΥΛΟΠΟΙΗΣΗ ΤΟΥ TRIE

1. Κάθε εσωτερικός κόμβος του δέντρου είναι ένας πίνακας μήκους k από δείκτες.
2. Κάθε θέση του πίνακα αντιστοιχίζεται σε ένα γράμμα του αλφαβήτου.
3. Κάθε θέση του πίνακα σε ένα κόμβο u σε βάθος i θα λάβει τιμή αν κάποιο από τα στοιχεία του S στην i-οστη θέση έχει τον αντίστοιχο χαρακτήρα

ΥΛΟΠΟΙΗΣΗ ΤΟΥ TRIE

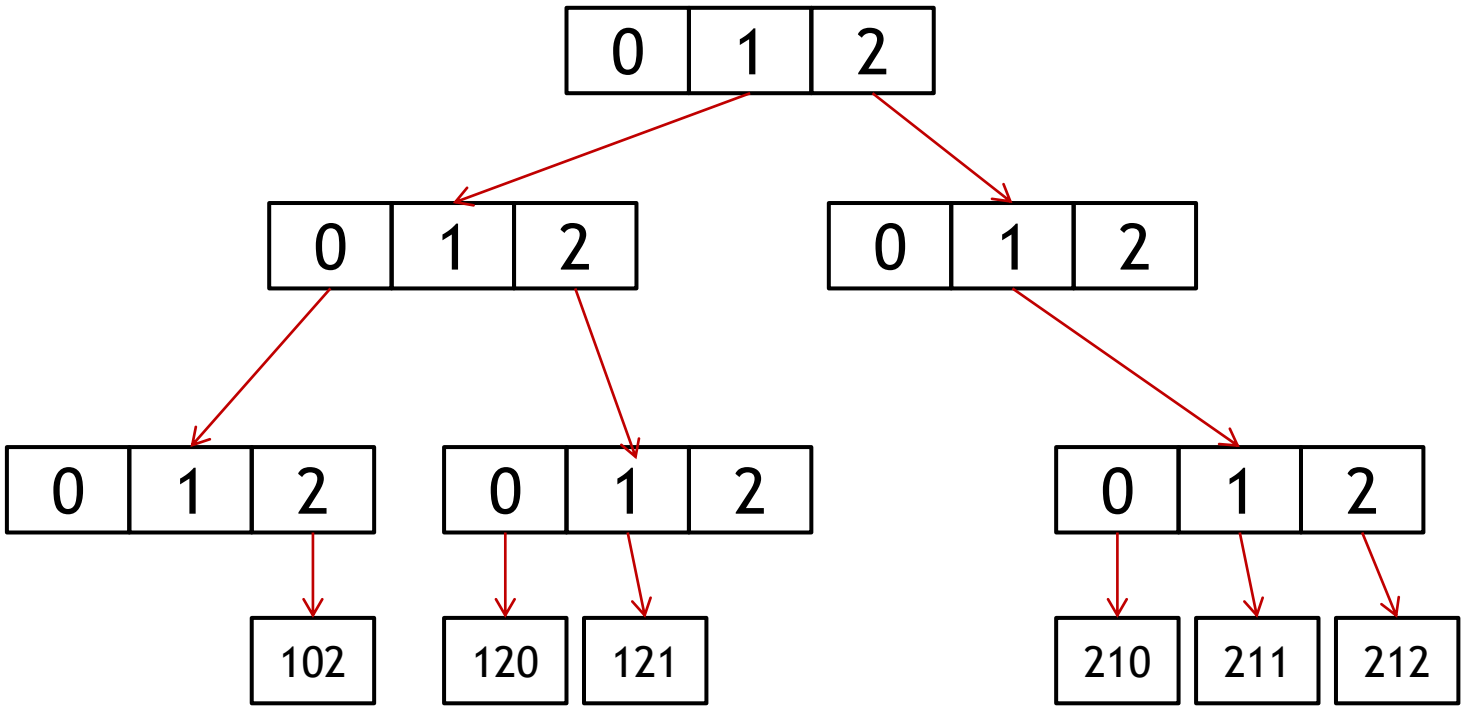
1. Κάθε εσωτερικός κόμβος του δέντρου είναι ένας πίνακας μήκους k από δείκτες.
2. Κάθε θέση του πίνακα αντιστοιχίζεται σε ένα γράμμα του αλφαβήτου.
3. Κάθε θέση του πίνακα σε ένα κόμβο u σε βάθος i θα λάβει τιμή αν κάποιο από τα στοιχεία του S στην i-οστη θέση έχει τον αντίστοιχο χαρακτήρα
4. Ύψος λ!
5. Χώρος O(k)

1ο ΠΑΡΑΔΕΙΓΜΑ



2ο ΠΑΡΑΔΕΙΓΜΑ

S = {102, 120, 121, 210, 211, 212}



ΠΟΛΥΠΛΟΚΟΤΗΤΕΣ

1. Οι **βασικές πράξεις** χρειάζονται χρόνο Ο(λ)

ΠΟΛΥΠΛΟΚΟΤΗΤΕΣ

1. Οι **<u>βασικές πράξεις</u>** χρειάζονται χρόνο O(λ)
2. Εξαρτάται από το μέγεθος του αλφάβητου και του σύμπαντος U: $O(\lambda) = O(\log_k N)$

ΠΟΛΥΠΛΟΚΟΤΗΤΕΣ

1. Οι **βασικές πράξεις** χρειάζονται χρόνο O(λ)
2. Εξαρτάται από το μέγεθος του αλφάβητου και του σύμπαντος U: $O(\lambda) = O(\log_k N)$
3. Ο χώρος που χρησιμοποιεί ένα TRIE όταν αναπαριστά |S| = n στοιχεία είναι O(nλk) στην χειρότερη περίπτωση. **Ποια είναι αυτή η περίπτωση;;**

ΠΟΛΥΠΛΟΚΟΤΗΤΕΣ

1. Οι **<u>βασικές πράξεις</u>** χρειάζονται χρόνο O(λ)
2. Εξαρτάται από το μέγεθος του αλφάβητου και του σύμπαντος U: $O(\lambda) = O(\log_k N)$
3. Ο χώρος που χρησιμοποιεί ένα TRIE όταν αναπαριστά |S| = n στοιχεία είναι O(nλk) στην χειρότερη περίπτωση. **<u>Ποια είναι αυτή η περίπτωση;;</u>**

Όταν τα στοιχεία δεν έχουν κανένα κοινό πρόθεμα. Οπότε:

- ***n* πλήρη μονοπάτια**
- **λ* *n* κόμβους συνολικά**
- **K * λ * *n* συνολικός χώρος**

ΛΥΣΗ – ΣΥΜΠΑΓΕΣ TRIE

Αποθηκεύονται ΜΟΝΟ οι κόμβοι του TRIE με βαθμό μεγαλύτερο ή ίσο του 2, ενώ οι αλυσίδες των κόμβων με βαθμό 1 αντικαθίστανται από έναν απλό αριθμό.

ΛΥΣΗ – ΣΥΜΠΑΓΕΣ TRIE

Αποθηκεύονται ΜΟΝΟ οι κόμβοι του TRIE με βαθμό μεγαλύτερο ή ίσο του 2, ενώ οι αλυσίδες των κόμβων με βαθμό 1 αντικαθίστανται από έναν απλό αριθμό.

Ο απλός αριθμός αποθηκεύεται στην (πρώτη) πλευρά που οδηγεί στην αλυσίδα και είναι ίσος με το πλήθος των κόμβων σε αυτή.

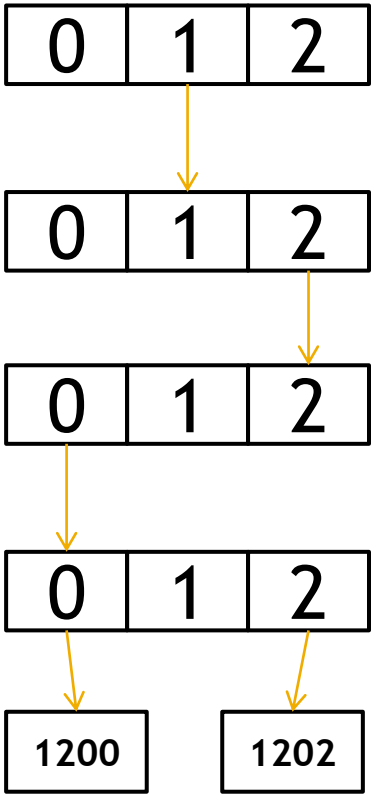
ΛΥΣΗ – ΣΥΜΠΑΓΕΣ TRIE

Αποθηκεύονται ΜΟΝΟ οι κόμβοι του TRIE με βαθμό μεγαλύτερο ή ίσο του 2, ενώ οι αλυσίδες των κόμβων με βαθμό 1 αντικαθίστανται από έναν απλό αριθμό.

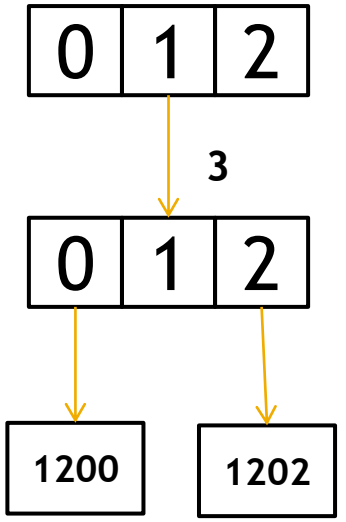
Ο απλός αριθμός αποθηκεύεται στην (πρώτη) πλευρά που οδηγεί στην αλυσίδα και είναι ίσος με το πλήθος των κόμβων σε αυτή.

Ο χώρος από O(nλk) γίνεται O(nk)

ΠΑΡΑΔΕΙΓΜΑ



TRIE



ΣΥΜΠΑΓΕΣ TRIE

ΔΕΝΔΡΟ ΕΠΙΘΕΜΑΤΩΝ – Suffix Tree

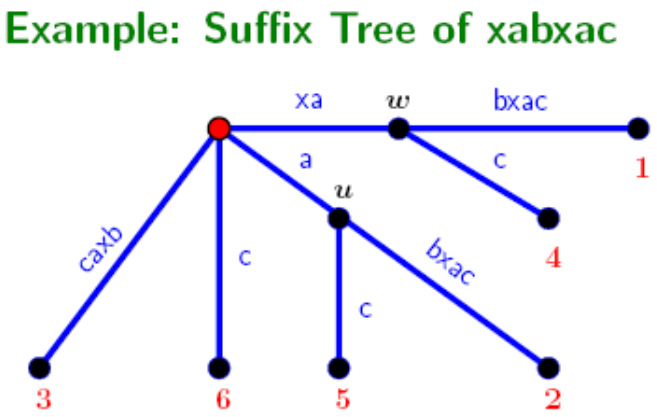
Ορισμός

Αποθηκεύει όλα τα δυνατά επιθέματα μιας συμβολοσειράς S.

ΔΕΝΔΡΟ ΕΠΙΘΕΜΑΤΩΝ – Suffix Tree

Ορισμός

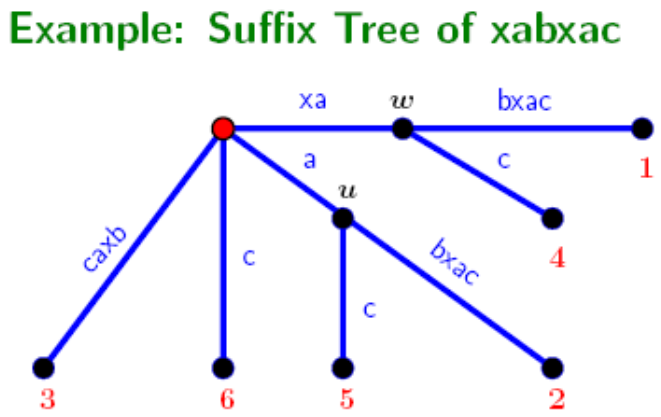
Αποθηκεύει όλα τα δυνατά επιθέματα μιας συμβολοσειράς S.



ΔΕΝΔΡΟ ΕΠΙΘΕΜΑΤΩΝ – Suffix Tree

Ορισμός

Αποθηκεύει όλα τα δυνατά επιθέματα μιας συμβολοσειράς S.



Το suffix tree μιας συμβολοσειράς S[1…n] είναι ένα *συμπαγές* **TRIE,** που περιέχει ως κλειδιά όλα τα επιθέματα S[i…n], 1≤i≤n.

ΚΑΤΑΣΚΕΥΗ Suffix Tree

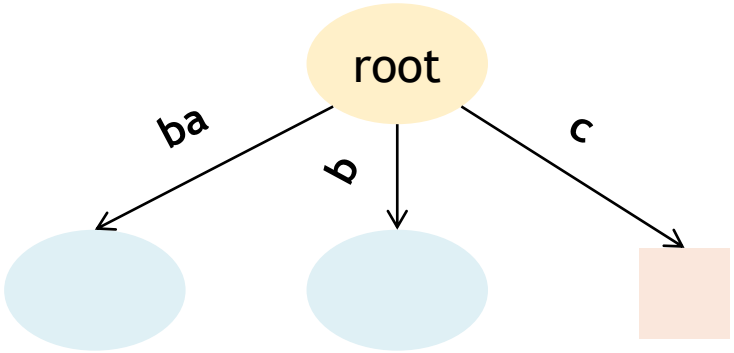
Έστω συμβολοσειρά X = ababc

ΚΑΤΑΣΚΕΥΗ Suffix Tree

Έστω συμβολοσειρά X = ababc

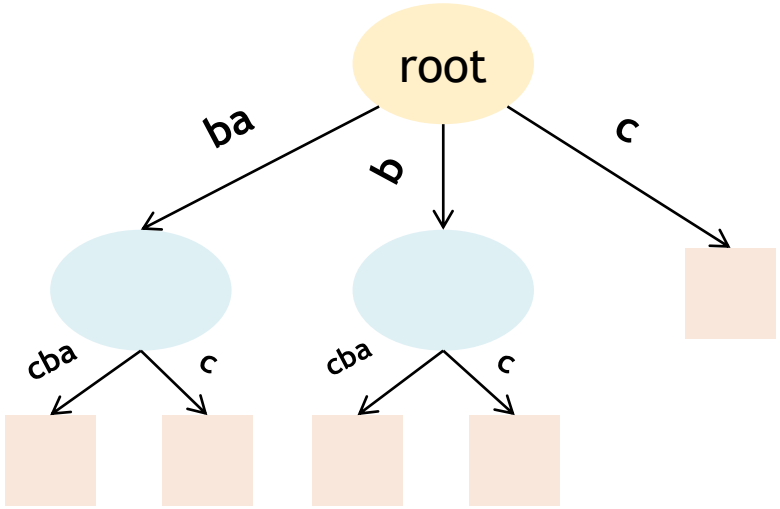
ΚΑΤΑΣΚΕΥΗ Suffix Tree

Έστω συμβολοσειρά X = ababc



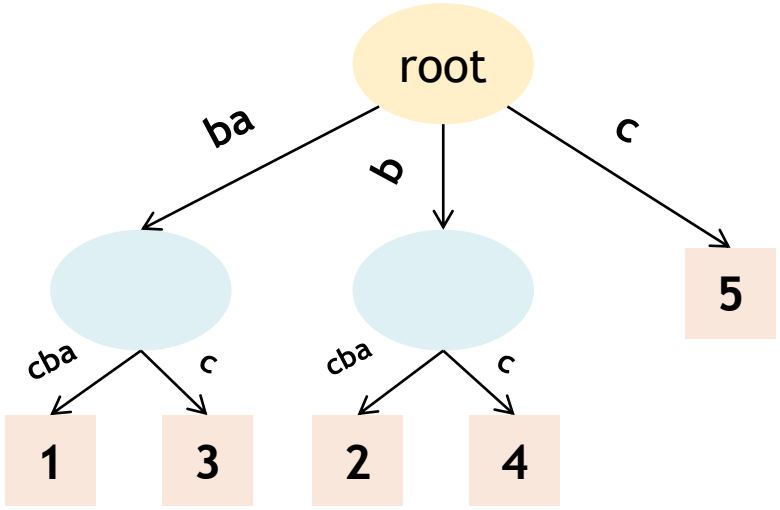
ΚΑΤΑΣΚΕΥΗ Suffix Tree

Έστω συμβολοσειρά X = ababc



ΚΑΤΑΣΚΕΥΗ Suffix Tree

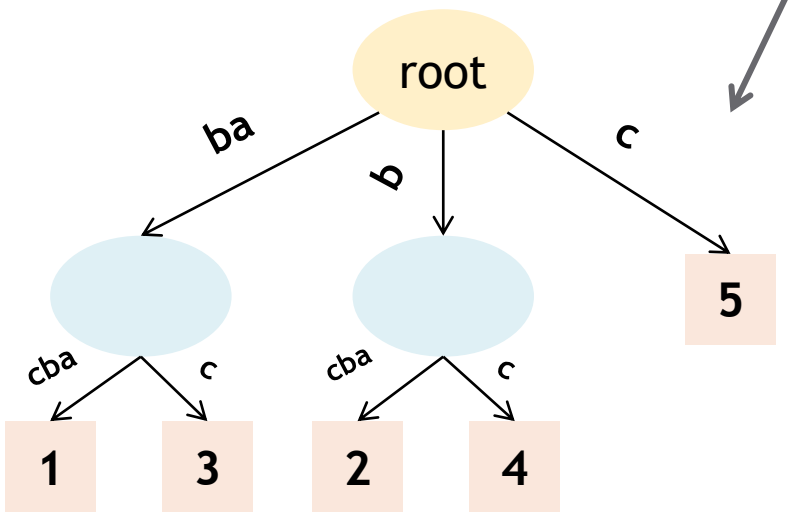
Έστω συμβολοσειρά X = ababc



ΚΑΤΑΣΚΕΥΗ Suffix Tree

Έστω συμβολοσειρά X = ababc

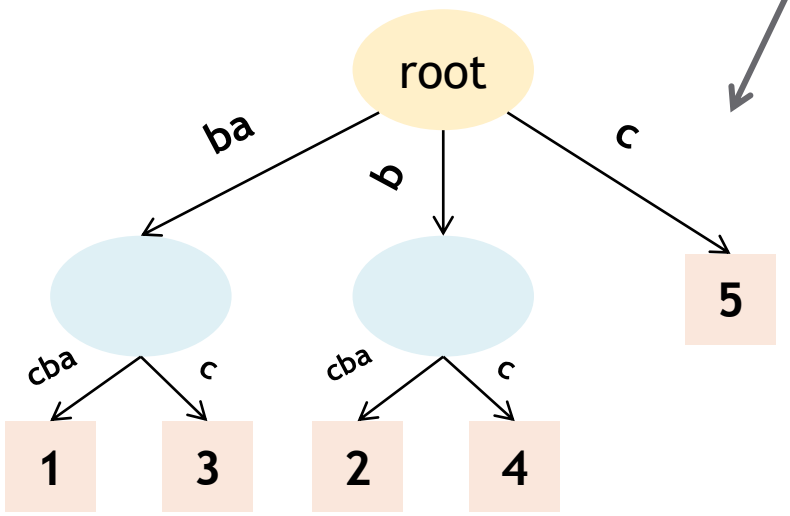
ΠΡΟΣΟΧΗ!
Από τον ίδιο κόμβο δεν εξέρχονται υπο-συμβολοσειρές με κοινό πρώτο χαρακτήρα



ΚΑΤΑΣΚΕΥΗ Suffix Tree

Έστω συμβολοσειρά X = ababc

ΠΡΟΣΟΧΗ!
Από τον ίδιο κόμβο δεν εξέρχονται υπο-συμβολοσειρές με κοινό πρώτο χαρακτήρα



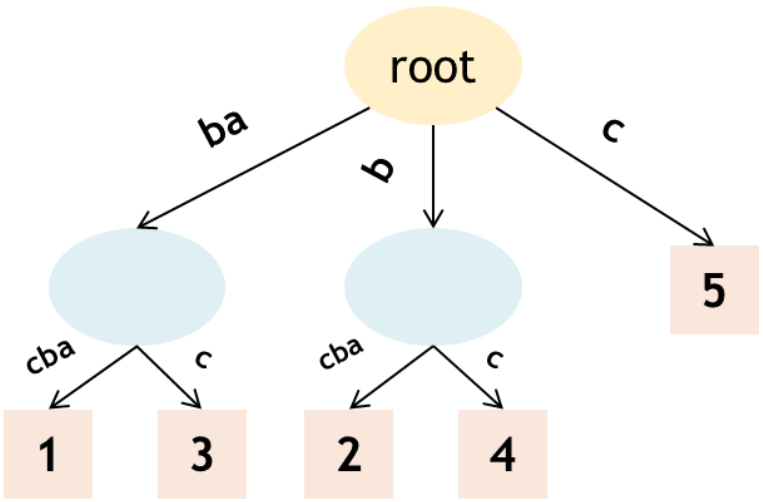
Γιατί αριθμήσαμε με αυτή τη σειρά τα φύλλα;

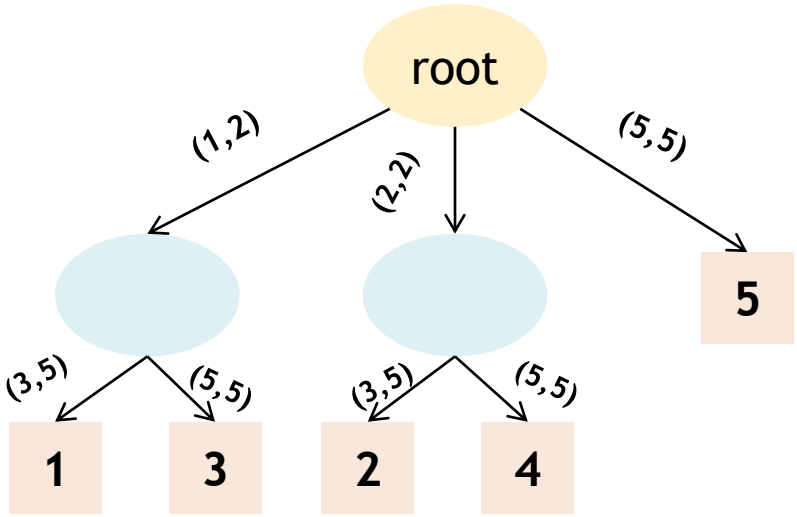
ΚΑΤΑΣΚΕΥΗ Suffix Tree – Αλλιώς

Έστω συμβολοσειρά X = ababc

ΚΑΤΑΣΚΕΥΗ Suffix Tree - Αλλιώς

Έστω συμβολοσειρά X = ababc





Εφαρμογές Suffix Tree

1. Ταίριασμα Προτύπου – Pattern matching
2. Μέγιστη Επαναλαμβανόμενη Υποσυμβολοσειρά – Longest Repeated Substring
3. Μέγιστη Κοινή Υποσυμβολοσειρά – Longest Common Substring