

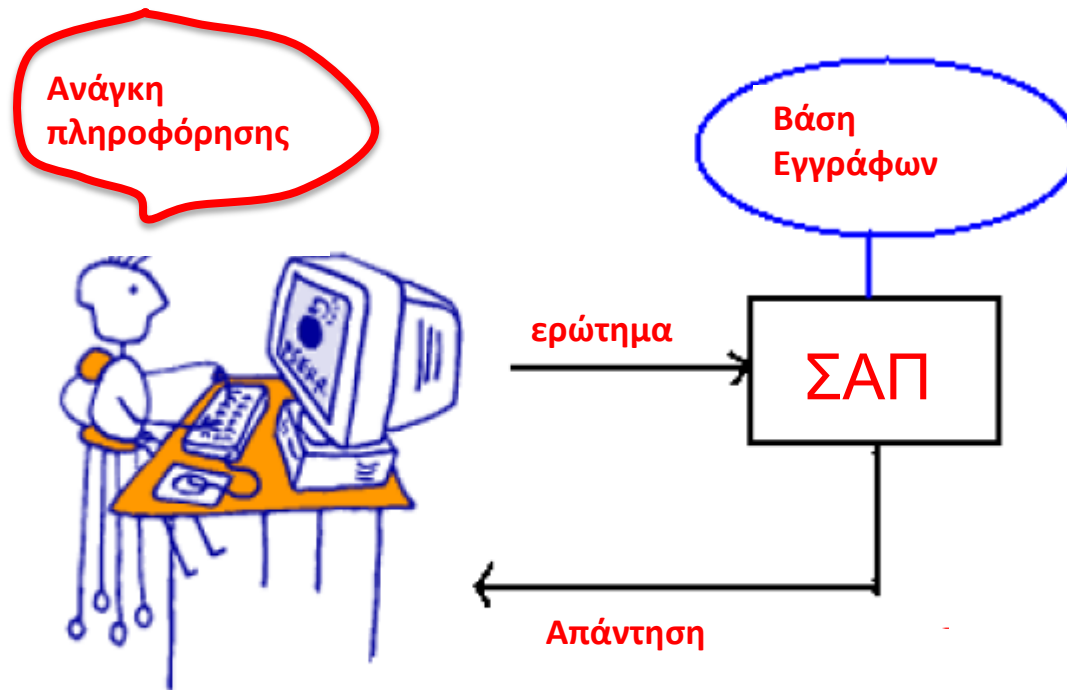
# Introduction to Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

*Διδάσκουσα: Ευαγγελία Πιτουρά*

Διάλεξη 2: Λεξιλόγιο Όρων και Λίστες Καταχωρήσεων

# Τι είναι η «Ανάκτηση Πληροφορίας»;



# Βασικά Βήματα

---

- Συλλογή των δεδομένων
- Ανάγκη πληροφόρησης, ερώτημα
- Μοντέλο (πχ Boolean) – διαβάθμιση
- Αξιολόγηση (recall-precision)
- Προ-επεξεργασία
  - Εύρεση των token, Γλωσσική επεξεργασία, Κατασκευή ευρετηρίου

# Βασικά Θέματα

---

- Ανάλυση Συνδέσεων (π.χ., pagerank) και μηχανές αναζήτησης
- Ομαδοποίηση και κατηγοριοποίηση
- Συστάσεις, επέκταση ερωτημάτων

# Επανάληψη

- Βασικά αντεστραμμένα ευρετήρια

- Δομή: Λεξικό και οι Λίστες Καταχωρήσεων

BRUTUS → 

1	2	4	11	31	45	173	174
---	---	---	----	----	----	-----	-----

CAESAR → 

1	2	4	5	6	16	57	132	...
---	---	---	---	---	----	----	-----	-----

CALPURNIA → 

2	31	54	101
---	----	----	-----

- Βασικό σημείο στην κατασκευή τους: Διάταξη (Sorting)

- Boolean επεξεργασία ερωτήσεων

- Τομή με βάση γραμμικού χρόνου συγχώνευση (merging)
- Απλές βελτιστοποιήσεις

# Λεξιλόγιο Όρων και Λίστες Καταχωρήσεων

# Τι θα δούμε σήμερα;

---

- Προ-επεξεργασία για τη δημιουργία του λεξιλογίου όρων
  - Έγγραφα
  - Tokenization – Μονάδες εγγράφου
  - Ποιους όρους να εισάγουμε στο ευρετήριο;
- Καταχωρήσεις
  - Γρηγορότερη συγχώνευση: *Λίστες Παράβλεψης* (skip lists)
  - Λίστες καταχωρήσεων με πληροφορίες θέσεων (Positional postings) και ερωτήματα φράσεων (phrase queries)

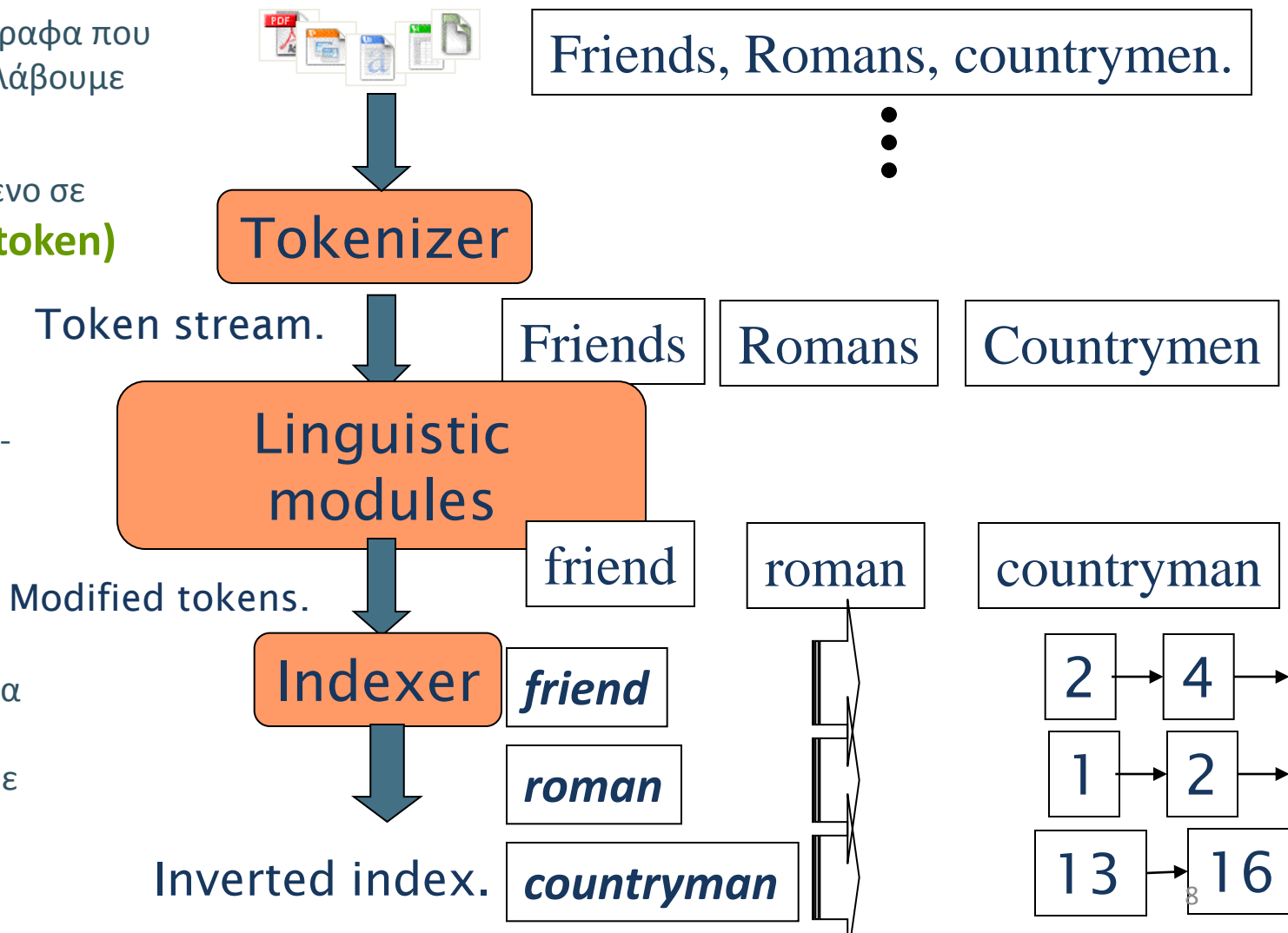
# Τα βασικά βήματα για την κατασκευή του ευρετηρίου

1. Συλλέγουμε τα έγγραφα που θέλουμε να συμπεριλάβουμε στο ευρετήριο

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

3. Γλωσσολογική προεπεξεργασία των συμβόλων

4. Ευρετηριάζουμε τα έγγραφα στα οποία περιλαμβάνεται κάθε όρος





# Parsing

---

Λήψη της ακολουθίας χαρακτήρων ενός εγγράφου

Ποια είναι τα θέματα;

- Σε τι format?

- pdf/word/excel/html ή και zip

Αν σε δυαδική μορφή - χρήση αποκωδικοποιητή (decoder)  
ώστε ακολουθία χαρακτήρων

- Σε ποια γλώσσα?

- Σε διαφορετικές κωδικοποιήσεις (σύνολο χαρακτήρων/character set)

- Π.χ., UTF-8

# Parsing

---

- Να αγνοήσουμε τα ειδικά σύμβολα (mark up)
  - JSON, XML
  - &amp; -> & (XML)

# Complications: Format/language

---

- Τα έγγραφα για τα οποία κατασκευάζουμε το ευρετήριο μπορεί να είναι γραμμένα σε διαφορετικές γλώσσες το καθένα
  - Στο ίδιο ευρετήριο μπορεί να υπάρχουν όροι από πολλές γλώσσες
- Πολλαπλές γλώσσες/format μπορεί να εμφανίζονται και σε ένα έγγραφο ή στα τμήματά του
  - *French email στα Γαλλικά με pdf attachment στα Γερμανικά.*

❖ Πως θα το καταλάβουμε;

Πρόβλημα ταξινόμησης (classification) αλλά στην πράξη συνήθως επιλογή από το χρήστη, χρήση μεταδεδομένων αρχείου κλπ

# Complications: όχι απαραίτητα σειριακά

---

ك ت ا ب ← كِتَابٌ  
un b ā t i k  
/kitābun/ 'a book'

Αραβικά: δισδιάστατη ακολουθία χαρακτήρων και χαρακτήρες σε μεικτή σειρά

Η αντίστοιχη **ακουστική** γραμμική ακολουθία

# Μονάδα εγγράφου

---

*Ποια θεωρείται η μονάδα εγγράφου που βάζουμε στο ευρετήριο;*

- Ένα αρχείο;
- Ένα email; (από τα πολλά στο mbox.)
- Ένα email με 5 συνημμένα έγγραφα (attachments); Αν το 1 συνημμένο σε μορφή zip;
- Ανάποδα: εργαλεία χωρίζουν ένα αρχείο σε πολλά, (PPT ή LaTeX σε πολλαπλές HTML σελίδες)

*Αναλυτικότητα ευρετηριοποίησης (indexing granularity)*

Π.χ., ποια πληροφορία για ένα βιβλίο έχουμε στο ευρετήριο (σε επίπεδο κεφαλαίου, παραγράφου, πρότασης;)

Ακρίβεια/ανάκληση

# **ΣΥΜΒΟΛΑ (TOKENS) ΚΑΙ ΟΡΟΙ (TERMS)**

# Tokenization – Διαίρεση σε Σύμβολα

---

- Είσοδος: “*Friends, Romans, Countrymen*”
- Έξοδος: Tokens
  - *Friends*
  - *Romans*
  - *Countrymen*
- Ένα σύμβολο (**token**) είναι μια ακολουθία από χαρακτήρες σε ένα κείμενο (που είναι ομαδοποιημένοι ως μια χρήσιμη σημασιολογικά μονάδα)
- Κάθε τέτοιο token είναι υποψήφιο για να εισαχθεί στο ευρετήριο μετά από περαιτέρω επεξεργασία

# Tokenization – Διαίρεση σε Σύμβολα

---

**Token** (λεκτική μονάδα)

**Type** (τύπος) μία ομάδα από tokens που αποτελείται από την ίδια ακολουθία χαρακτήρων

**Term** (όρος) συχνά κανονικοποιημένος τύπος που εισάγεται στο ευρετήριο του συστήματος



# Tokenization – Διαίρεση σε Σύμβολα

---

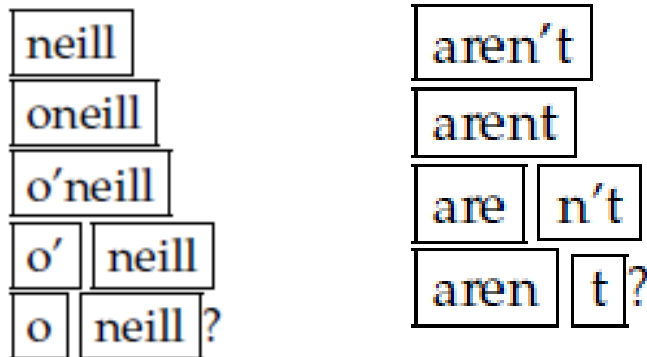
***Αλλά ποια είναι τα κατάλληλα tokens;***

*Αρκεί να χωρίσουμε το κείμενο στα κενά και στα σημεία στίξης;*

*Εξαρτάται από τη γλώσσα*

# Tokenization: Θέματα

- Αγγλικά: απόστροφος (σύντμηση και γενική κτητική)
  - *Finland's capital* → *Finland? Finlands? Finland's?*
  - *Mr. O' Neill thinks that the boys' stories about Chile's capital aren't amusing*



- καθορίζουν ποιες *Boolean* ερωτήσεις θα απαντούν  
*Την ίδια πολιτική και στην ερώτηση και στο κείμενο*

# Tokenization: Θέματα

---

- Ενωτικό (hyphen):
    - *Hewlett-Packard* → *Hewlett* και *Packard* ως δύο tokens ?
    - *state-of-the-art* ή *the-hold-him-back-and-drag-him-away maneuver* (να διασπάσουμε την ακολουθία;)
    - *co-education* (χωρισμός φωνηέντων)
    - *lowercase, lower-case, lower case* ?
  - Διάσπαση στο κενό σύμβολο
    - *San Francisco, Los Angeles York University vs New York University* (διάσπαση ονομάτων) αλλά πως μπορούμε να το καταλάβουμε;
  - Ή και συνδυασμός
    - *San Francisco-Los Angeles*
- ✓ Την ίδια πολιτική και στην ερώτηση και στο κείμενο
  - ✓ Χρήστες πάντα το – όταν θέλουν να εξεταστούν όλες οι περιπτώσεις
  - ✓ Φράσεις (πχ *lower, case, lowercase*)

# Tokenization: Αριθμοί

---

- *3/12/91*                      *Mar. 12, 1991*                      *12/3/91*
- *55 B.C.*
- *B-52*
- *My PGP key is 324a3df234cb23e*
- *(800) 234-2333*
  - Συχνά περιέχουν ενδιάμεσα κενά
  - Τα παλιότερα συστήματα μπορεί να μη έβαζαν στο ευρετήριο τους αριθμούς
    - Συχνά όμως είναι χρήσιμοι, πχ αναζήτηση για κώδικες λάθους error codes/stacktraces στο web, IP διευθύνσεις, package tracking numbers
    - (Χρήση n-grams)
  - Ευρετηριοποίηση των μεταδεδομένων ξεχωριστά
    - Ημερομηνία δημιουργίας, format, κλπ

# Tokenization

---

- Επίσης ειδικές λέξεις
  - M\*A\*S\*H
  - C++
  - C#

# Tokenization: άλλες γλώσσες

---

- Γαλλικά
  - *L'ensemble* → (σύντμηση άρθρου)
    - *L ? L' ? Le ?*
    - Θα θέλαμε τα *l'ensemble* να ταιριάζει με το *un ensemble*
      - Έως το 2003, δεν το υποστήριζε το Google
        - **Internationalization!**
- Γερμανικά (οι σύνθετες λέξεις δεν διαχωρίζονται)
  - *Lebensversicherungsgesellschaftsangestellter* (life insurance company employee)
  - Τα Γερμανικά συστήματα ανάκτησης πληροφορίας χρησιμοποιούν μια μονάδα **compound splitter**
    - Βελτίωση της απόδοσης κατά 15%

# Tokenization: άλλες γλώσσες

---

莎拉波娃现在居住在美国东南部的佛罗里达。今年4月9日，莎拉波娃在美国第一大城市纽约度过了18岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Κινέζικα: δεν υπάρχουν κενά

# Tokenization: άλλες γλώσσες

---

- Τα Κινέζικα και τα Ιαπωνικά δεν έχουν κενούς χαρακτήρες ανάμεσα στις λέξεις:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。

## Χωρισμός σε λέξεις (word segmentation)

- Διάφορες τεχνικές: χρήση λεξικού και ταίριασμα της μεγαλύτερης ακολουθίας, μηχανική μάθηση

Αλλά δεν υπάρχει πάντα μια μοναδική tokenization



# Tokenization: άλλες γλώσσες

---

和尚

Κινέζικα: είτε ως ακολουθία δύο λέξεων “and” και “still” ή ως μια λέξη “monk”

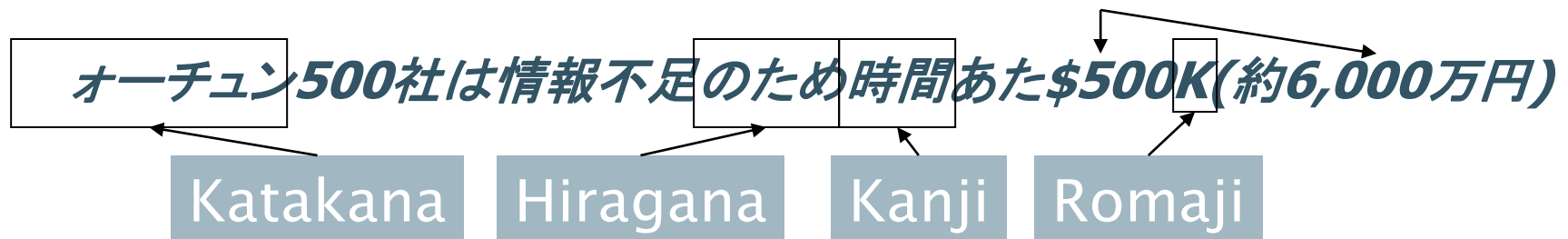
# Tokenization: άλλες γλώσσες

---

Ευρετηριοποίηση όλων των ακολουθιών  
χαρακτήρων (k-grams)

# Tokenization: άλλες γλώσσες

- Ακόμα πιο δύσκολο στα Ιαπωνικά, ανάμιξη πολλαπλών αλφάβητων
  - Ημερομηνίες/ποσά σε πολλά formats



Ο χρήστης μπορεί να διατυπώσει την ερώτηση μόνο σε hiragana!

# Tokenization: άλλες γλώσσες

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるMOTTAINAIキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを800字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて10月20日までにお送りください。大賞受賞者には、50万円相当の旅行券とエコ製品2点の副賞が贈られます。

Γιαπωνέζικα - 4 διαφορετικά “αλφάβητα”:

*Chinese characters, hiragana syllabary for inflectional endings and functional words, katakana syllabary for transcription of foreign words and other uses, and latin. No spaces (as in Chinese).*

*End user can express query entirely in hiragana!*

# Tokenization: άλλες γλώσσες

- Τα Αραβικά και στα Εβραϊκά γράφονται από τα δεξιά προς τα αριστερά, αλλά με συγκεκριμένα τμήματα (πχ αριθμοί) να γράφονται από τα αριστερά στα δεξιά
- Οι λέξεις διαχωρίζονται αλλά τα γράμματα μέσα στις λέξεις περίπλοκοι χαρακτήρες

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 χρήση Unicode, η αποθηκευμένη μορφή είναι απλοποιημένη

# Stop words (Διακόπτουσες λέξεις)

---

- Χρήση stop list, αποκλείουμε από το λεξικό τις πιο κοινές λέξεις. Γιατί;
  - Έχουν μικρό σημασιολογικό περιεχόμενο: *a, an, and, are, as, at, be, by, for, from, has, he, in, is, it, its, of, on, that, the, to, was, were, will, with*
  - Είναι πάρα πολλές: ~30% των καταχωρήσεων αφορούν τις πιο συχνές 30 λέξεις

# Stop words (Διακόπτουσες λέξεις)

---

- Ωστόσο η τάση είναι να μη χρησιμοποιούνται λίστες:
  - Καλές τεχνικές συμπίεσης οδηγούν στο να ελαχιστοποιούν το χώρο που χρειάζεται για την αποθήκευση τους
  - Καλές τεχνικές για την επεξεργασία ερωτημάτων (βάρη όρων) μειώνουν το κόστος στην εκτέλεση μιας ερώτησης εξαιτίας των stop words.

Είναι χρήσιμα για:

- Φράσεις: “King of Denmark”
- Τίτλους τραγουδιών, κλπ.: “Let it be”, “To be or not to be”
- “Σχεσιακά” ερωτήματα: “flights to London”

# Κανονικοποίηση (Token normalization)

---

- Χρειάζεται να «κανονικοποιήσουμε» τις λέξεις στο κείμενο αλλά και στις ερωτήσεις στην ίδια μορφή
  - Π.χ, θέλουμε να ταιριάξουμε το **U.S.A.** ακαι **USA**
- Το αποτέλεσμα είναι ένας **όρος** (term).
  - Ένας όρος είναι ένα (κανονικοποιημένος) τύπος λέξης που εισάγεται στο Λεξικό του Συστήματος Ανάκτησης Πληροφορίας



# Κανονικοποίηση (Token normalization)

---

- Συχνά ορίζουμε έμμεσα (*equivalence classes*) για τους όρους, π.χ.,
  - Σβήνουμε τις τελείες από έναν όρο
    - *U.S.A., USA ( USA*
  - Σβήνουμε τα ενωτικά από έναν όρο *anti-discriminatory, antidiscriminatory*

Αρκούν απλοί κανόνες αντιστοίχισης

Μερικές φορές δεν είναι εύκολο να εντοπιστεί πότε χρειάζεται προσθήκη χαρακτήρων

# Κανονικοποίηση σε όρους

---

- Μια εναλλακτική προσέγγιση στις λίστες ισοδυναμίας είναι να κρατάμε όλα τα μη κανονικοποιημένα *token*
    - (ή και «Χειροποίητες» λίστες συνωνύμων )
  - 1. Διεύρυνση κατά την ερώτηση (διάζευξη)  
Enter: *windows* Search: *Windows, windows, window*  
Enter: *car* Search: *car automobile*
  - 2. Εναλλακτικά, καταχωρούμε το έγγραφο στις λίστες καταχώρησης κάθε συνώνυμου (πχ έγγραφο που περιέχει το *car* καταχωρείται και στο *automobile*)
- Το 1 ή το 2 είναι καλύτερο;

# Κανονικοποίηση σε όρους

---

- Μη συμμετρική διεύρυνση
- Ένα παράδειγμα όπου αυτό μπορεί να φανεί χρήσιμο
  - Enter: *window*      Search: *window, windows*
  - Enter: *windows*      Search: *Windows, windows, window*
  - Enter: *Windows*      Search: *Windows*
- Θεωρητικά πιο ισχυρό από τις λίστες αλλά λιγότερο αποδοτικό

# Κανονικοποίηση: άλλες γλώσσες

60% ιστοσελίδων στα Αγγλικά (2007) – 1/3 των χρηστών του διαδικτύου - 10% του παγκόσμιου πληθυσμού μιλούν Αγγλικά

- Accents: π.χ., Γαλλικά *résumé* vs. *resume*.
- Umlauts: π.χ., Γερμανικά: ***Tuebingen*** vs. ***Tübingen***
  - Πρέπει να είναι ισοδύναμα
- Πιο βασικό κριτήριο:
  - Πως προτιμούν οι χρήστες να γράφουν αυτές τις λέξεις στα ερωτήματά τους
- Ακόμα και σε γλώσσες που έχουν accents, οι χρήστες δεν τα πληκτρολογούν
  - Οπότε συχνά είναι καλύτερο να κανονικοποιούμε ή να αφαιρούμε το accent από ένα όρο
    - ***Tuebingen, Tübingen, Tubingen \ Tubingen***

# Κανονικοποίηση: άλλες γλώσσες

- Κανονικοποίηση σε περιπτώσεις όπως οι ημερομηνίες
  - **7月30日 vs. 7/30**
  - **Japanese use of kana vs. Chinese characters**
- Tokenization και οι κανονικοποίηση μπορεί να εξαρτάται από τη γλώσσα όποτε μαζί με αναγνώριση γλώσσας

*Morgen will ich in MIT ...*

Is this  
German “mit

- Βασικό: Πρέπει το κείμενο που θα ευρετηριοποιηθεί και οι όροι στο ερώτημα να κανονικοποιούνται με τον ίδιο τρόπο

# Μετατροπή σε κεφαλαία/μικρά

- Μετατροπή όλων των γραμμάτων σε μικρά
  - εξαίρεση: κεφαλαία στη μέση της πρότασης;
    - e.g., *General Motors*
    - *Fed vs. fed*
    - *Bush vs. bush*
  - Πρακτικά μετατροπή όλων σε μικρά, αφού συχνά οι χρήστες χρησιμοποιούν μικρά ανεξάρτητα της «σωστής» χρήσης των κεφαλαίων Παράδειγμα από τη Google:
- Δοκιμάστε την ερώτηση C.A.T.
  - #1 αποτέλεσμα για “cat”



# Θησαυροί (Thesauri) και soundex

- Πως χειριζόμαστε τα συνώνυμα και τα ομώνυμα;
  - Π.χ., κατασκευάζοντας **λίστες ισοδυναμίας** με το χέρι
    - *car = automobile*    *color = colour*
  - Μπορούμε να το ξαναγράψουμε (rewrite) για να δημιουργήσουμε κλάσεις ισοδυναμίας όρων
    - Καταχωρούμε το έγγραφο στις λίστες καταχώρησης κάθε συνώνυμου (πχ έγγραφο που περιέχει το *car* καταχωρείται και στο *automobile* και το ανάποδο)
  - Ή να διευρύνουμε το ερώτημα
    - Όταν το ερώτημα περιέχει *automobile*, ψάξε και για το *car*
- Τι γίνεται με τα ορθογραφικά λάθη (spelling mistakes)?
  - Μια προσέγγιση είναι το soundex, που σχηματίζει κλάσεις ισοδυναμίας από λέξεις βασιζόμενες σε ακουστικούς ευριστικούς κανόνες phonetic heuristics

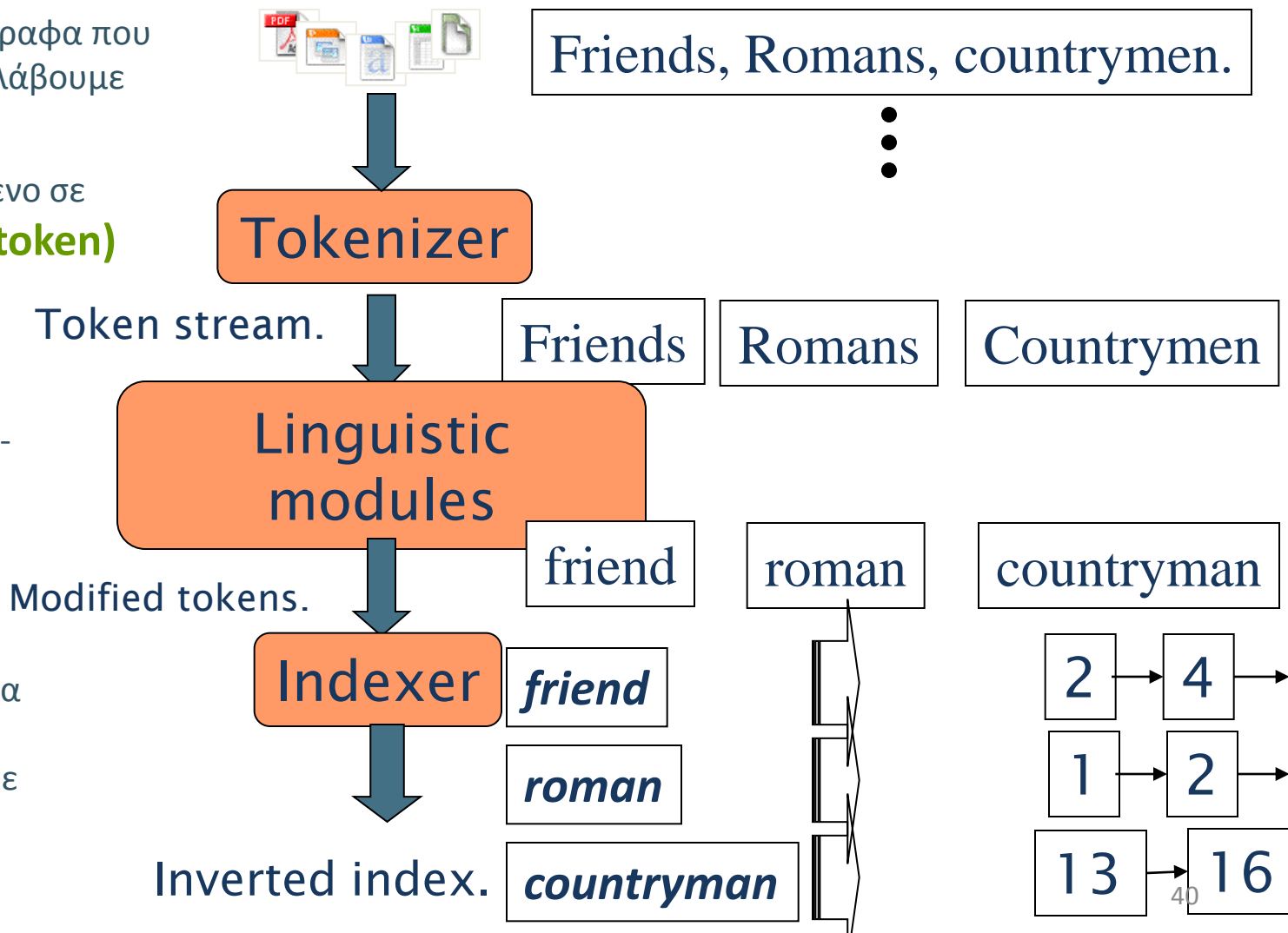
# Τα βασικά βήματα για την κατασκευή του ευρετηρίου

1. Συλλέγουμε τα έγγραφα που θέλουμε να συμπεριλάβουμε στο ευρετήριο

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

3. Γλωσσολογική προεπεξεργασία των συμβόλων

4. Ευρετηριάζουμε τα έγγραφα στα οποία περιλαμβάνεται κάθε όρος





# Λημματοποίηση (Lemmatization)

---

- Περικοπή κλιτικών καταλήξεων και αναγωγή παράγωγων μορφών μιας λέξης σε κοινή 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 λημματοποίηση προϋποθέτει «ορθή» αναγωγή που χρησιμοποιεί λεξιλόγιο και μορφολογική ανάλυση των λέξεων και επιστρέφει τη βασική μορφή της λέξης, το λήμμα

# Stemming (Περιστολή)

- Αναγωγή των όρων στις ρίζες του πριν την εισαγωγή τους στο ευρετήριο
- “Stemming” υπονοεί ωμό κόψιμο των καταλήξεων
  - εξαρτάται από τη γλώσσα
  - π.χ., *automate(s)*, *automatic*, *automation* όλα ανάγονται 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 διαδεδομένος αλγόριθμος stemming για τα Αγγλικά
  - Τα αποτελέσματα δείχνουν ότι είναι τουλάχιστον τόσο καλός όσο οι άλλες επιλογές
- Συμβάσεις + 5 φάσεις περικοπών
  - Οι φάσεις εφαρμόζονται διαδοχικά
  - Κάθε φάση αποτελείται από ένα σύνολο κανόνων
- Παράδειγμα σύμβασης: Επιλογή εκείνου του κανόνα από κάθε ομάδα που μπορεί να εφαρμοστεί στο μεγαλύτερο επίθεμα .

# Χαρακτηριστικοί κανόνες του Porter

- *sses* → *ss*
  - *ies* → *i*
  - *ational* → *ate*
  - *tional* → *tion*
- Παράδειγμα
- caresses → caress
  - ponies → poni
  - caress → caress
  - cats → cat
- Οι κανόνες χρησιμοποιούν ένα είδους μέτρου (*measure*) που ελέγχει το πλήθος των συλλαβών
  - ( $m > 1$ ) *ELEMENT* →
    - *replacement* → *replac*
    - *cement* → *cement*

[www.tartarus.org/~martin/PorterStemmer](http://www.tartarus.org/~martin/PorterStemmer)

# Άλλοι stemmers

---

- Υπάρχουν και άλλου π.χ., Lovins stemmer
  - <http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm>
  - Ένα πέρασμα, αφαίρεση της μεγαλύτερης κατάληξης (περίπου 250 κανόνες)
- Πλήρη μορφολογική ανάλυση – περιορισμένα οφέλη
- Βοηθά το stemming και οι άλλοι κανονικοποιητές;
  - English: ανάμικτα αποτελέσματα. Βοηθά την ανάκληση αλλά βλάπτει την ακρίβεια
    - operative (dentistry) ⇒ oper
    - operational (research) ⇒ oper
    - operating (systems) ⇒ oper
- Οπωσδήποτε χρήσιμο για Ισπανικά, Γερμανικά, Φινλανδικά
  - 30% βελτίωση για τα Φινλανδικά

# Άλλοι stemmers: σύγκριση

---

- Sample text:* Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation
- Porter stemmer:* such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation
- Lovins stemmer:* such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation
- Paice stemmer:* such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

# Εξάρτηση από τη γλώσσα

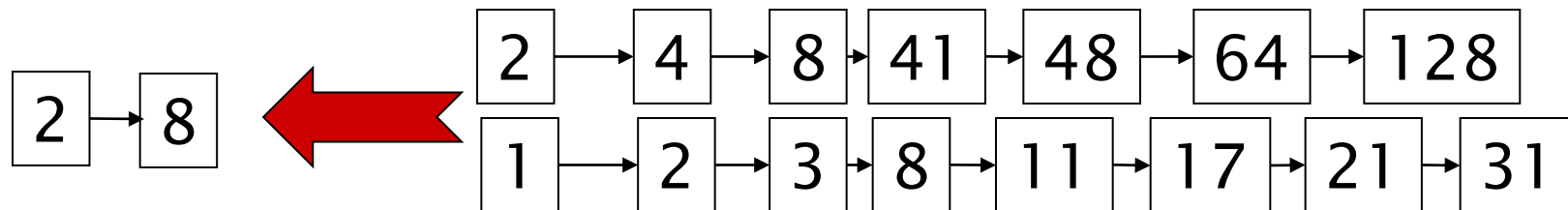
---

- Πολλά από τα παραπάνω περιλαμβάνουν μετασχηματισμούς που
  - Εξαρτώνται από τη γλώσσα και
  - Συχνά από την εφαρμογή
- Με τη μορφή “plug-in” πριν τη διαδικασία δεικτοδότησης
- Ελεύθερου λογισμικού και εμπορικά

# ΛΙΣΤΕΣ ΚΑΤΑΧΩΡΗΣΕΩΝ

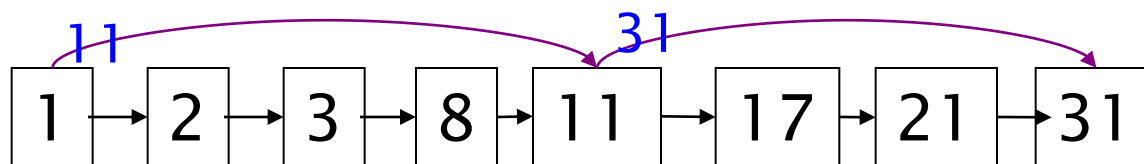
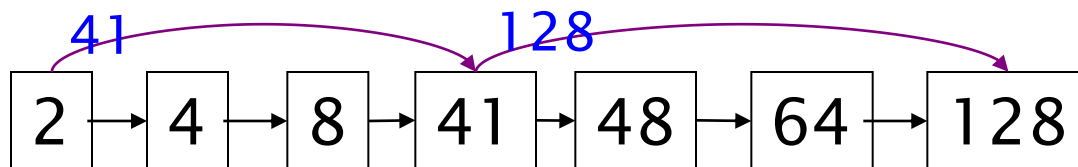


# Βασική συγχώνευση



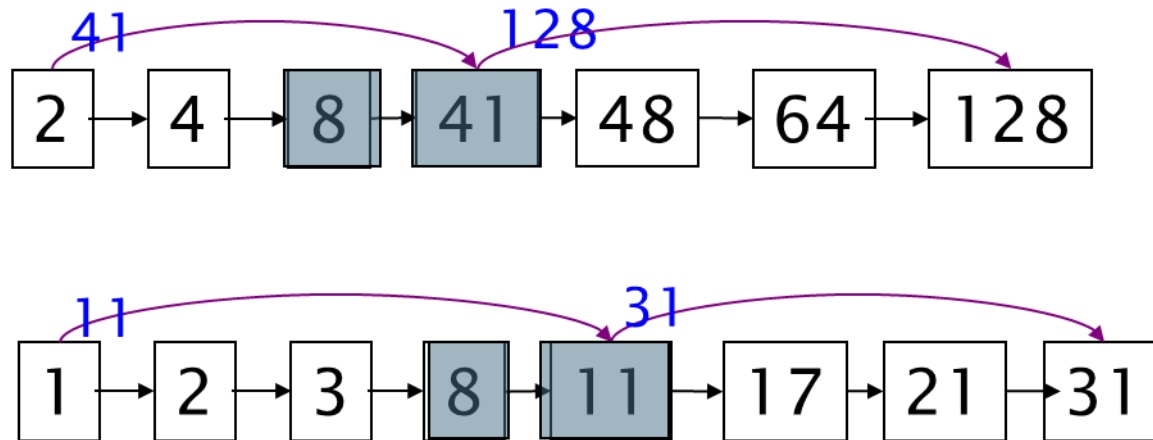
Αν τα μήκη των λιστών είναι  $m$  και  $n$ ,  $O(m+n)$

# Επέκταση των λιστών με δείκτες παράλειψης *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

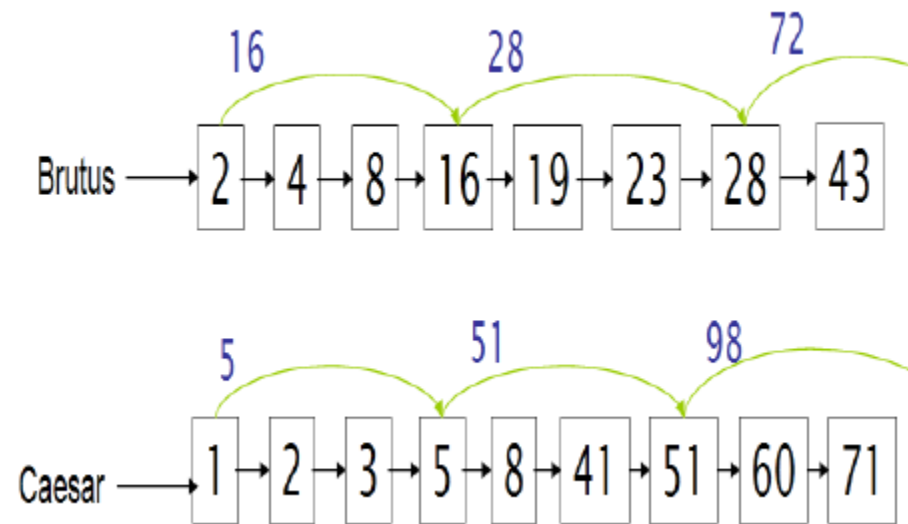


Υποθέστε ότι έχουμε διατρέξει τις λίστες και έχουμε βρει το κοινό στοιχείο **8** σε κάθε λίστα, το ταιριάζουμε και προχωράμε

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

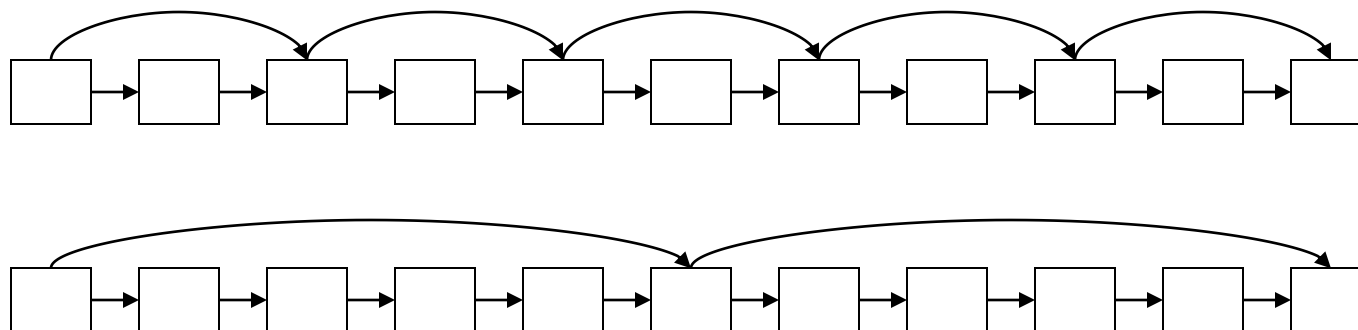
Αλλά ο δείκτης παράλειψης του **11** είναι το **31**, οπότε μπορούμε να παραβλέψουμε τις ενδιάμεσες καταχωρήσεις

# Επεξεργασία ερωτήματος με skip pointers



# 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  $\sqrt{L}$  ισαπέχοντες δείκτες παράβλεψης.
- Αγνοεί την κατανομή των όρων της ερώτησης.
- Εύκολο αν το ευρετήριο είναι σχετικά στατικό; Δύσκολο αν το  $L$  αλλάζει συνεχώς λόγω τροποποιήσεων .
- Βοηθούσε (λόγω πιο αργής CPU); Όχι τόσο με το νέο υλικό εκτός αν memory-based
  - Το I/O κόστος για να φορτωθεί μια μεγαλύτερη λίστα καταχωρήσεων μπορεί να υπερβαίνει το κέρδος από τη γρηγορότερη συγχώνευση

# ΕΥΡΕΤΗΡΙΑ ΦΡΑΣΕΩΝ

# Ερωτήματα Φράσεων

---

- Θέλουμε να μπορούμε να απαντάμε σε ερωτήματα όπως “*stanford university*” – ως φράση
- Οπότε η φράση “*I went to university at Stanford*” δεν αποτελεί ταίριασμα.
  - Η έννοια των ερωτήσεων 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 έμμεσες ερωτήσεις φράσεις
- Για να τις 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.



false positives!

# 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
- 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***

# Θέματα

---

- False positives
- 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 tokens του:

<***term***, number of docs containing ***term***;

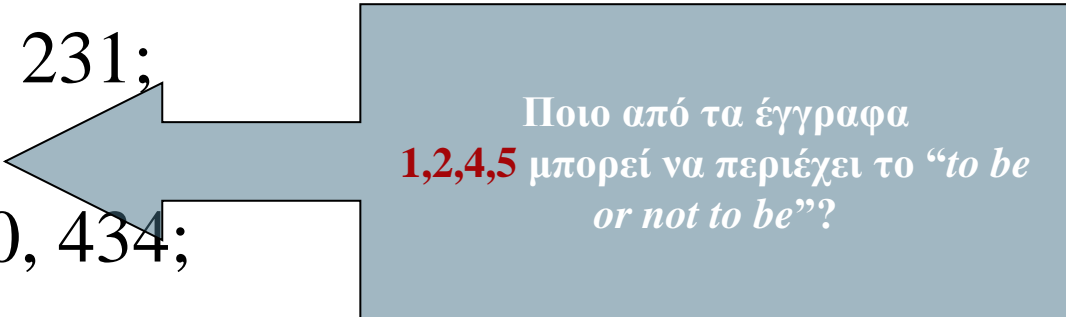
*doc1*: position1, position2 ... ;

*doc2*: position1, position2 ... ;

etc.>

# Παράδειγμα

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



Ποιο από τα έγγραφα  
**1,2,4,5** μπορεί να περιέχει το “*to be  
or not to be*”?

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

# Επεξεργασία ερωτήματος φράσης

---

- Βρες τις εγγραφές του ευρετηρίου για κάθε διαφορετικό όρο: ***to, be, or, not***.
- Συγχώνευσε τις *doc:position* λίστες για απαρίθμηση όλων των πιθανών θέσεων του “***to be or not to be***”.
  - ***to:***
    - 2:1,17,74,222,551; **4:8,16,190,429,433**; 7:13,23,191; ...
  - ***be:***
    - 1:17,19; **4:17,191,291,430,434**; 5:14,19,101; ...
- Η ίδια γενική μέθοδος για ερωτήματα γειτονικότητας (proximity searches)

# Ερωτήματα γειτονικότητας (Proximity queries)

---

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



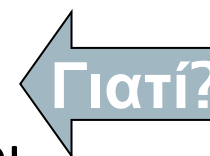
# Μέγεθος ευρετηρίου

---

- Μπορούμε να συμπίεσουμε τα position values/offsets
- Παρόλα αυτά, σημαντική αύξηση του χώρου αποθήκευση 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 ανά έγγραφο ,
- Το μέγεθος του ευρετηρίου εξαρτάται από το μέσο μέγεθος του αρχείου
  - Μέσο μέγεθος web σελίδας <1000 όροι
  - SEC filings, books, even some epic poems ... πάνω από 100,000 όρους
- Έστω ένας όρος με συχνότητα 0.1%



Document size	Postings	Positional postings
1 000	1	1
1 00,000	1	1 00

# Rules of thumb

---

- Ένα ευρετήριο θέσεων είναι 2–4 μεγαλύτερο από ένα απλό ευρετήριο
- Το μέγεθος του 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”**) οι συνεχείς συγχώνευση καταχωρήσεων ευρετηρίου θέσεων δεν είναι αποδοτική
    - Ακόμα περισσότερο για φράσεις όπως **“The Who”**

---

## ΤΕΛΟΣ 2<sup>ου</sup> Μαθήματος

### Ερωτήσεις?

*Χρησιμοποιήθηκε κάποιο υλικό των:*

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