Ανάκληση Πληροφορίας

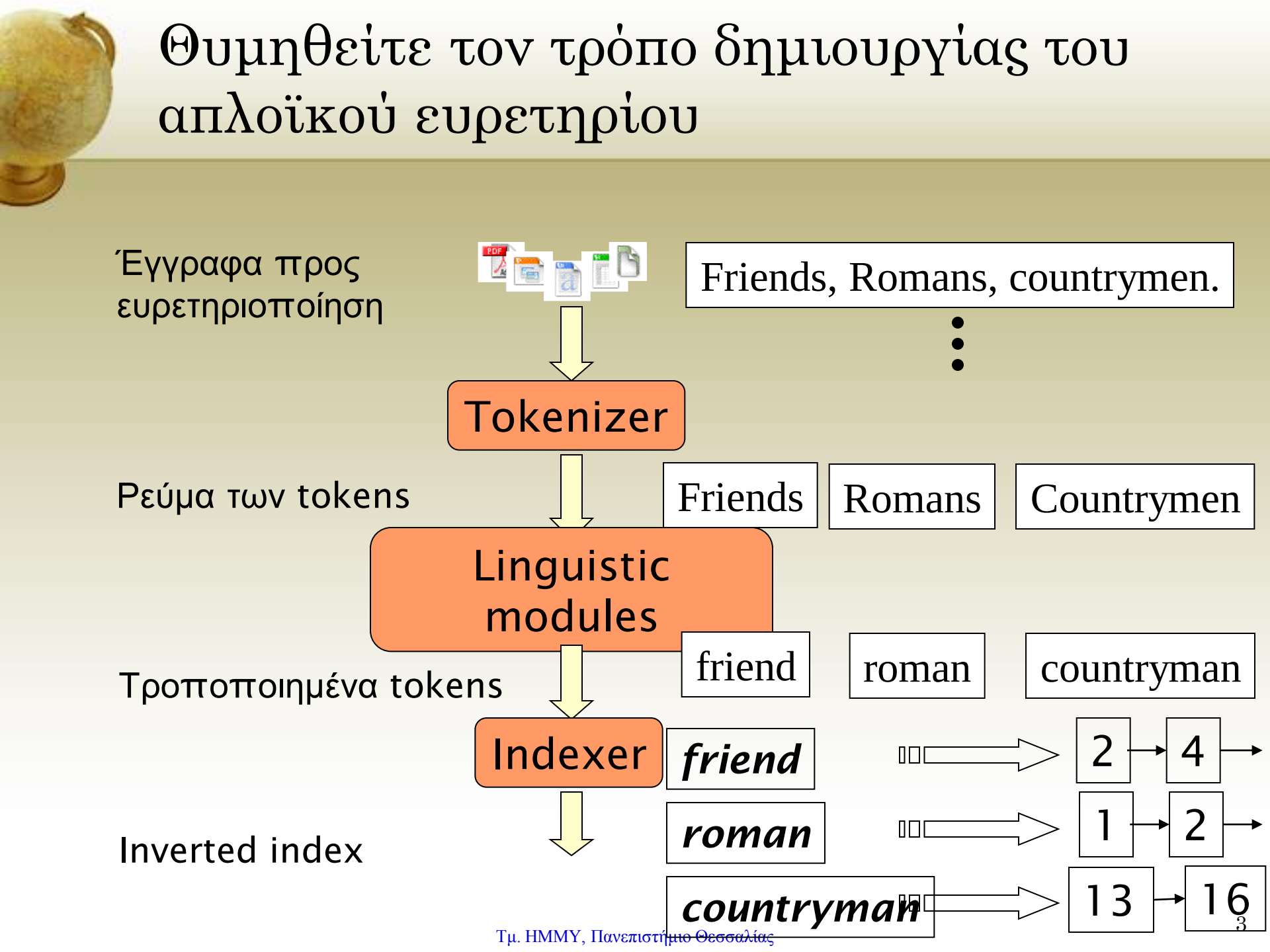
Information Retrieval

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 3η

Dictionary και Postings

Θυμηθείτε τον τρόπο δημιουργίας του απλοϊκού ευρετηρίου



Parsing ενός εγγράφου

- Ποιο είναι το format του;
  - pdf/word/excel/html;
- Σε ποια γλώσσα είναι γραμμένο;
- Σε ποιο character set;

Κάθε μια από αυτές τις ερωτήσεις εισάγει ένα πρόβλημα κατηγοριοποίησης (classification problem), που δεν θα μελετήσουμε στο παρόν μάθημα, αλλά που περιγράφεται στο βιβλίο Introduction to Information Retrieval

Συνήθως, αυτές οι εργασίες γίνονται ευρεστικά …

Επιπλοκές: Μορφή/γλώσσα

- Το έγγραφα που θα μπουν στο ευρετήριο μπορεί να έχουν γραφτεί σε διάφορες γλώσσες
  - Το ίδιο ευρετήριο μπορεί να περιέχει όρους (terms) προερχόμενους από διάφορες γλώσσες
- Ενα έγγραφο ή τμήματά του μπορεί να περιέχει κείμενο γραμμένο σε διαφορετικές γλώσσες
  - Email στα Γαλλικά με επισυναπτόμενο ένα pdf στα Γερμανικά
- Πώς ορίζεται το "έγγραφο-μονάδα";
  - Είναι ένα αρχείο;
  - Ένα email;
  - Ένα email με 5 επισυναπτόμενα;
  - Μια ομάδα αρχείων (PPT ή LaTeX σε HTML)

Tokenization

Tokenization

- Είσοδος: "***Friends, Romans and Countrymen***"
- Έξοδος: Tokens
  - ***Friends***
  - ***Romans***
  - ***Countrymen***
- Κάθε τέτοιο token είναι υποψήφιο να γίνει όρος του λεξικού του ευρετηρίου, μετά από κάποια επιπλέον επεξεργασία
  - που περιγράφεται στη συνέχεια
- Ποια όμως είναι τα έγκυρα tokens;

Tokenization

- Ζητήματα στην διαδικασία tokenization:
  - ***Finland's capital*** →

    Finland ή ***Finlands*** ή ***Finland's***
  - ***Hewlett-Packard*** → ***Hewlett*** and ***Packard*** ως δυο tokens;
    - ***State-of-the-art***: διάσπαση ως ακολυθία λέξεων διαχωρισμένων με παύλες;
    - ***co-education*** ;
    - ***the hold-him-back-and-drag-him-away-maneuver*** ;
  - ***San Francisco***: ένα ή δυο tokens; Πώς αποφασίζεις ότι πρόκειται για ένα token;

Αριθμοί

- ***3/12/91*** ***Mar. 12, 1991***
- ***55 B.C.***
- ***B-52***
- ***My PGP key is 324a3df234cb23e***
- ***100.2.86.144***
  - Συχνά, δεν τους εισάγουμε στο ευρετήριο ως κείμενο
    - Αλλά ίσως είναι χρήσιμο: σκεφτείτε την περίπτωση όταν αναζητείτε error codes/stacktraces
    - (Θα μπορούσαμε να το κάνουμε με χρήση των n-grams: δείτε επόμενες διαλέξεις)
  - Δημιουργία ξεχωριστού ευρετηρίου για "meta-data"
    - Ημερομηνία δημιουργίας, format, κ.τ.λ.

Tokenization: Ζητήματα γλώσσας

- ***L'ensemble*** → ένα ή δυο tokens;
  - ***L*** ή ***L'*** ή ***Le*** ;
  - Επιθυμούμε το ***l'ensemble*** να ταύτιστεί με το ***un ensemble***

- Τα "πολυ-σύνθετα" ουσιαστικά που επιτρέπει η Γερμανική γλώσσα δεν διασπώνται
  - Lebensversicherungsgesellschaftsangestellter
  - 'life insurance company employee'

Tokenization: Ζητήματα γλώσσας

- Τα αραβικά (και Εβραϊκά) γράφονται από δεξιά προς αριστερά. αλλά μερικά στοιχεία όπως οι αριθμοί γράφονται από αριστερά προς δεξιά
- Οι λέξεις διαχωρίζονται, αλλά τα γράμματα μέσα σε μια λέξη σχηματίζουν σύνθετες μορφές
- استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.
- ← → ← → ← start
- 'Algeria achieved its independence in 1962 after 132 years of French occupation.'
- Με Unicode, η "επιφανειακή" αναπαράσταση είναι σύνθετη, αλλά η αποθηκευμένη μορφή είναι προφανής

Κανονικοποιήση

- Χρειάζεται να “κανονικοποιήσουμε” τους όρους που θα μπουν στο ευρετήριο, αλλά και τα ερωτήματα, ώστε να έχουν την ίδια μορφή
  - Επιθυμούμε να ταιριάξουν το ***U.S.A.*** και το ***USA***
- Ορίζουμε έμμεσα κλάσεις ισοδυναμίας των όρων
  - Π.χ., με διαγραφή των τελειών από έναν όρο
- Η εναλλακτική είναι να εκτελέσουμε ασυμμετρική επέκταση:
  - Εισάγουμε: ***window*** Αναζητείται: ***window, windows***
  - Εισάγουμε : ***windows*** Αναζητείται : ***Windows, windows***
  - Εισάγουμε : ***Windows*** Αναζητείται : ***Windows***
- Εν δυνάμει πιο ισχυρό, αλλά λιγότερο αποδοτικό

Κανονικοποιήση: Άλλες γλώσσες

- Προφορά: ***résumé*** vs. ***resume***.
- Το πιο σημαντικό κριτήριο:
  - Πώς αναμένεις ότι θα γράψουν οι χρήστες ερωτήματα γι' αυτές τις λέξεις;
- Ακόμη και για γλώσσες που υποστηρίζουν εκ της φύσεώς τους "σημεία στίξης" που καθορίζουν την προφορά, πολλές φορές οι χρήστες δεν τα χρησιμοποιούν όταν συντάσσουν κάποιο ερώτημα
- Στα γερμανικά: Tuebingen vs. Tübingen
  - Θα έπρεπε να είναι ισοδύναμα

Κανονικοποιήση: Άλλες γλώσσες

- Χρειάζεται να "κανονικοποιήσουμε" το κείμενο που μπαίνει στο ευρετήριο καθώς επίσης και το ερώτημα στην ίδια μορφή

7月30日 vs. 7/30

- Αναγνώριση αλφαβήτου σε επίπεδο χαρακτήρα και κατόπιν μετατροπή
  - Η tokenization δεν είναι χωριστή διαδιακασία από αυτή
  - Μερικές φορές διφορούμενο

Morgen will ich in MIT …

Το "mit" είναι στα Γερμανικά;

Κεφαλαία-μικρά γράμματα

- Μετροπή όλων των γραμμάτων σε μικρά
  - εξαίρεση: Κεφαλαία (εντός προτάσεως;)
    - Π.χ., ***General Motors***
    - ***Fed*** vs. ***fed***
    - ***SAIL*** vs. ***sail***
  - Συχνά είναι καλύτερο να μετρέπουμε τα πάντα σε μικρά, αφού οι χρήστες θα χρησιμοποιήσουν μικρά γράμματα ανεξάρτητα από το ένα το ερώτημά τους περιέχει λέξη/όρο που κανονικά θα γραφόταν με κεφαλαίο...

Τερματικές-λέξεις (Stop words)

- Με μια stop list, εξαιρούμε από το λεξικό τις πιο κοινές λέξεις. Διαίσθηση:
  - Μεταφέρουν μικρό σημασιολογικό περιεχόμενο: *the, a, and, to, be*
  - Καταλαμβάνουν πολύ χώρο: ~30% των postings για τις top-30
- Αλλά η τάση είναι διαφορετική:
  - Οι καλές τεχνικές συμπίεσης σημαίνουν ότι ο χώρος για την συμπερίληψη των stopwords σε ένα σύστημα είναι μικρός
  - Οι καλές τεχνικές βελτιστοποίησης ερωτημάτων σημαίνει ότι πληρώνουμε μικρό τίμημα κατά την υποβολή του ερωτήματος για την συμπερίληψη των stop words
  - Τις χρειαζόμαστε για:
    - Ερωτήματα-φράσεις: “King of Denmark”
    - Διάφορους τίτλους τραγουδιών, κ.τ.λ.: “Let it be”, “To be or not to be”
    - “Σχεσιακά” ερωτήματα: “flights to London”

Thesauri και soundex

- Χειρίζεται συνώνυμα και ομόνυμα
  - Χειρωνακτική κατασκευή των κλάσεων ισοδυναμίας
    - Π.χ., ***car*** = ***automobile***
    - ***color*** = ***colour***
- Δημιουργία ευρετηρίου για αυτές τις κλάσεις ισοδυναμίας
  - Όταν το έγγραφο περιέχει την λέξη ***automobile***, το βάζουμε στο ευρετήριο και κάτω από το ***car*** επίσης (συνήθως, και ανάποδα)
- Και επέκταση ενός ερωτήματος;
  - Όταν το ερώτημα περιέχει τον όρο ***automobile***, κοιτάζουμε και κάτω από το ***car*** επίσης

Soundex

- Είναι η παραδοσιακή κλάση των ευρεστικών μεθόδων για επέκταση ενός ερωτήματος στα φωνητικά του αντίστοιχα
  - Εξαρτάται από την γλώσσα – κυρίως χρησιμοποιείται για ουσιαστικά
  - Π.χ., ***chebyshev → tchebycheff***

- Περισσότερα επ' αυτού σε επόμενη διάλεξη …

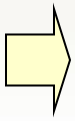
Lemmatization

- Μετατροπή πληθωριστικών/παραλλαγμένων μορφών στην βασική μορφή
- Π.χ.,
  - *am, are, is* → *be*
  - *car, cars, car's, cars'* → *car*
- *the boy's cars are different colors* → *the boy car be different color*
- Η lemmatization υπονοεί την εκτέλεση "κατάλληλης" μετατροπής στην μορφή των λέξεων του dictionary

Stemming

- "Μετατροπή" των όρων στην ρίζα τους, πριν εισαχθούν στο ευρετήριο
- Το "stemming" υπονοεί αποκοπή της κατάληξης
  - Εξαρτάται φυσικά από την γλώσσα
  - Π.χ., ***automate(s), automatic, automation*** όλα "ελαττώνονται" στο ***automat***.

for example compressed and compression are both accepted as equivalent to compress.



for exampl compress and compress ar both accept as equival to compress

Ο αλγόριθμος του Porter

- Ο πιο κοινός αλγόριθμος για stemming της αγγλικής
  - Αποτελέσματα υποδεικνύουν ότι είναι τουλάχιστον τόσο καλός όσο και οι άλλοι stemmers
- Συμβάσεις + 5 φάσεις από reductions
  - Οι φάσεις εφαρμόζονται η μια μετά την άλλη
  - Κάθε φάση αποτελείτει από ένα σύνολο εντολών
  - Δειγματική σύμβαση: *Of the rules in a compound command, select the one that applies to the longest suffix*

Τυπικοί κανόνες στον αλγόριθμο Porter

- *sses* → *ss*
- *ies* → *i*
- *ational* → *ate*
- *tional* → *tion*

- Weight of word sensitive rules
- *(m>1) EMENT* →
  - *replacement* → *replac*
  - *cement* → *cement*

Άλλοι stemmers

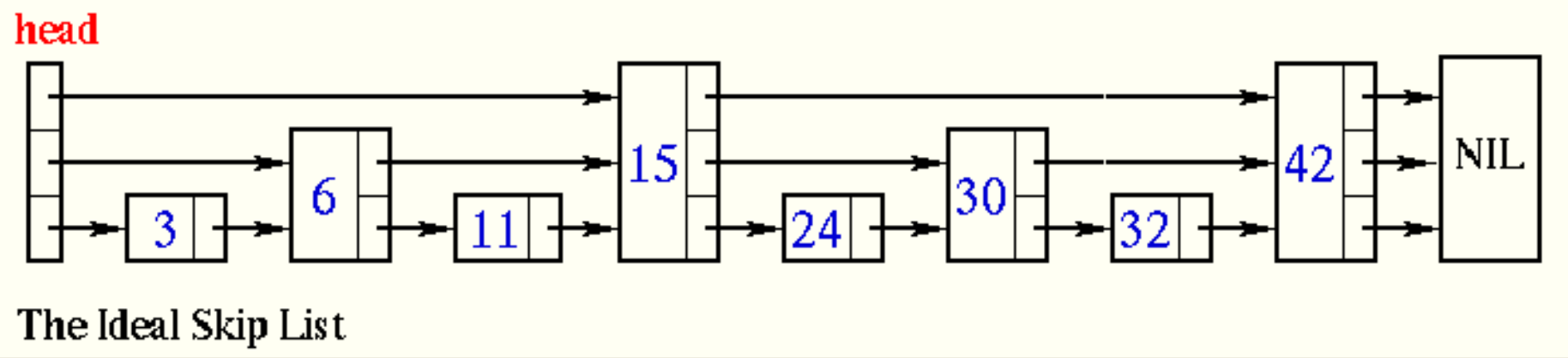
- Υπάρχουν και άλλοι stemmers, π.χ., του Lovins
  http://snowball.tartarus.org/algorithms/lovins/stemmer.html
  - Είναι ενός περάσματος, λειτουργεί με αφαίρεση του μακρύτερου suffix (περίπου 250 κανόνες)
  - Κατάλληλο κυρίως για γλωσσολόγους, αλλά και για IR
- Πλήρης μορφολογική ανάλυση – με μέτρια οφέλη για να την ανάκτηση
- Το stemming και οι κανονικοποιήσεις βοηθούν;
  - Είναι αντικρουόμενες οι απόψεις: σίγουρα βοηθούν στο *recall* (ποσότητα ανακτώμενων σχετικών εγγράφων) για μερικά ερωτήματα, αλλά βλάπτουν το *precision* (ποσοστό ανακτώμενων σχετικών εγγράφων) για άλλα

Language-specificity

- Πολλά από τα προηγούμενα χαρακτηριστικά αξιοποιούν μετασχηματισμούς, οι οποίοι είναι
  - Εξειδικευμένοι στην κάθε γλώσσα
  - (Συχνά) Ειδικοί για την κάθε εφαρμογή
- Αυτοί είναι “plug-in” προσθήκες στην διαδικασία δημιουργίας του ευρετηρίου
- Υπάρχουν διαθέσιμοι και open source και εμπορικοί plug-ins για τον χειρισμό τους

Γρηγορότερη συγχώνευση των postings: Δείκτες Παράκαμψης (Skip pointers)

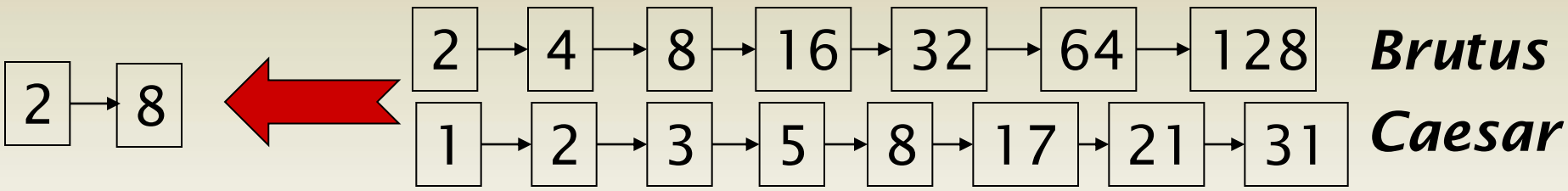
Υπόβαθρο: Skip lists



The Ideal Skip List

Θυμηθείτε την βασική συγχώνευση

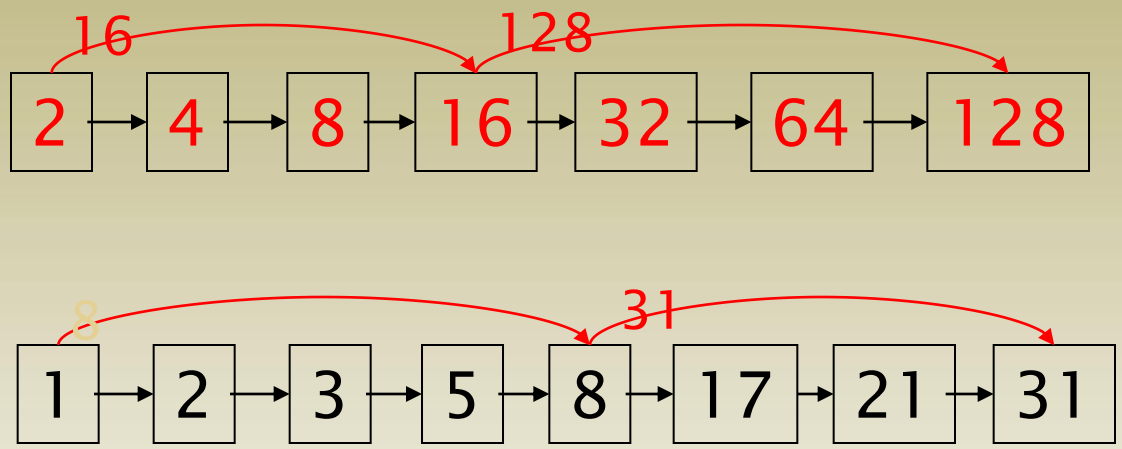
- Διασχίζουμε τις δυο λίστες των postings παράλληλα, σε χρόνο γραμμικό ως προς τον συνολικό αριθμό των postings



Εάν τα μήκη των λιστών είναι *m* και *n*, η συγχώνευση απαιτεί O(*m*+*n*) λειτουργίες

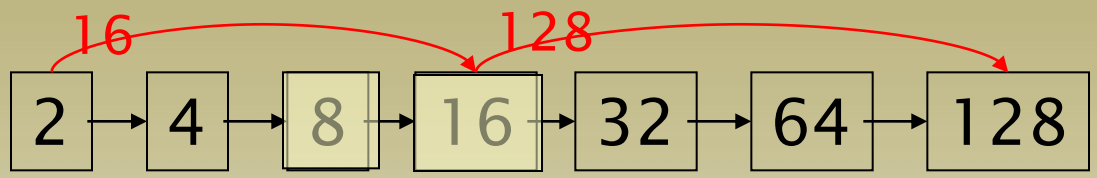
Μπορούμε κάτι καλύτερο;
Ναι, εάν το ευρετήριο δεν αλλάζει πολύ γρήγορα

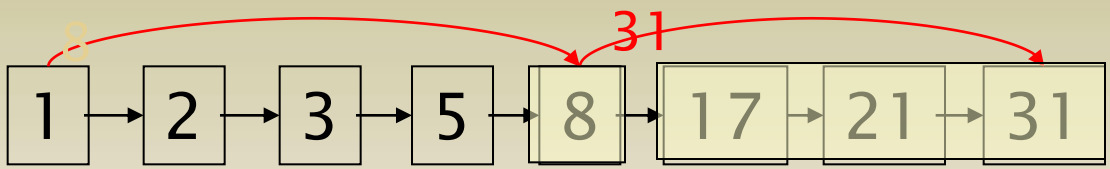
Επαύξηση των postings με skip pointers (όταν δημιουργείται το ευρετήριο)



- Γιατί;
- Για να παρακάμψουμε postings που δεν θα συνεισφέρουν στο αποτέλεσμα τις αναζήτησης
- Πώς;
- Πού τοποθετούμε skip pointers;

Επεξεργασία ερωτήματος με τους skip pointers





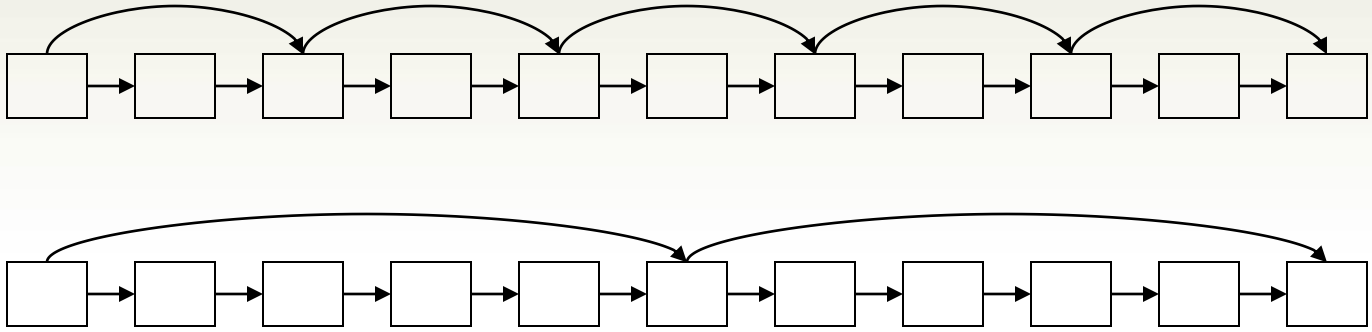
Υποθέστε ότι έχουμε διασχίσει τις λίστες μέχρι να φτάσουμε στην επεξεργασία του **8** σε κάθε λίστα

Όταν φτάνουμε στο **16** στην πάνω λίστα, βλέπουμε ότι το επόμενο στοιχείο είναι το **32**

Αλλά το επόμενο του **8** με βάση τον skip pointer στην κάτω λίστα είναι το **31**, οπότε μπορούμε παρακάμψουμε τα postings της κάτω λίστας μέχρι εκεί που δείχνει ο skip pointer

Πού τοποθετούμε “παρακάμψεις”;

- Tradeoff:
  - Περισσότερα skips → μικρότερες απόστάσεις μεταξύ των skips ⇒ πιο πιθανή η χρήση ενός skip. Αλλά, πολλές συγκρίσεις ώστε να γίνει χρήση των skip pointers
  - Λιγότερα skips → λιγότερες συγκρίσεις δεικτών, αλλά μεγαλύτερες αποστάσεις μεταξύ των ⇒ λιγότερα χρησιμοποιημένα skips



Τοποθέτηση των skips

- Απλό ευρεστικό: για μια λίστα postings με μήκος L, χρησιμοποιήστε $\sqrt{L}$ ισαπέχοντες skip pointers
- Αυτή η προσέγγιση αγνοεί την κατανομή των όρων που εμφανίζονται στο ερώτημα
- Εύκολη η προσθήκη τους εάν το ευρετήριο είναι σχετικά στατικό, αλλά δύσκολο εάν το L αλλάζει διαρκώς εξαιτίας ενημερώσεων του ευρετηρίου
- Σίγουρα βοηθά η προσθήκη τους: Στο σύγχρονο hardware ίσως όχι
  - Το κόστος φόρτωσης μιας μεγαλύτερης λίστας postings μπορεί να ισορροπήσει ή να υπερκεράσει το όφελος από μια ταχύτερη εντός-μνήμης συγχώνευση