Ανάκληση Πληροφορίας

Information Retrieval

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 2η

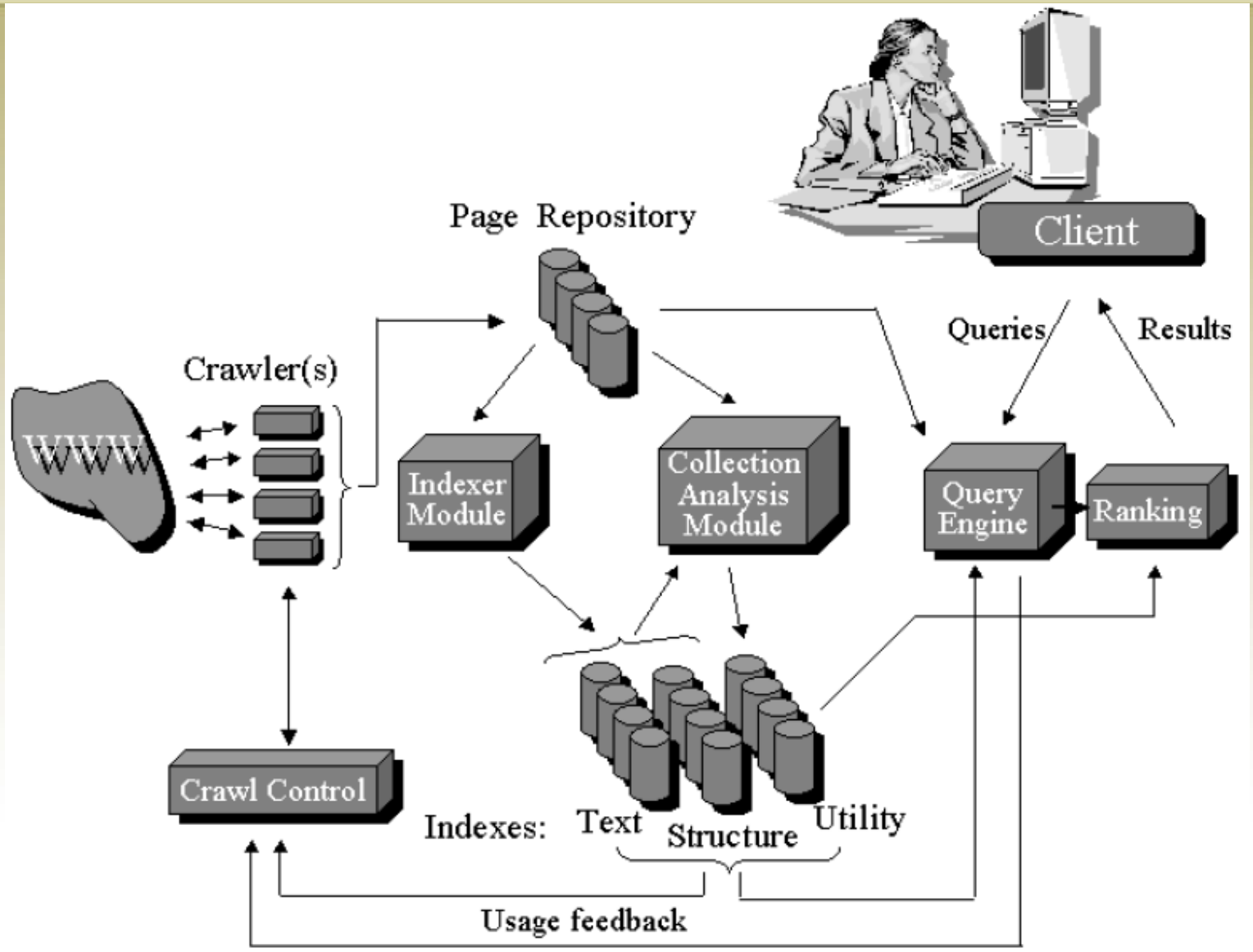
Μεγάλες συλλογές (corpora)

- Έστωσαν $N = 1\text{M}$ έγγραφα, το κάθε ένα με περίπου 1K όρους
- Avg 6 bytes/term, συμπεριλαμβανόμενων των κενών/στίξη
  - Άρα, 6GB δεδομένα στα έγγραφα
- Έστω ότι υπάρχουν $m = 500\text{K}$ *διακριτοί* όροι μεταξύ αυτών

Ο πίνακας πρακτικά δεν χτίζεται

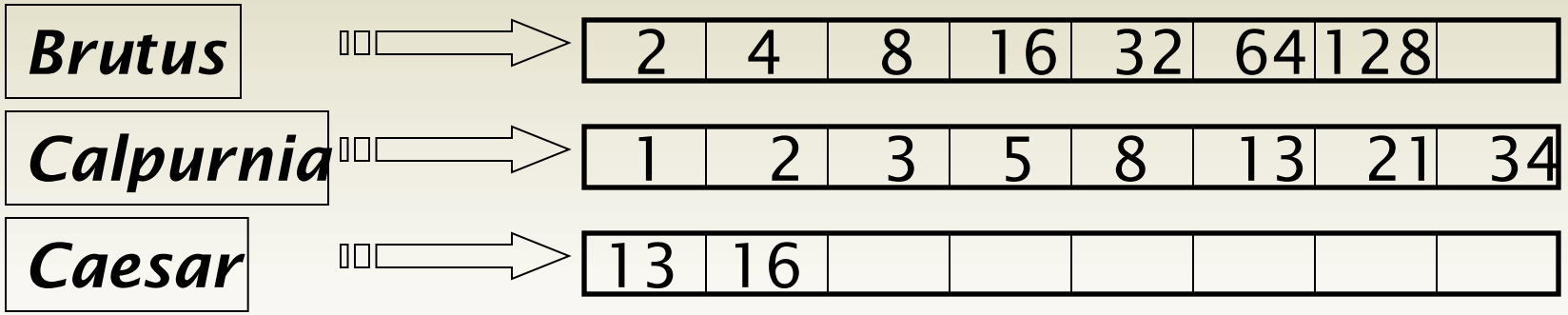
- Πίνακας με 500K x 1M έχει μισό-τρις 0's και 1's
- Αλλά έχει πάνω από ένα δις 1's
  - Ο πίνακας είναι εξαιρετικά αραιός
- Ποια θα ήταν καλύτερη αναπαράσταση; ⇐ Γιατί;
  - Καταγράφουμε μόνο τις θέσεις των 1

Αρχιτεκτονική Μηχανής Αναζήτησης



Inverted index (αντεστραμμένο ευρετήριο)

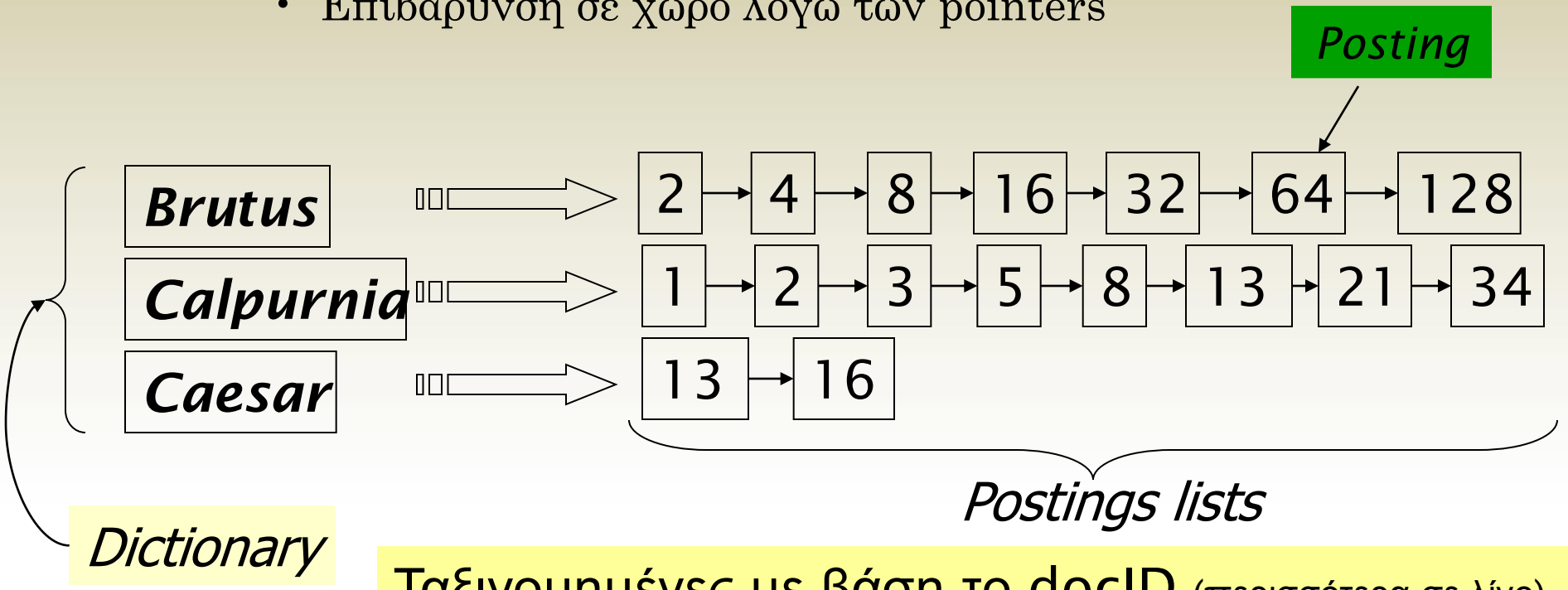
- Για κάθε όρο *T*, πρέπει ν' αποθηκεύσουμε όλα τα έγγραφα που περιέχουν τον *T*
- Να χρησιμοποιήσουμε πίνακα ή λίστα για τον σκοπό αυτό;



Τι θα συμβεί εάν ο όρος ***Caesar*** προστεθεί στο έγγραφο14;

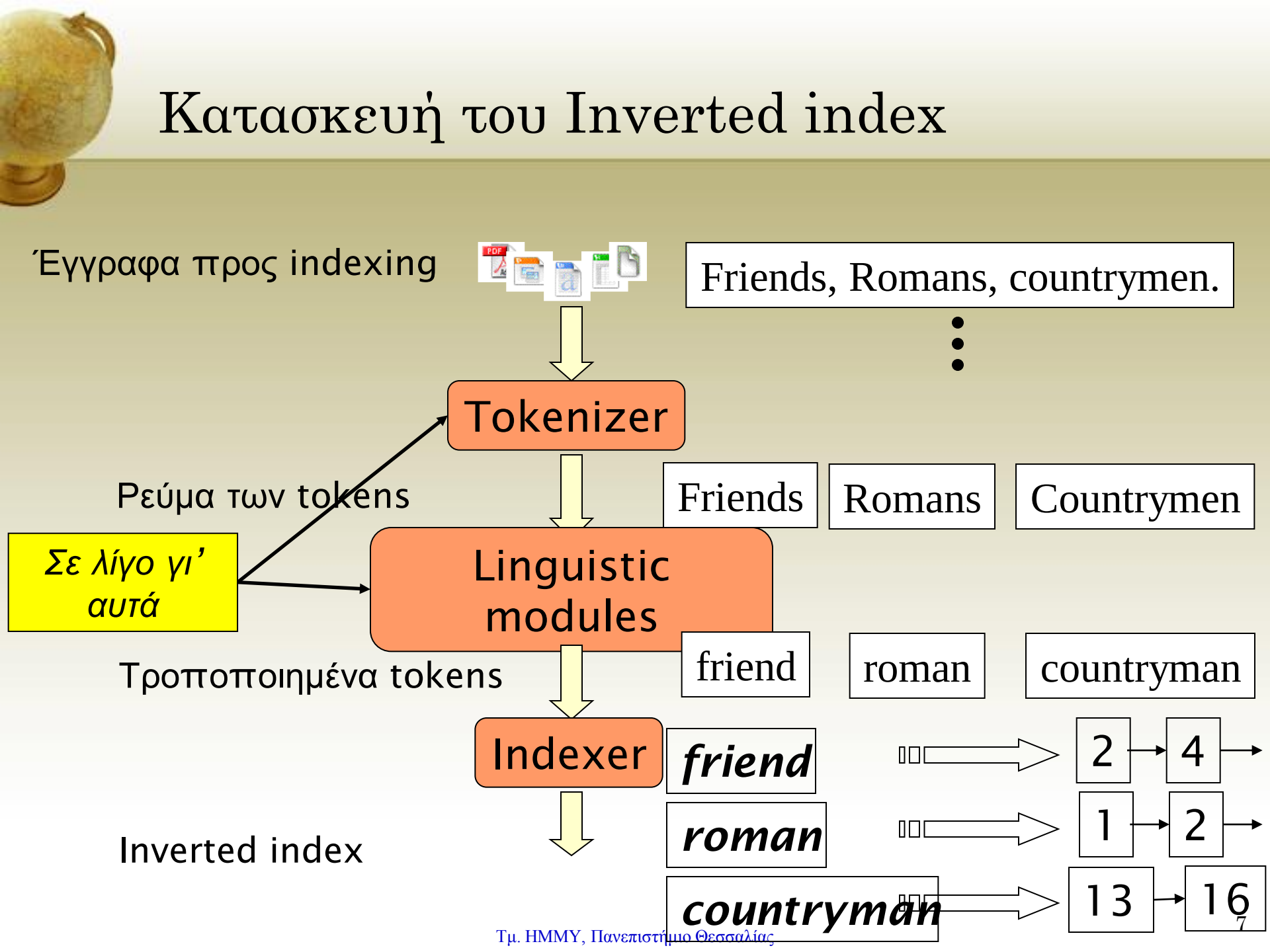
Inverted index

- Οι συνδεδεμένες λίστες γενικώς είναι προτιμότερες έναντι των πινάκων
  - Δυναμική δέσμευση χώρου
  - Εισαγωγή όρων στα έγγραφα είναι εύκολη
  - Επιβάρυνση σε χώρο λόγω των pointers



Ταξινομημένες με βάση το docID (περισσότερα σε λίγο)

Κατασκευή του Inverted index



Βήματα του indexer

Ακολουθία των ζευγών (Modified token, DocumentID)

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	Doc #
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 κατά όρο

Βήμα-θεμέλιο του indexing

Term	Doc #
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	Doc #
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

- Πολλαπλές εμφανίσεις του ίδιου όρου στο ίδιο έγγραφο συγχωνεύονται
- Προστίθεται πληροφορία σχετική με την συχνότητα εμφάνισης του όρου

Γιατί πληροφορία συχνότητας;
Θα το συζητήσουμε αργότερα

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

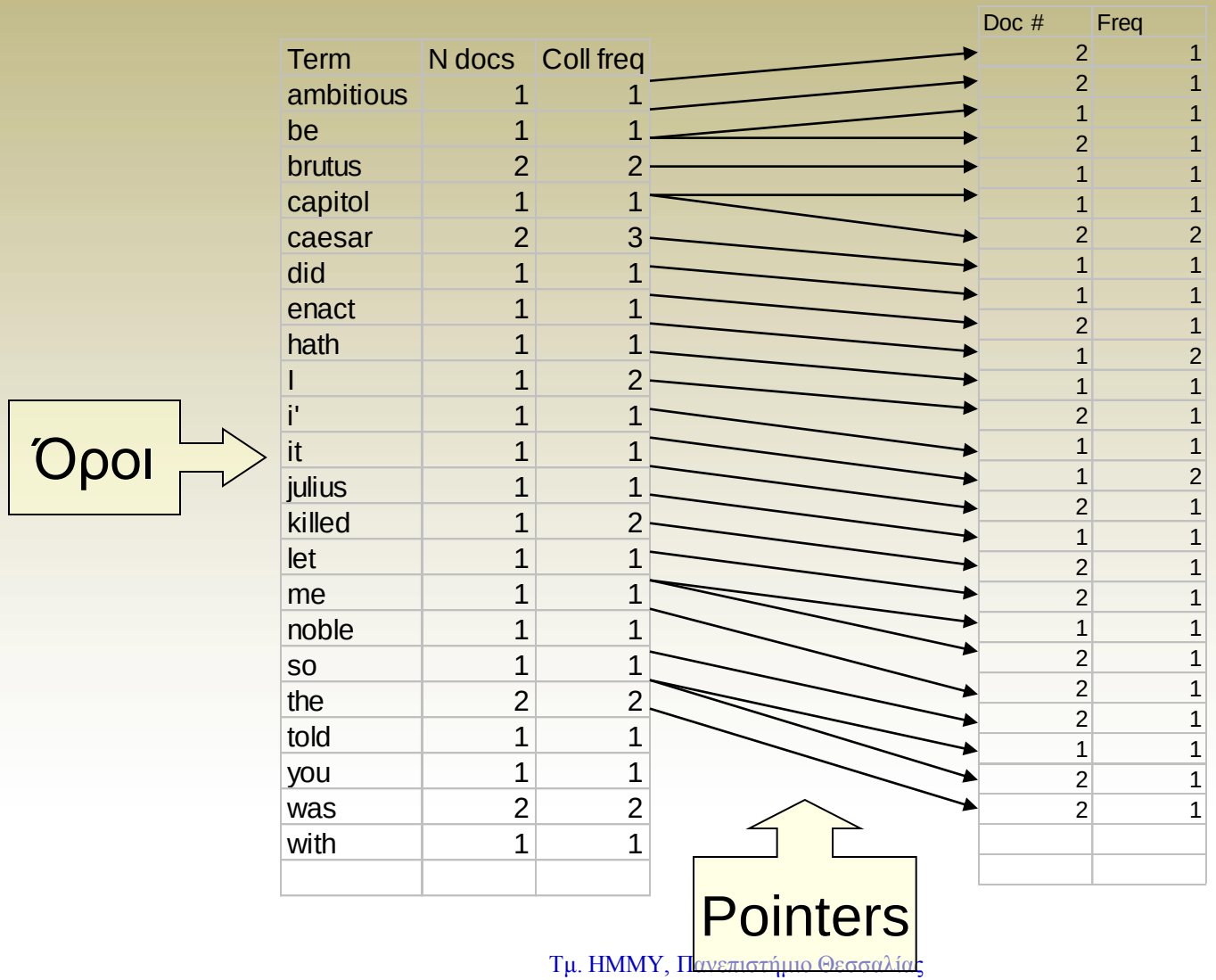
Το αποτέλεσμα διαμερίζεται στο *Dictionary* και στο αρχείο με τα *Postings*

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

Term	N docs	Coll freq
ambitious	1	1
be	1	1
brutus	2	2
capitol	1	1
caesar	2	3
did	1	1
enact	1	1
hath	1	1
I	1	2
i'	1	1
it	1	1
julius	1	1
killed	1	2
let	1	1
me	1	1
noble	1	1
so	1	1
the	2	2
told	1	1
you	1	1
was	2	2
with	1	1

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

Τι πληρώνουμε για αποθήκευση;



Term	N docs	Coll freq
ambitious	1	1
be	1	1
brutus	2	2
capitol	1	1
caesar	2	3
did	1	1
enact	1	1
hath	1	1
I	1	2
i'	1	1
it	1	1
julius	1	1
killed	1	2
let	1	1
me	1	1
noble	1	1
so	1	1
the	2	2
told	1	1
you	1	1
was	2	2
with	1	1

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

Θα ποσοτικοποιήσουμε την αποθήκευση αργότερα

Το ευρετήριο (index) που μόλις χτίσαμε

- Πώς επεξεργαζόμαστε ένα ερώτημα;
  - Αργότερα – τι είδους ερωτήματα μπορούμε να επεξεργαστούμε;

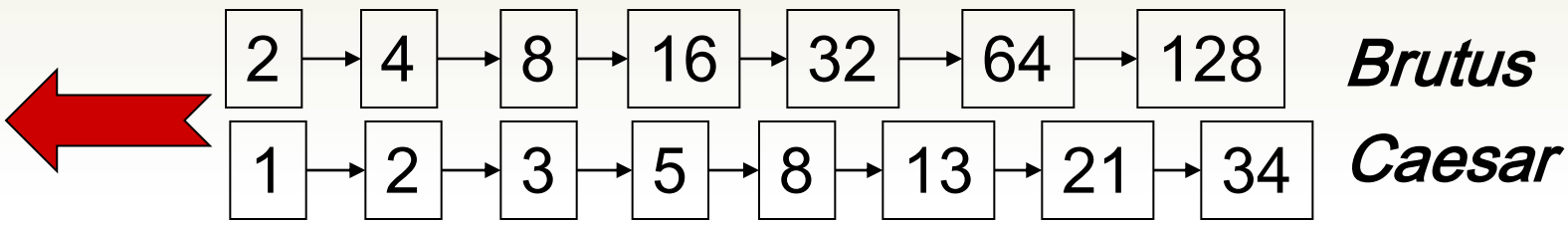
Η εστίασή μας σήμερα

Επεξεργασία ερωτήματος: AND

- Θεωρήστε το ερώτημα:

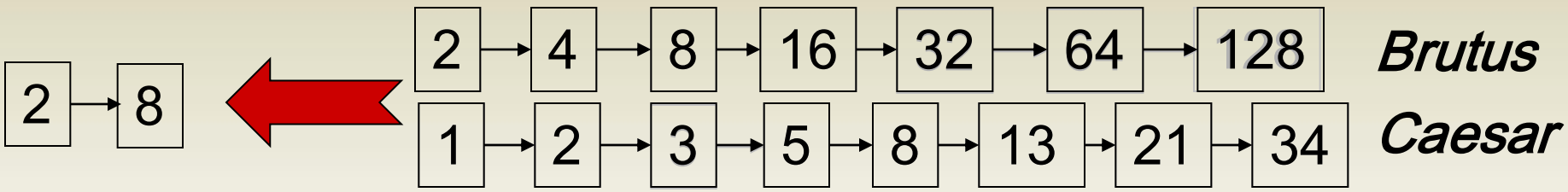
  Brutus** AND **Caesar

  - Εντόπισε τον όρο ***Brutus*** στο Dictionary
    - Ανάκτησε την λίστα με τα postings που του αντιστοιχεί
  - Εντόπισε τον όρο *Caesar* στο Dictionary
    - Ανάκτησε την λίστα με τα postings που του αντιστοιχεί
  - "Συγχώνευσε" τις δυο λίστες των postings:



Η συγχώνευση

- Διασχίστε τις δυο λίστες ταυτόχρονα, σε χρόνο γραμμικό σε σχέση με τον συνολικό αριθμό των postings entries



Εάν τα μήκη των λιστών είναι *x* και *y*, η συγχώνευση παίρνει O(*x*+*y*) λειτουργίες

Προσοχή: τα postings είναι ταξινομημένα ως προς το docID

Boolean ερωτήματα: Ακριβές ταίριασμα

- Το Boolean μοντέλο ανάκτησης είναι ικανό να θέτει ένα ερώτημα το οποίο είναι μια Boolean έκφραση:
  - Τα Boolean ερωτήματα είναι ερωτήματα που φτιάχνοναι με χρήση των τελεστών *AND, OR* και *NOT* οι οποίοι ενώνουν τους όρους του ερωτήματος
    - Βλέπει κάθε έγγραφο ως ένα σύνολο από λέξεις
    - Είναι ακριβές: τα ανακτηθέντα έγγραφο είτε ικανοποιούν ακριβώς τις συνθήκες είτε όχι
- Ήταν το κυρίαρχο εμπορικό μοντέλο ανάκτησης για πάνω από 3 δεκαετίες
- Στις "επαγγελματικές" αναζητήσεις (π.χ., δικηγόροι) χρησιμοποιούνται ακόμη Boolean ερωτήματα

Παράδειγμα: WestLaw http://www.westlaw.com/

- Η μεγαλύτερη εμπορική (συνδρομητική) νομική υπηρεσία αναζήτησης (ξεκίνησε το 1975; το ranking προστέθηκε το 1992)
- Πολλά terabytes δεδομένων; 700,000 χρήστες
- Η πλειονότητα των χρηστών χρησιμοποιεί *ακόμα* boolean ερωτήματα
- Παράδειγμα ερωτήματος:
  - What is the statute of limitations in cases involving the federal tort claims act?
  - LIMIT! /3 STATUTE ACTION /S FEDERAL /2 TORT /3 CLAIM
- /3 = within 3 words, /S = in same sentence

Παράδειγμα: WestLaw http://www.westlaw.com/

- Ένα ακόμη παράδειγμα ερωτήματος:
  - Requirements for disabled people to be able to access a workplace
  - disabl! /p access! /s work-site work-place (employment /3 place
- Παρατηρήστε ότι το ΚΕΝΟ σημαίνει διάζευξη και όχι σύζευξη!
- Μακρά, ακριβή ερωτήματα; τελεστές εγγύτητας (proximity); Όχι όπως η Web search

Boolean ερωτήματα: Πιο γενικές συγχωνεύσεις

- Άσκηση: Προσαρμόστε την συγχώνευση για τα ερωτήματα:

  Brutus** AND NOT **Caesar

  Brutus** OR NOT **Caesar

Μπορούμε και εδώ να εκτελέσουμε την συγχώνευση σε χρόνο $O(x+y)$ ή τι μπορούμε να επιτύχουμε;

Συγχώνευση

Τι ισχύει για τυχείες Boolean εκφράσεις;

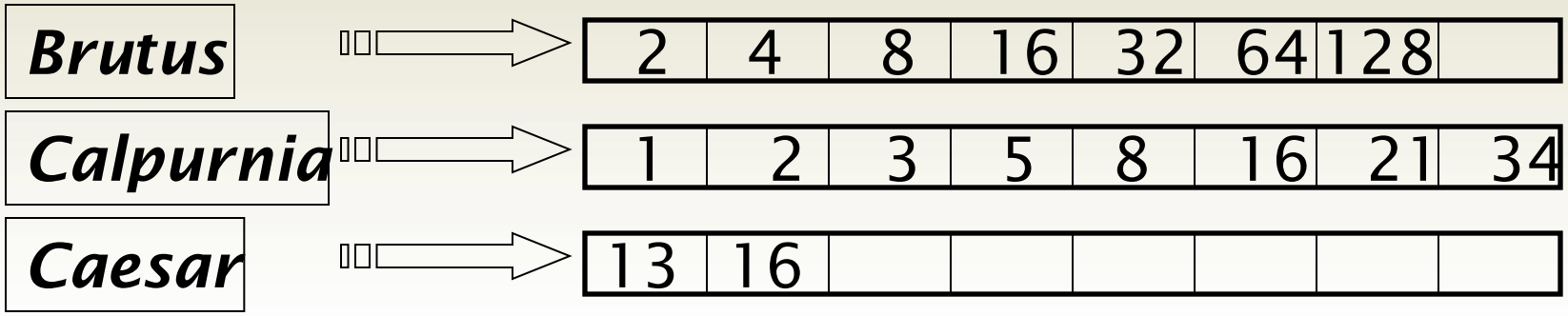
***(Brutus** OR **Caesar)** AND NOT*

(Antony** OR **Cleopatra)

- Μπορούμε πάντα να εκτελέσουμε συγχώνευση σε "γραμμικό" χρόνο;
  - Γραμμικό ως προς τι;
- Μπορούμε να επιτύχουμε κάτι καλύτερο;

Βελτιστοποίηση επεξεργασίας ερωτημάτων

- Ποια είναι η καλύτερη σειρά για την επεξεργασία;
- Θεωρήστε ένα ερώτημα που αποτελείται από όρους συνδεόμενους με *AND*
- Για κάθε έναν από τους όρους, ανακτήστε την λίστα με τα postings, κατόπιν κάντε *AND* μεταξύ τους

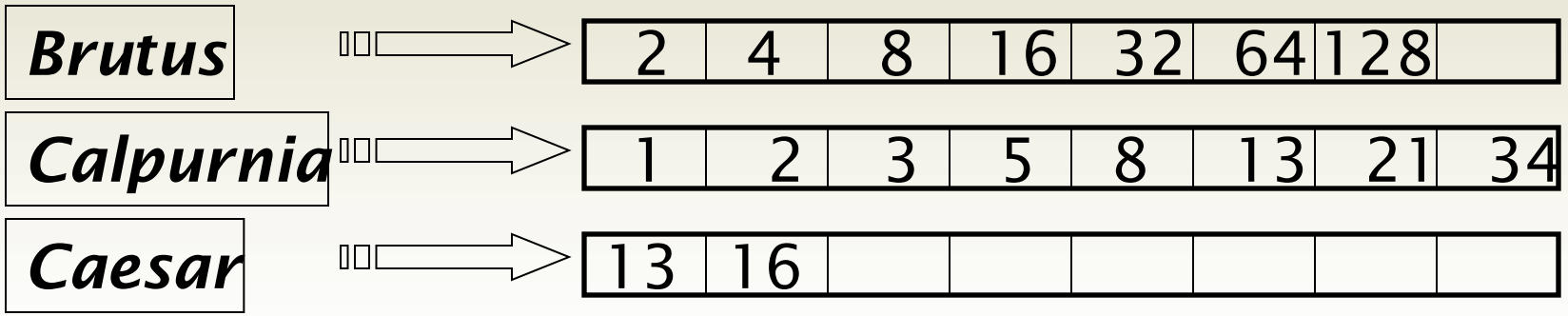


Ερώτημα: ***Brutus*** *AND* ***Calpurnia*** *AND* ***Caesar***

Παράδειγμα βελτιστοποίησης επεξεργασίας ερωτήματος

- Επεξεργασία με αυξανόμενη freq:
  - *Ξεκινήστε με το μικρότερο σύνολο, κατόπιν το αμέσως μεγαλύτερο, κ.τ.λ.*

Αυτός είναι ο λόγος που αποθηκεύσαμε πληροφορία συχνότητας στο dictionary



Επεξεργαστείτε το ερώτημα ως (***Caesar*** *AND* ***Brutus)*** *AND* ***Calpurnia***.

Πιο γενική βελτιστοποίηση

- Π.χ., ***(madding** OR **crowd)** AND **(ignoble** OR **strife)***
- Ανακτήστε την freq όλων των όρων
- Εκτιμήστε το μέγεθος κάθε *OR* ως να είναι το άθροισμα των freq τους (συντηρητικό)
- Επεξεργαστείτε το ερώτημα σε αυξανόμενα μεγέθη των *OR*

Άσκηση

Προτείνετε μια σειρά επεξεργασία του ερωτήματος

***(tangerine** OR **trees)** AND*
***(marmalade** OR **skies)** AND*
(kaleidoscope** OR **eyes)

Term	Freq
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

Ασκήσεις στην επεξεργασία ερωτημάτων

- Εάν το ερώτημα είναι ***friends*** *AND* ***romans*** *AND (NOT* ***countrymen****)*, πώς μπορούμε να αξιοποιήσουμε την freq του ***countrymen***;
- Άσκηση: Επεκτείνετε την τεχνική της συγχώνευσης για ένα τυχαίο Boolean ερώτημα. Μπορούμε πάντα να εγγυηθούμε γραμμικό χρόνο ως προς το συνολικό μέγεθος των postings;
  - Υπόδειξη: Ξεκινήστε με την περίπτωση ενός Boolean *formula* ερωτήματος: σε ένα τέτοιο ερώτημα, κάθε όρος στο ερώτημα εμφανίζεται μόνο μια φορά στο ερώτημα

Τι άλλο στην IR; Πέρα από την αναζήτηση όρων

- Φράσεις:
  - ***Stanford University***
- Εγγύτητα (proximity): Βρείτε έγγραφα που περιέχουν τον όρο ***Gates*** *NEAR* ***Microsoft***
  - Χρειαζόμαστε ευρετήριο για να καταγράφουμε και να αξιοποιούμε πληροφορία τοποθεσίας στα έγγραφα. Περισσότερα πάνω σ' αυτό σε επόμενες διαλέξεις
- Ζώνες (zones) στα έγγραφα: Βρείτε έγγραφαμε (*author* = ***Ullman***) *AND* (text contains ***automata***)

Διάταξη ή διαβάθμιση (ranking) των αποτελεσμάτων της αναζήτησης

- Τα Boolean ερωτήματα δουλεύουν με inclusion ή exclusion των εγγράφων
- Συχνά (σχεδόν πάντα) επιθυμούμε να διατάξουμε/ομαδοποιήσουμε (rank/group) τα αποτελέσματα
  - Χρειάζεται να μετρήσουμε την εγγύτητα από το ερώτημα προς κάθε έγγραφο
  - Χρειάζεται ν' αποφασίσουμε εάν τα έγγραφα που παρουσιάζονται στους χρήστες είναι singletons, ή μια ομάδα εγγράφων ποτ καλύπτουν διαφορετικές πτυχές του ερωτήματος

Clustering και classification

- Δεδομένου ενός συνόλου εγγράφων, ομαδοποιήστε τα με βάση τα περιεχόμενά τους

- Δεδομένου ενός συνόλο θεμάτων (topics), και ενός νέου εγγράφου D, αποφασίστε σε ποιο/α θέμα/τα ανήκει το D