ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Προεπεξεργασία

Ακαδημαϊκό Έτος 2023-2024

Προεπεξεργασία/Κατασκευή του ευρετηρίου

Επεξεργασία εγγράφου
- λεξικό/λεξιλόγιο

Κατασκευή ευρετηρίου
- ανεστραμμένο ευρετήριο

Κανονική Λειτουργία

Επεξεργασία Ερωτημάτων

Ενημέρωση ευρετηρίου

Τα βασικά βήματα για την κατασκευή του ευρετηρίου

1. Συλλέγουμε τα έγγραφα που θέλουμε να συμπεριλάβουμε στο ευρετήριο

2. Διαιρούμε το κείμενο σε γλωσσικά σύμβολα **(token)**

3. Γλωσσολογική προ-επεξεργασία των συμβόλων

4. Ευρετηριάζουμε τα έγγραφα στα οποία περιλαμβάνεται κάθε όρος

Friends, Romans, countrymen.

Tokenizer

Token stream. Friends Romans Countrymen

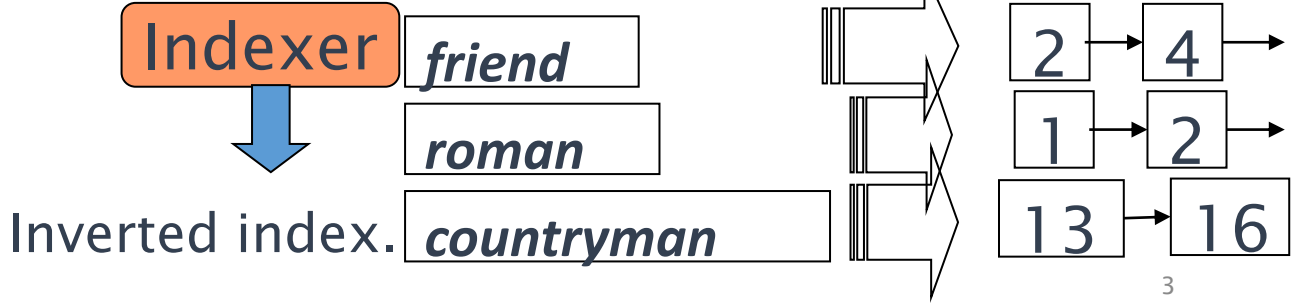
Linguistic modules

friend roman countryman

Terms

Indexer

Inverted index.



Ακολουθία εγγράφων

Περίληψη

Έγγραφο 1 (d1) : Το Παν. Ιωαννίνων ιδρύθηκε το 1970.

Έγγραφο 2 (d2) : Τα Ιωάννινα είναι η μεγαλύτερη πόλη της Ηπείρου.

Έγγραφο 3 (d3) : Η πτυχιακή εξεταστική στο Τμήμα Μηχ. Η/Υ και Πληροφορικής θ΄ αρχίσει την 1η Φεβρουαρίου.

Έγγραφο 4 (d4) : Οι μαθητές των Ιωαννίνων αρίστευσαν στις εξετάσεις για την εισαγωγή στα Πανεπιστήμια.

Έγγραφο 5 (d5): Το 2017 ιδρύθηκε Πολυτεχνική Σχολή στο ΠΙ.

Granularity: Μονάδα εγγράφου

↓

Token (λεκτικές μονάδες)

Θέματα

- Που σταματάμε: κενό/σημείο στίξης αλλά και απόστροφοι/όχι κενό/παύλα, κλπ
- Stop words (το, και?)
- Κανονικοποίηση
  - Κεφαλαία/μικρά
  - Τόνοι
  - Κανόνες vs Λίστες ισοδυναμίας

↓

Όροι (terms) που θα εισαχθούν στο ευρετήριο

- **Περιστολή (stemming)** περικοπή καταλήξεων
- **Λημματοποίηση (lemmatization)** γλωσσική/μορφολογική επεξεργασία και αναγωγή της λέξης στη ρίζα της

✓ Ίδια πολιτική και στο κείμενο και στην ερώτηση

Επανάληψη (ερωτήσεις)

Άσκηση 2.1

Are the following statements true or false?

1. In a Boolean retrieval system, stemming never lowers precision.

1Α Σωστό 1Β Λάθος

2. In a Boolean retrieval system, stemming never lowers recall.

2Α Σωστό 2Β Λάθος

Επανάληψη (ερωτήσεις)

Άσκηση 2.1

Are the following statements true or false?

3. Stemming increases the size of the vocabulary

3A Σωστό 3B Λάθος

4. Stemming should be invoked at indexing time but not while processing a query.

4A Σωστό 4B Λάθος

Βήματα του Indexer: Ακολουθία Token

- Ακολουθία από ζεύγη (Modified token, Document ID).

Doc 1

I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Doc 2

So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

Βήματα του Indexer: Ταξινόμηση (sort)

- Ταξινόμηση με βάση τους όρους
  - Και μετά το docID

Βασικό βήμα της ευρετηριοποίησης

Term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

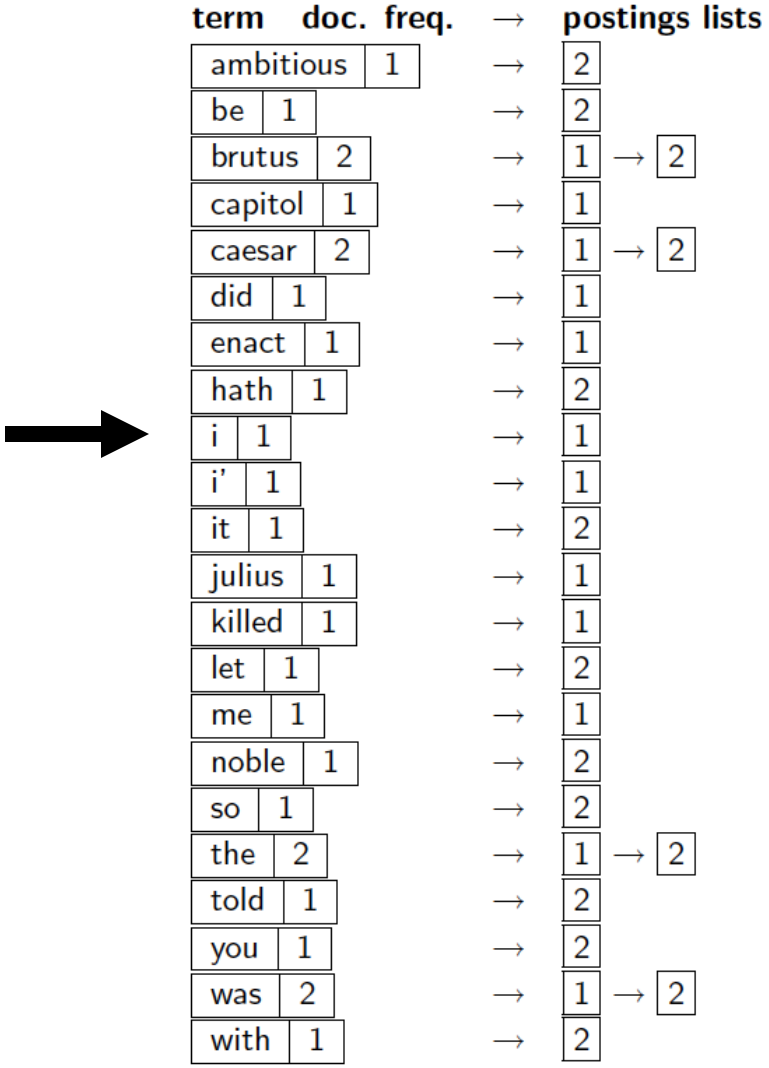
Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

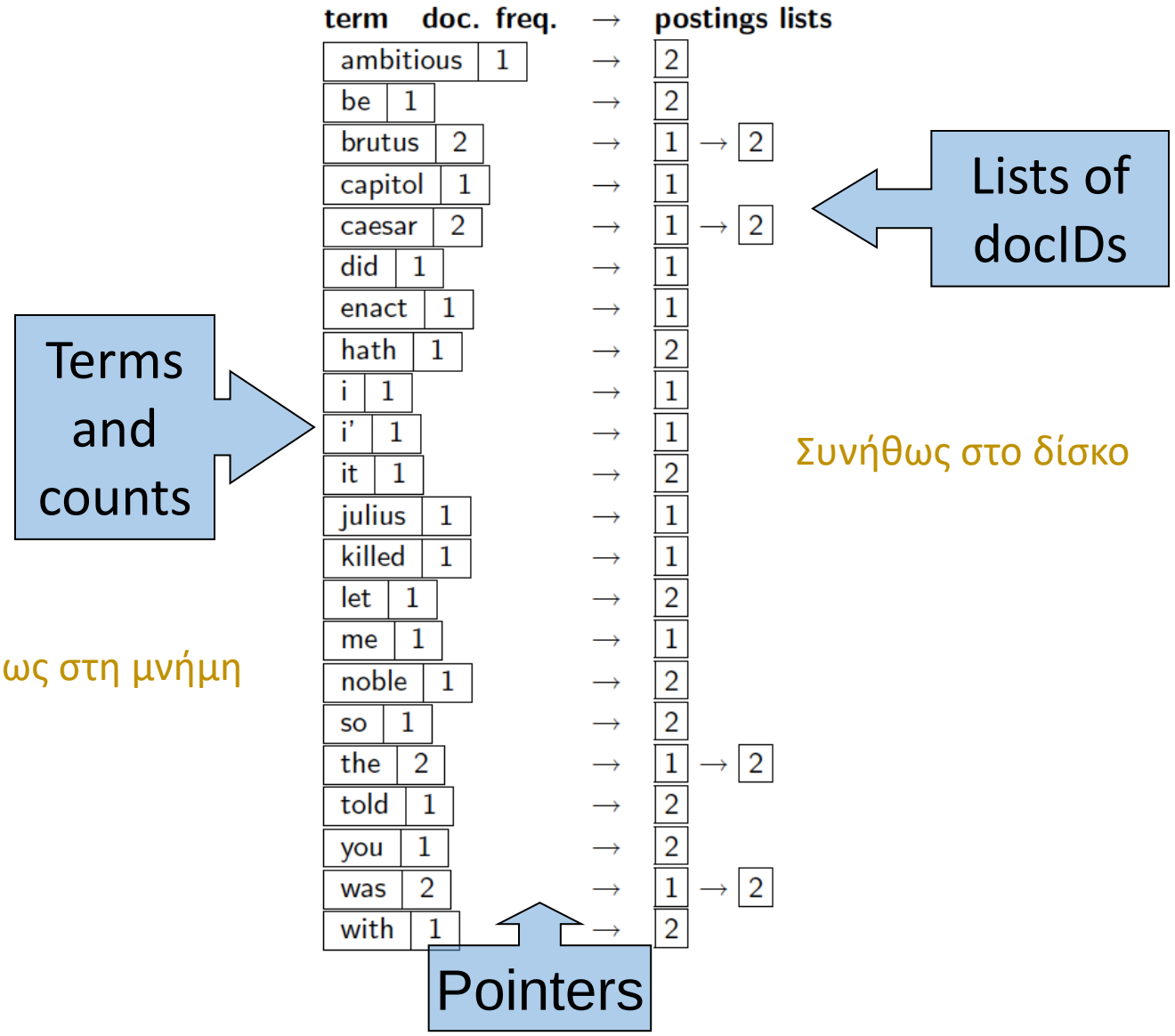
Βήματα του Indexer: Λεξικό & Καταχωρήσεις

- Πολλαπλές εμφανίσεις του όρου σε ένα έγγραφο συγχωνεύονται (merged).
- Διαχωρισμός σε *λεξικό* και *καταχωρήσεις*
- Προσθέτουμε και πληροφορία για τη συχνότητα εγγράφου (doc. frequency).

Γιατί τη συχνότητα;
Επίσης, συχνότητα όρου (term frequency)

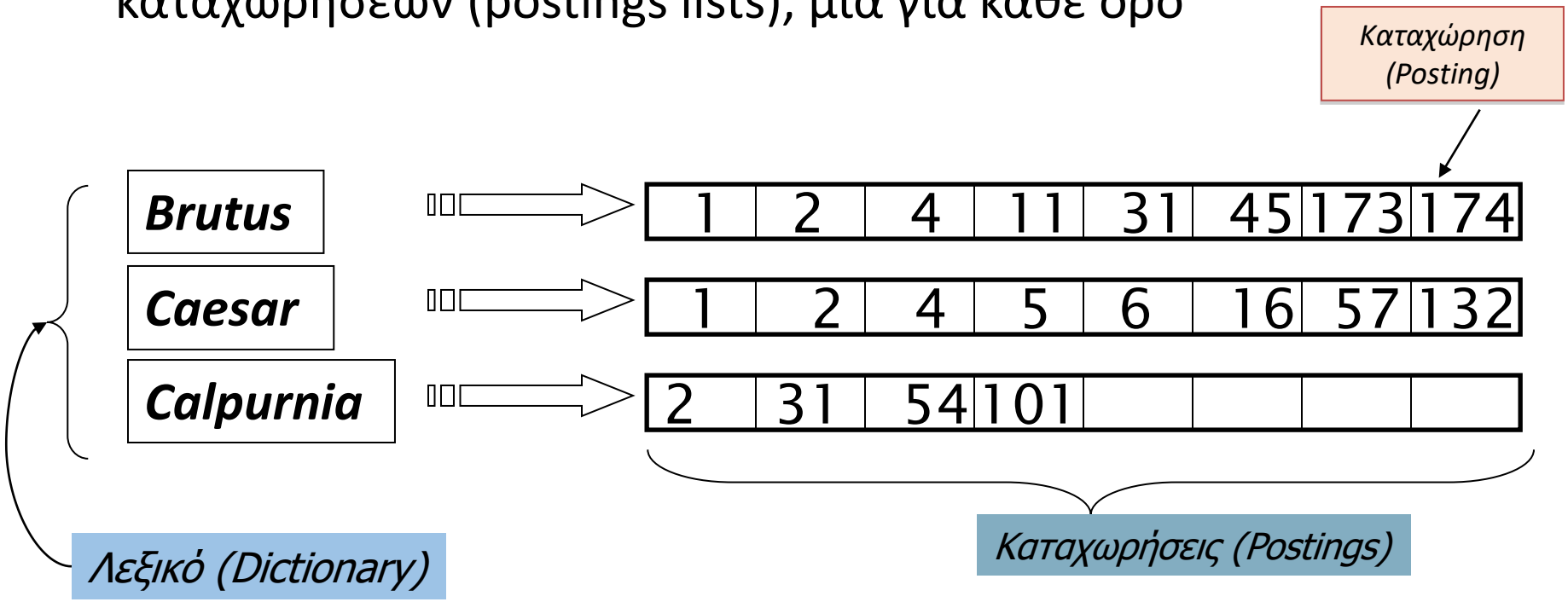
Term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2




term doc. freq. → postings lists
ambitious 1 → 2
be 1 → 2
brutus 2 → 1 → 2
capitol 1 → 1
caesar 2 → 1 → 2
did 1 → 1
enact 1 → 1
hath 1 → 2
i 1 → 1
i' 1 → 1
it 1 → 2
julius 1 → 1
killed 1 → 1
let 1 → 2
me 1 → 1
noble 1 → 2
so 1 → 2
the 2 → 1 → 2
told 1 → 2
you 1 → 2
was 2 → 1 → 2
with 1 → 2
Terms and counts
Lists of docIDs
Συνήθως στο δίσκο
Συνήθως στη μνήμη
Pointers

Βασικές Δομές

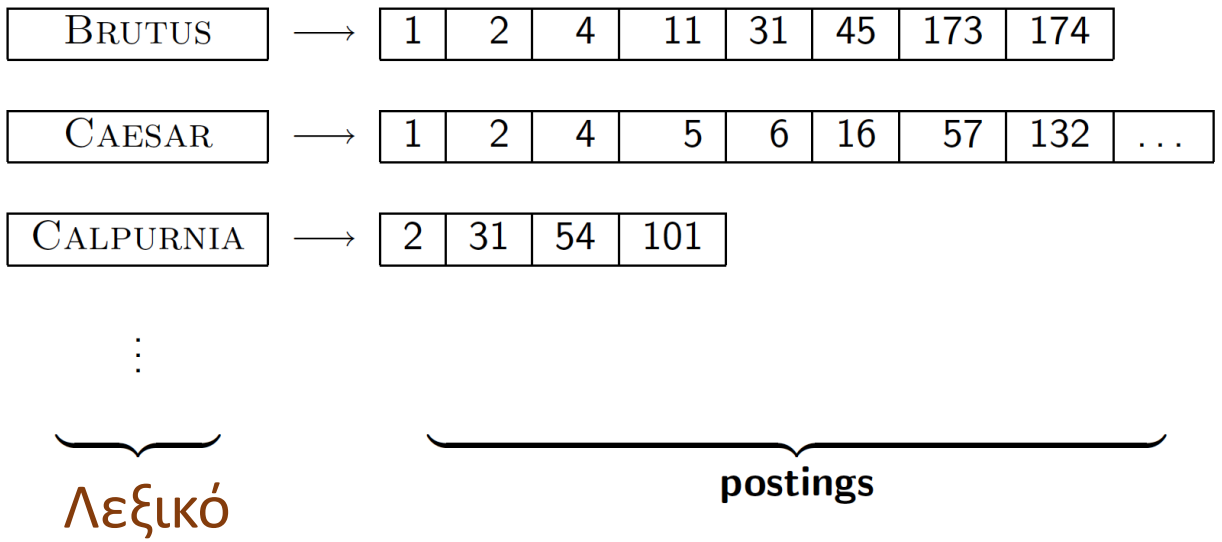
- Λεξικό
- Ανεστραμμένο ευρετήριο: αποτελείται από λίστες καταχωρήσεων (postings lists), μία για κάθε όρο



Σε διάταξη με βάση το docID

Δομές Δεδομένων για Λεξικά

- Το *λεξικό* περιέχει: το *λεξιλόγιο όρων* -- για κάθε όρο: τη συχνότητα εγγράφου (document frequency) και δείκτη στη λίστα καταχωρήσεων
- Σε κάθε ερώτημα, *αναζήτηση στο λεξικό* αν υπάρχει ο όρος και σε ποιες λίστες



Ποια δομή δεδομένων είναι κατάλληλη;

Δομές Δεδομένων για Λεξικά

Λεξιλόγιο (vocabulary): το σύνολο των όρων

Λεξικό (dictionary): μια δομή για την αποθήκευση του λεξιλογίου

Πως αποθηκεύουμε ένα λεξικό (στη μνήμη) αποδοτικά;

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Δομές για Λεξικά

Ακαδημαϊκό Έτος 2022-2023

Μια απλοϊκή λύση

- array of struct:

term	document frequency	pointer to postings list
a	656,265	⟶
aachen	65	⟶
. . .	. . .	. . .
zulu	221	⟶

char[20] int Postings *

20 bytes 4/8 bytes 4/8 bytes

- Πως αναζητούμε έναν όρο (κλειδί, key) στο λεξικό γρήγορα κατά την εκτέλεση του ερωτήματος;

Δομές Δεδομένων για Λεξικά

Κριτήρια για την επιλογή δομής:

(ποιες λειτουργίες, workload, μέγεθος)

- Αποδοτική αναζήτηση ενός όρου (κλειδιού) στο λεξικό.
- Είναι στατικό ή έχουμε συχνά εισαγωγές/διαγραφές όρων ή και τροποποιήσεις; Μόνο εισαγωγές (insert only – append only)
- Πόσοι είναι οι όροι
- Σχετικές συχνότητες προσπέλασης των κλειδιών (πιο γρήγορα οι συχνοί όροι;)

Δομές Δεδομένων για το Λεξικό

- Δυο βασικές επιλογές:
  - Πίνακες Κατακερματισμού (Hashtables)
  - Δέντρα (Trees)

- Μερικά συστήματα ανάκτησης πληροφορίας χρησιμοποιούν πίνακες κατακερματισμού άλλα δέντρα

Πίνακες Κατακερματισμού

Κάθε όρος του λεξιλογίου κατακερματίζεται σε έναν ακέραιο

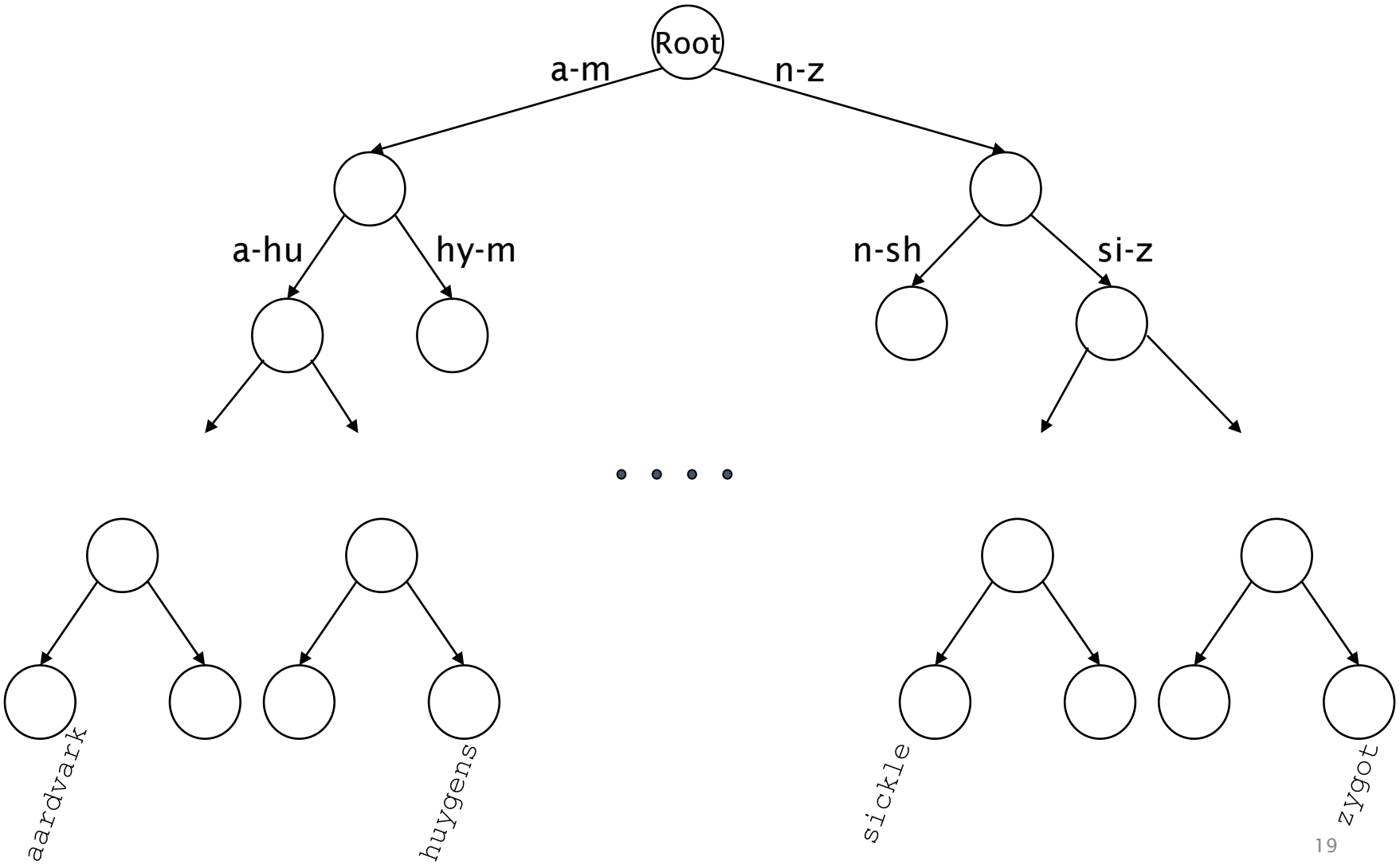
+:

- Η αναζήτηση είναι πιο γρήγορη από ένα δέντρο: O(1)

- :

- Δεν υπάρχει εύκολος τρόπος να βρεθούν μικρές παραλλαγές ενός όρου
  - judgment/judgement, *resume* vs. *résumé*
- Μη δυνατή η προθεματική αναζήτηση [ανεκτική ανάκληση]
- Αν το λεξιλόγιο μεγαλώνει συνεχώς, ανάγκη για να γίνει κατακερματισμός από την αρχή

Δέντρα Αναζήτησης: Δυαδικό Δέντρο

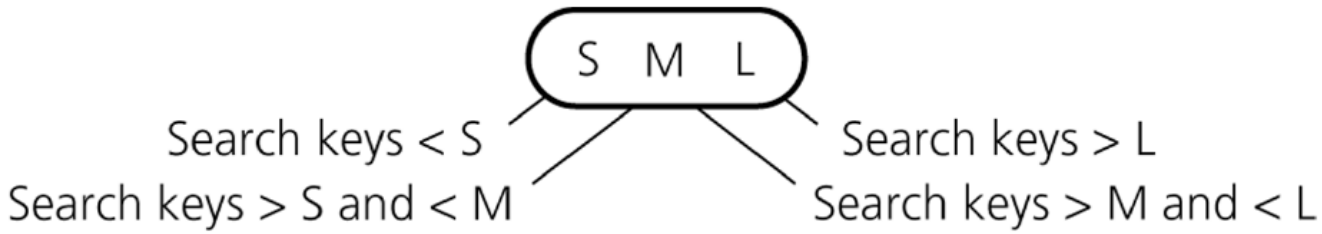


Δέντρα Αναζήτησης: Δυαδικό Δέντρο

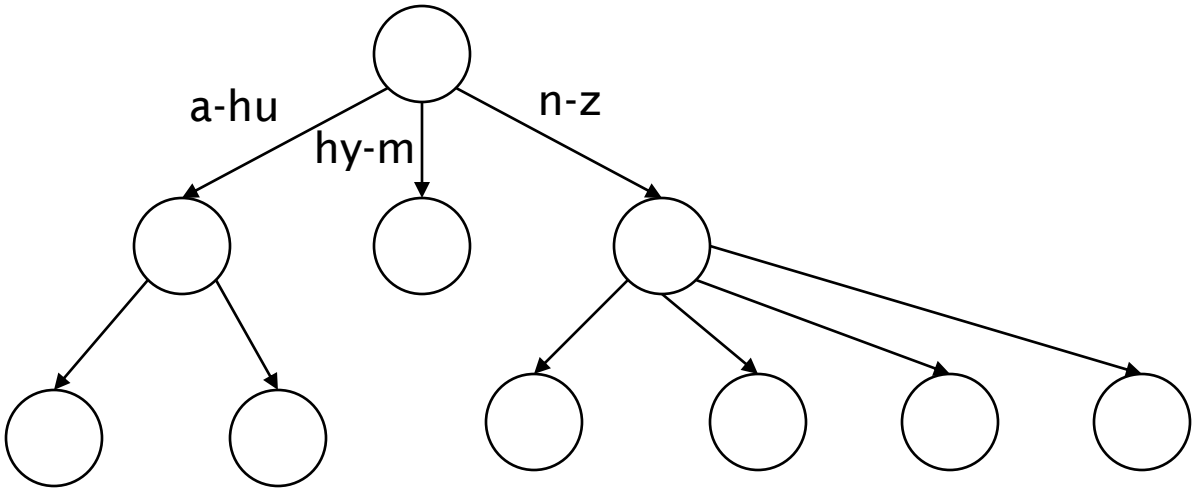
- O(log M), M: αριθμός των όρων (το μέγεθος του λεξικού)
  - προϋποθέτει ισοζύγιση

Δέντρα: B-δέντρα

Ορισμός: Κάθε εσωτερικός κόμβος έχει έναν αριθμό από παιδιά στο διάστημα [*a*, *b*] όπου *a, b* είναι κατάλληλοι φυσικοί αριθμοί, π.χ., [2, 4]



Δέντρα: B-δέντρα



Ορισμός: Κάθε εσωτερικός κόμβος έχει έναν αριθμό από παιδιά στο διάστημα [*a*, *b*] όπου *a, b* είναι κατάλληλοι φυσικοί αριθμοί, π.χ., [2, 4]

Δέντρα

- Το απλούστερο: δυαδικό δέντρο
- Το πιο συνηθισμένο: B-δέντρα
- Τα δέντρα απαιτούν ένα δεδομένο τρόπο διάταξης των χαρακτήρων (αλλά συνήθως υπάρχει ή μπορεί να οριστεί)

+:

- Λύνουν το πρόβλημα προθέματος (π.χ., όροι που αρχίζουν με *hyp*)
- Πλεονεκτούν όταν το λεξικό αποθηκεύεται στο δίσκο (τότε τα *a* και *b* καθορίζονται από το μέγεθος του block)

-:

- Πιο αργή: O(log *M*) [και αυτό απαιτεί (*ισοζυγισμένα* (*balanced)* δέντρα)
- Η επανα-ισοζύγιση (rebalancing) των δυαδικών δέντρων είναι ακριβή
  - Αλλά τα B-δέντρα καλύτερα

ΜΥΕ003: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Λίστες Καταχωρήσεων: Επεκτάσεις

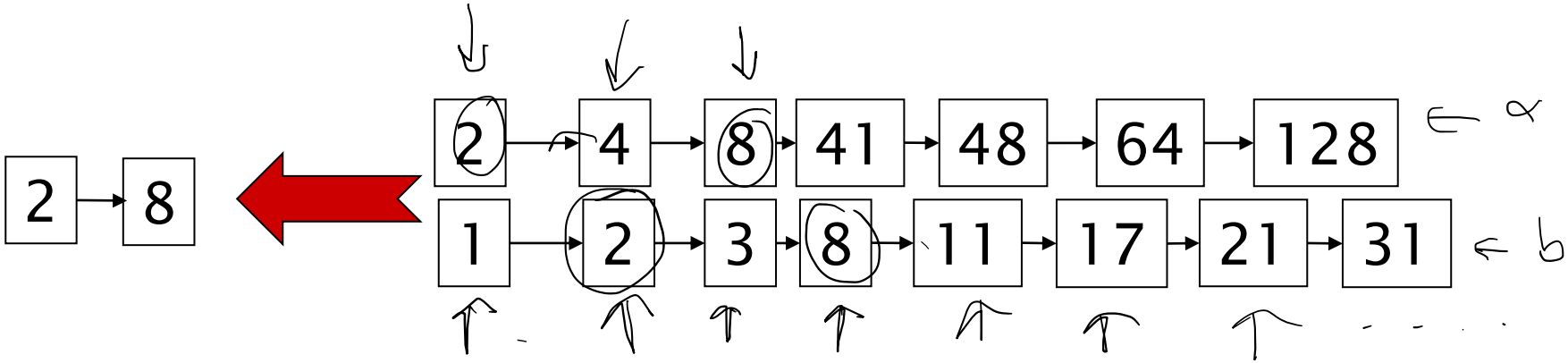
Ακαδημαϊκό Έτος 2023-2024

Τι θα δούμε σήμερα;

Ανεστραμμένο ευρετήριο

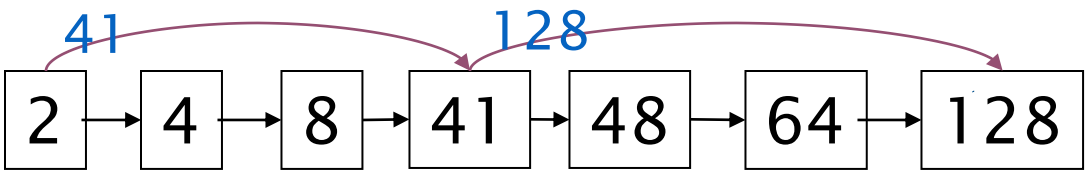
- Γρηγορότερη συγχώνευση: *Λίστες Παράβλεψης* (skip lists)
- Λίστες καταχωρήσεων *με πληροφορίες θέσεων* (*positional postings*) και ερωτήματα φράσεων (*phrase queries*)

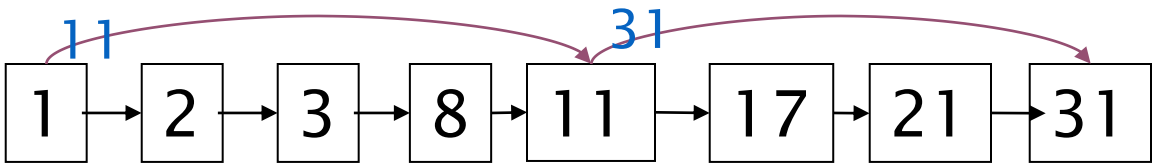
Βασική συγχώνευση



Αν τα μήκη των λιστών είναι m και n, O(*m+n*)

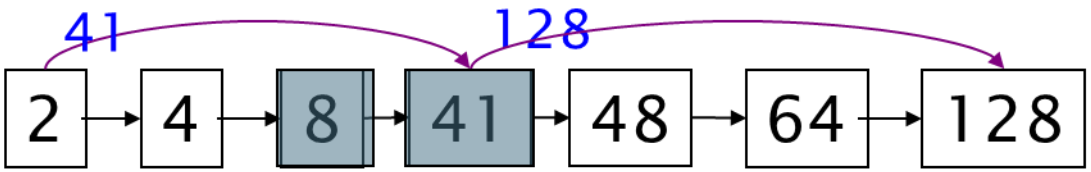
Επέκταση των λιστών με δείκτες παράβλεψης skip pointers (κατά την κατασκευή του ευρετηρίου)

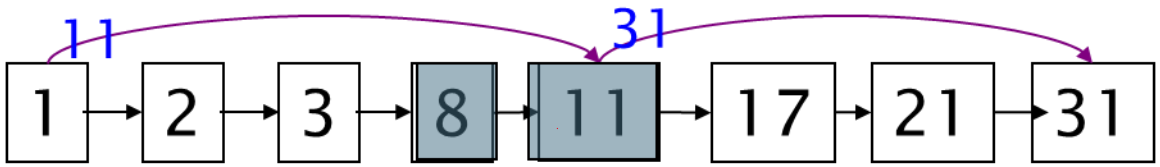




- Γιατί?
  - *Για να αποφύγουμε (skip) καταχωρήσεις που δεν θα εμφανιστούν στο αποτέλεσμα της αναζήτησης.*
- Πως?
- Που να τοποθετήσουμε αυτούς τους δείκτες?

Επεξεργασία ερωτήματος με skip pointers



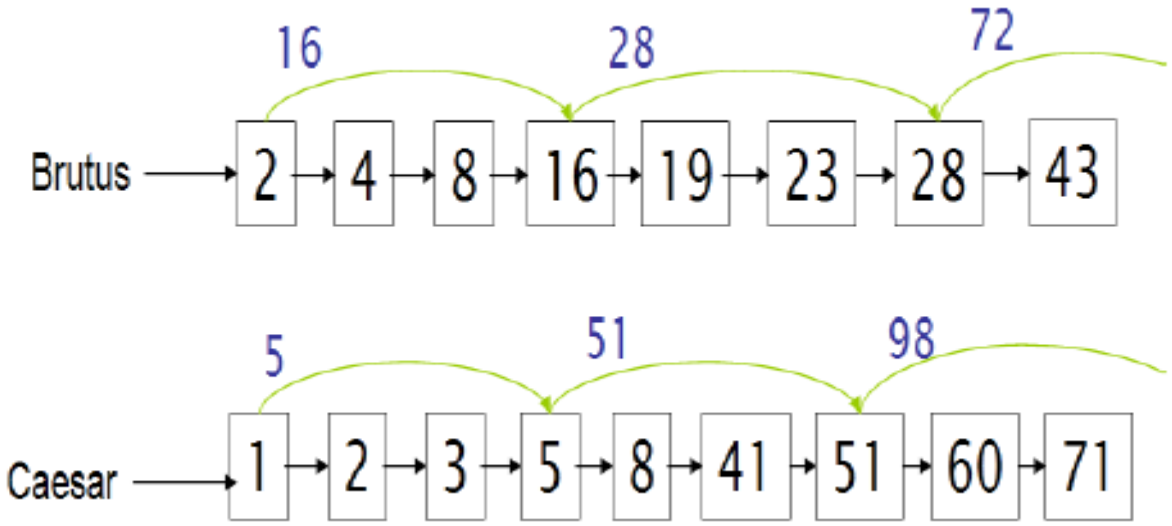


Υποθέστε ότι έχουμε διατρέξει τις λίστες και έχουμε βρει το κοινό στοιχείο **8** σε κάθε λίστα, το ταιριάζουμε και προχωράμε

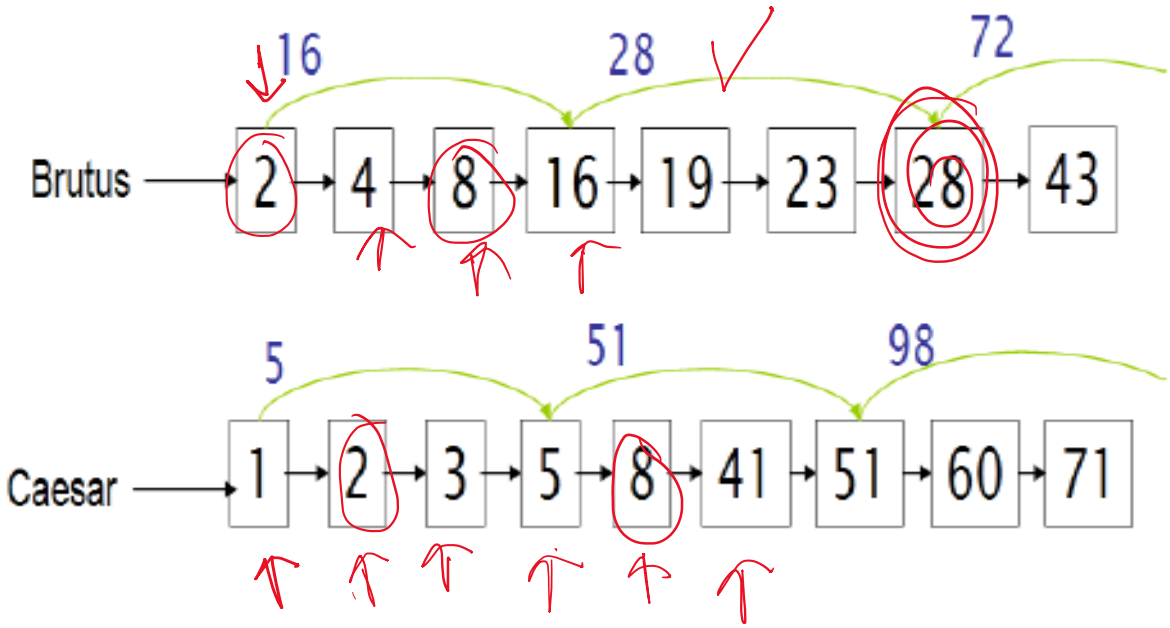
Έχουμε **41** και **11**. Το **11** είναι το μικρότερο.

Ο δείκτης παράλειψης του **11** είναι το **31**, οπότε μπορούμε να παραβλέψουμε τις ενδιάμεσες καταχωρήσεις

Επεξεργασία ερωτήματος με skip pointers

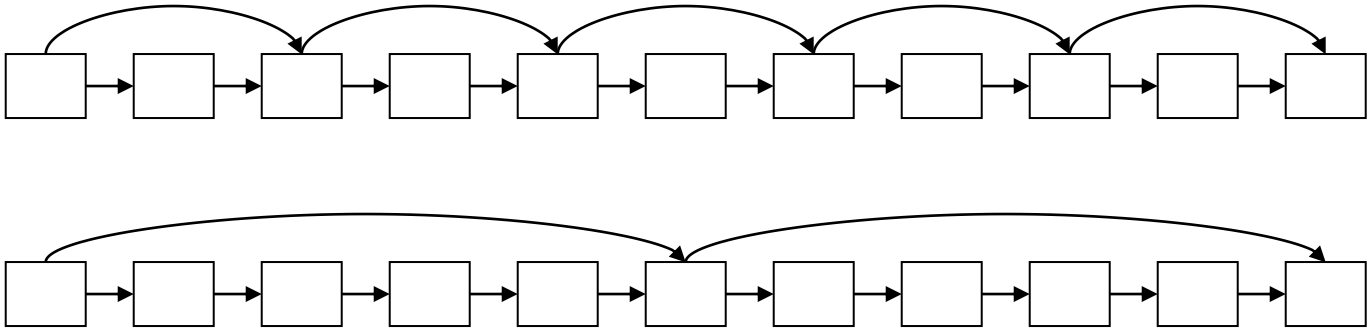


Αριθμός συγκρίσεων χωρίς και με χρήση δεικτών παράβλεψης


16
28
72
Brutus
2
4
8
16
19
23
28
43
5
51
98
Caesar
1
2
3
5
8
41
51
60
71

Που να τοποθετήσουμε τους δείκτες?

- Tradeoff:
  - Πολλοί δείκτες παράβλεψης $\rightarrow$ μικρότερα διαστήματα παράβλεψης $\Rightarrow$ μεγαλύτερη πιθανότητα παράβλεψης. Πολλές συγκρίσεις για να παραλείψουμε δείκτες.
  - Λιγότεροι δείκτες παράβλεψης $\rightarrow$ λιγότερες συγκρίσεις δεικτών αλλά μεγαλύτερα διαστήματα $\Rightarrow$ λίγες επιτυχημένες παραβλέψεις.



Τοποθέτηση των δεικτών

Απλώς ευριστικός: για καταχωρήσεις μήκους L, χρησιμοποίησε $\sqrt{L}$ *δείκτες* παράβλεψης σε ίδια απόσταση μεταξύ τους (evenly spaced), δηλαδή σε *απόσταση* $\sqrt{L}$

- Αγνοεί την κατανομή των όρων της ερώτησης.
- Εύκολο αν το ευρετήριο είναι σχετικά στατικό. Δύσκολο αν το L αλλάζει συνεχώς λόγω τροποποιήσεων.
- Βοηθάει αν *memory-based*
  - Το I/O κόστος για να φορτωθεί μια μεγαλύτερη (λόγω skip pointers) λίστα καταχωρήσεων μπορεί να υπερβαίνει το κέρδος από τη γρηγορότερη συγχώνευση

Ευρετήρια φράσεων

Ερωτήματα Φράσεων (phrase queries)

- Θέλουμε να μπορούμε να απαντάμε σε ερωτήματα όπως “***stanford university***” – ως φράση
- Οπότε το έγγραφο *“I went to university at Stanford”* δεν αποτελεί ταίριασμα.
  - Η έννοια των ερωτημάτων φράσεων έχει αποδειχθεί πολύ δημοφιλής και εύκολα κατανοητή από τους χρήστες,
  - Από τις λίγες μορφές αναζήτησης πέρα της βασικής που υιοθετήθηκαν (ερωτήσεις με «» αποτελούν το 10%)
  - Ακόμα περισσότερες είναι *έμμεσα* ερωτήματα φράσεων
- Για να τα υποστηρίξουμε, δεν αρκούν εγγραφές της μορφής

  <*term* : *docs*>

Μια πρώτη προσέγγιση: Ευρετήρια ζευγών λέξεων (Biword indexes)

- Εισήγαγε στο ευρετήριο *κάθε διαδοχικό ζεύγος όρων* στο κείμενο ως φράση
- Για παράδειγμα το κείμενο “Friends, Romans, Countrymen” παράγει τα biwords
  - ***friends romans***
  - ***romans countrymen***
- Κάθε τέτοιο biword είναι τώρα ένας όρος του ευρετηρίου
- Επιτρέπει την επεξεργασία ερωτημάτων φράσεων με δύο λέξεις.

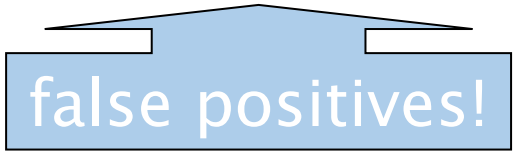
Μεγαλύτερες φράσεις

- Οι μεγαλύτερες φράσεις με κατάτμηση:

stanford university palo alto μπορεί να διασπαστεί ως ένα Boolean ερώτημα με biwords:

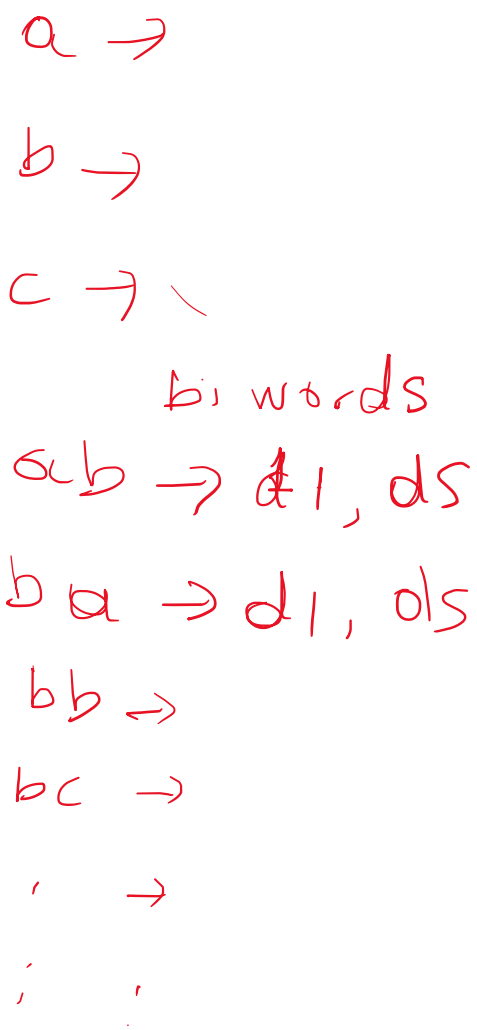
stanford university *AND* ***university palo*** *AND* ***palo alto***

Χωρίς να εξετάσουμε τα έγγραφα, δεν μπορούμε να εξακριβώσουμε ότι τα έγγραφα που ικανοποιούν το παραπάνω ερώτημα περιέχουν τη φράση.



Παράδειγμα

d1: a b a
d2: b b c a
d3: b c d c
d4: a c d b
d5: a c b a b



Διευρυμένα biwords

- Επεξεργασία του κειμένου και εκτέλεση part-of-speech-tagging (POST).
- Ομαδοποιούμε τους όρους (έστω) σε ουσιαστικά- Nouns (N) και άρθρα/προθέσεις (X).

- Διευρυμένο biword: κάθε ακολουθία όρων της μορφής NX*N
  - Κάθε τέτοιο διευρυμένο biword είναι τώρα ένας όρος του λεξικού

Διευρυμένα biwords

- Παράδειγμα: ***catcher in the rye***

  N X X N

- Επεξεργασία ερωτήματος: χώρισε το σε N και X
  - Διαίρεσε την ερώτηση σε διευρυμένα biwords
  - Αναζήτησε στο ευρετήριο το: ***catcher rye***

- Παράδειγμα: ***cost overruns on a power plant***
  - “cost overruns” “overruns power” “power plant”

Θέματα

- False positives
- Περισσότερους από 2 όρους -> Phrase index (ευρετήριο φράσης)
- Δημιουργούνται πολύ μεγάλα λεξικά
  - Δεν είναι δυνατόν για μεγαλύτερες φράσεις από 2 λέξεις, μεγάλα ακόμα και για αυτές
- Τα ευρετήρια biword δεν είναι η συνήθης λύση (για όλα τα biwords) αλλά χρησιμοποιούνται *ως μέρος πιο σύνθετων λύσεων*

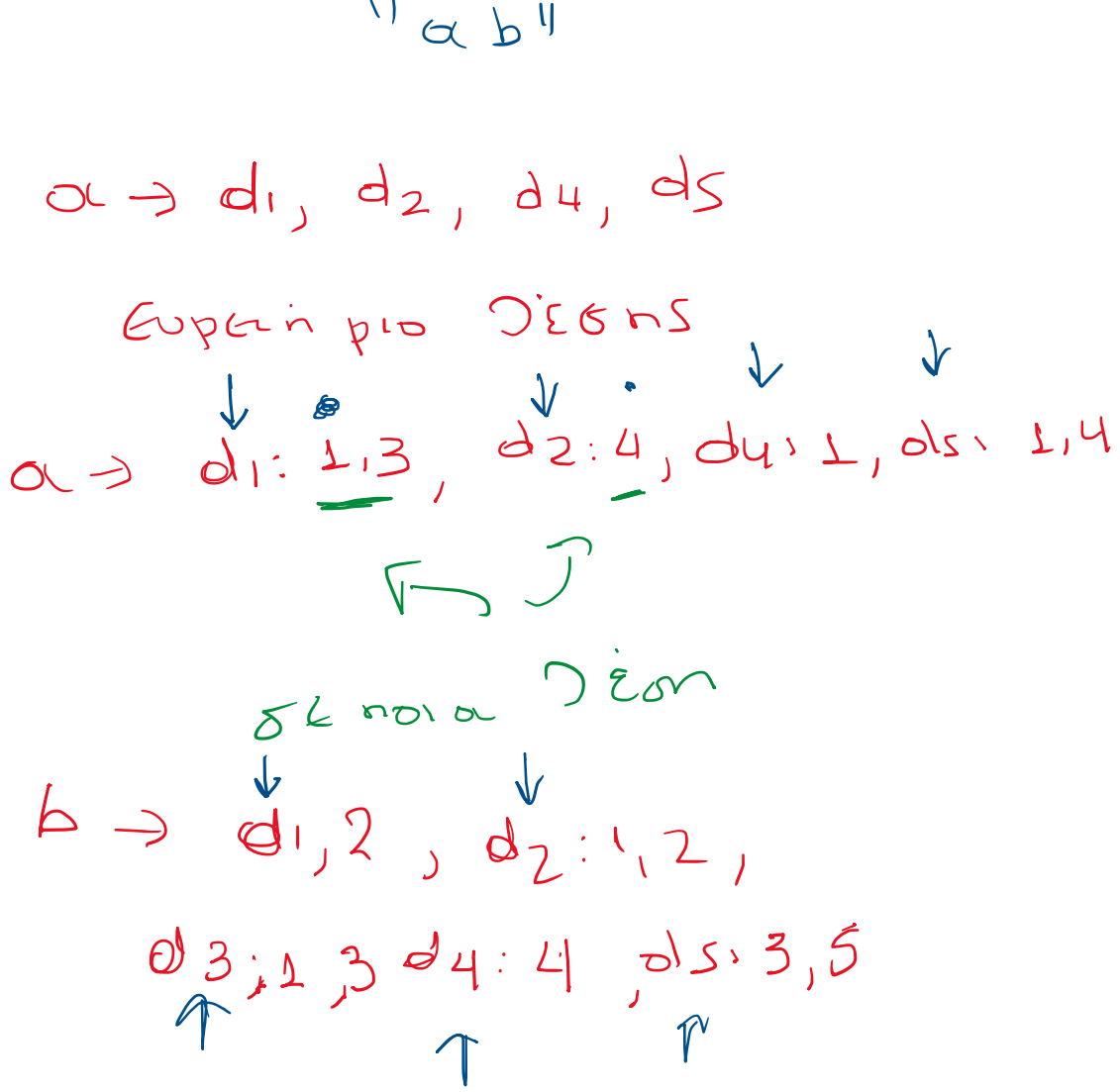
Λύση 2: Positional indexes (Ευρετήρια Θέσεων)

- Στις καταχωρήσεις, με κάθε όρο, αποθηκεύουμε και τη θέση (θέσεις) όπου εμφανίζονται τα tokens του:

<***term***, number of docs containing ***term***;
doc1: position1, position2 … ;
doc2: position1, position2 … ;
etc.>

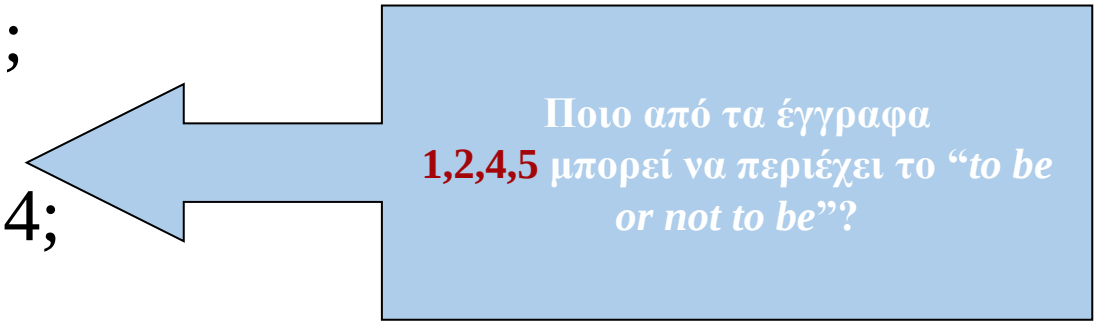
Παράδειγμα

- d1: a b a
- d2: b b c a
- d3: b c d c
- d4: a c d b
- d5: a c b a b



Παράδειγμα

<***be***: 993427;
1: 7, 18, 33, 72, 86, 231;
2: 3, 149;
4: 17, 191, 291, 430, 434;
5: 363, 367, …>



- Για ερωτήματα φράσεων, χρησιμοποιούμε έναν αλγόριθμο φράσεων αναδρομικά στο επίπεδο εγγράφου
- Αλλά τώρα δεν αρκεί η ισότητα των doc id

Επεξεργασία ερωτήματος φράσης

- Βρες τις εγγραφές του ευρετηρίου για τους όρους του ερωτήματος
- Συγχώνευσε τις doc:position λίστες για απαρίθμηση όλων των πιθανών θέσεων

- Δύο συγχωνεύσεις: όρο και θέση

Παράδειγμα

“a b”

d1: a b a
d2: b b c a
d3: b c d c
d4: a c d b
d5: a c b a b

Επεξεργασία ερωτήματος φράσης

Παράδειγμα ερωτήματος: *“to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*”*

TO, 993427:

‹ **1**: ‹7, 18, 33, 72, 86, 231›; **2**: ‹1, 17, 74, 222, 255›; **4**: ‹8, 16, 190, 429, 433›; 5: ‹363, 367›; 7: ‹13, 23, 191›; . . . ›

BE, 178239:

‹ **1**: ‹17, 25›; **4**: ‹17, 191, 291, 430, 434›; **5**: ‹14, 19, 101›; . . . ›

Ερωτήματα γειτονικότητας (Proximity queries)

- Η ίδια γενική μέθοδος για ερωτήματα γειτονικότητας (proximity searches)

- LIMIT! /3 STATUTE /3 FEDERAL /2 TORT
  - Πάλι, */k* means "within *k* words of".
- Μπορούμε να χρησιμοποιήσουμε ευρετήρια θέσεων αλλά όχι ευρετήρια biword.

Πολυπλοκότητα ερώτησης

- Αυξάνει την πολυπλοκότητα της ερώτησης από O(T), T αριθμός εγγράφων σε O(N), N αριθμός token.

Μέγεθος ευρετηρίου

- Μπορούμε να συμπιέσουμε τα position values/offsets
- Παρόλα αυτά, σημαντική αύξηση του χώρου αποθήκευσης των λιστών καταχωρήσεων
- Αλλά χρησιμοποιείται ευρέως

Η σχετική θέση των όρων χρησιμοποιείται και εμμέσως για την κατάταξη των αποτελεσμάτων.

Μέγεθος ευρετηρίου

- Χρειάζεται μια εγγραφή για κάθε εμφάνιση στο έγγραφο αντί για μια για κάθε έγγραφο
- Το *μέγεθος του ευρετηρίου* εξαρτάται από το *μέσο μέγεθος του αρχείου*
  - Μέσο μέγεθος web σελίδας < 1000 όροι
  - SEC filings, books, ακόμα και μερικά επικά ποιήματα … πάνω από 100,000 όρους
- Έστω ένας όρος με συχνότητα 0.01% (1 ανά 1000 όρους) σε ένα έγγραφο

Document size	Postings	Positional postings
1000	1	1
100,000	1	?

? A 1
B 100
C 1000

Rules of thumb

- Ένα ευρετήριο θέσεων είναι 2–4 μεγαλύτερο από ένα απλό ευρετήριο
- Το μέγεθος του *συμπιεσμένου* ευρετηρίου είναι το 35–50% του όγκου του αρχικού κειμένου
- Αυτά αφορούν την Αγγλική (και παρόμοιες) γλώσσες

Συνδυαστικές μέθοδοι

- Αυτές οι δυο προσεγγίσεις μπορεί να συνδυαστούν
  - Για συγκεκριμένες φράσεις (***"Michael Jackson", "Britney Spears"***) η συνεχής συγχώνευση καταχωρήσεων ευρετηρίου θέσεων δεν είναι αποδοτική
    - Ακόμα περισσότερο για φράσεις όπως ***"The Who"***

Πότε biwords αντί για positional indexes?

- Αυτά που συναντώνται συχνά
- Τις ποιο «ακριβές»

ΤΕΛΟΣ 2ου Κεφαλαίου

Ερωτήσεις?

Χρησιμοποιήθηκε υλικό των:

✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*