

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΑΝΟΙΚΤΑ ΑΚΑΔΗΜΑΪΚΑ ΜΑΘΗΜΑΤΑ

ανοικτάμαθήματα
opencourses

Ανάκτηση Πληροφορίας

Λεξιλόγιο Ορων και Λίστες Καταχωρίσεων

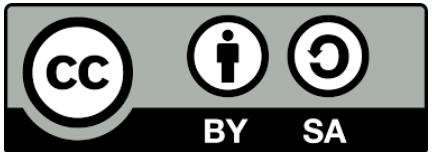
Διδάσκων: Καθ. Ε. Πιτουρά

CC BY SA

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Αδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Introduction to

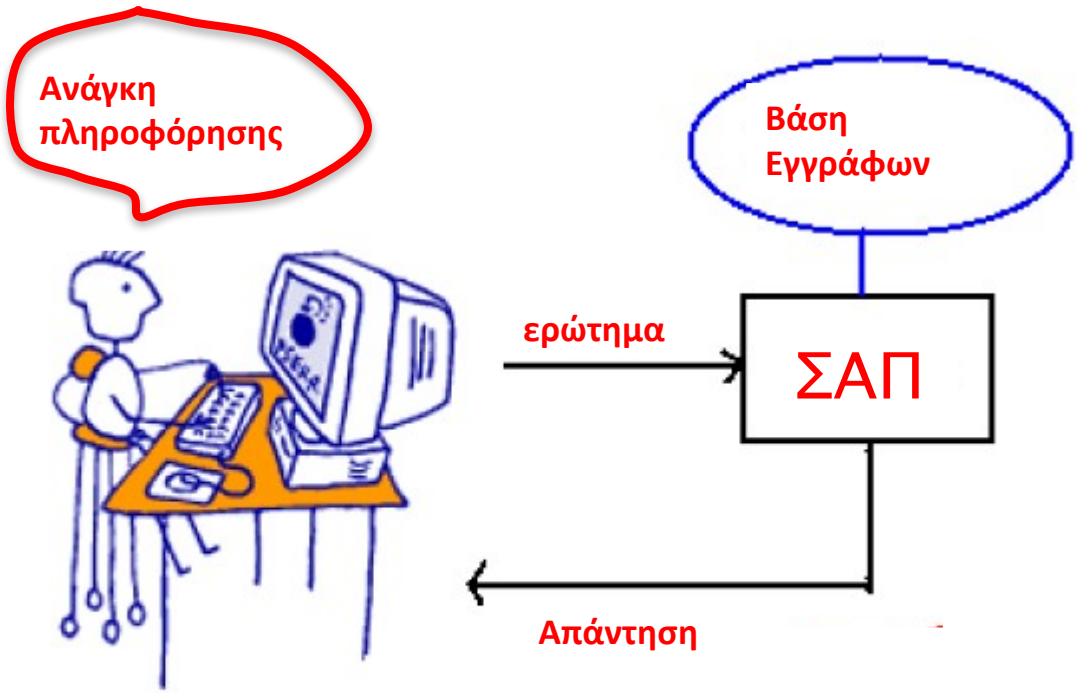
Information Retrieval

ΠΛΕ70: Ανάκτηση Πληροφορίας

Διδάσκουσα: Ευαγγελία Πιτουρά

Διάλεξη 2: Λεξιλόγιο Ορων και Λίστες Καταχωρίσεων

Τι είναι η «Ανάκτηση Πληροφορίας»;



Βασικά Βήματα

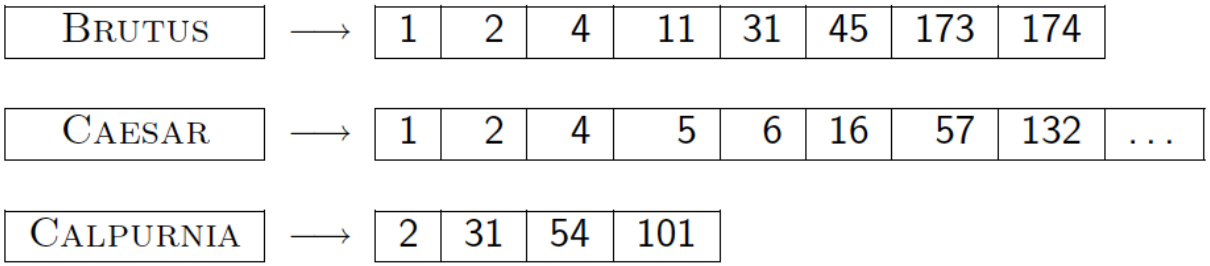
- Συλλογή των δεδομένων
- Ανάγκη πληροφόρησης, ερώτημα
- Μοντέλο (πχ Boolean) – διαβάθμιση
- Αξιολόγηση (recall-precision)
- Προ-επεξεργασία
  - Εύρεση των token, Γλωσσική επεξεργασία, Κατασκευή ευρετηρίου.

Βασικά Θέματα

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

Επανάληψη

- Βασικά ανεστραμμένα ευρετήρια
  - Δομή: Λεξικό και οι Λίστες Καταχωρίσεων



  - Βασικό σημείο στην κατασκευή τους: Διάταξη (Sorting).

- Boolean επεξεργασία ερωτήσεων
  - Τομή με συγχώνευση γραμμικού χρόνου (merging).
  - Απλές βελτιστοποιήσεις.

Λεξιλόγιο Ορων *και* Λίστες Καταχωρίσεων

Τι θα δούμε σήμερα;

- Προ-επεξεργασία για τη δημιουργία του λεξιλογίου όρων
  - Εγγραφα.
  - Tokenization – Μονάδες εγγράφου.
  - Ποιους *όρους* να εισάγουμε στο ευρετήριο;
- Καταχωρίσεις
  - Γρηγορότερη συγχώνευση: *Λίστες Παράβλεψης* (skip lists).
  - Λίστες καταχωρίσεων *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 λέξεις **(token).**

3. Γλωσσολογική προ-επεξεργασία των λέξεων.

4. Ευρετηριοποιούμε τα έγγραφα στα οποία περιλαμβάνεται κάθε όρος.



Friends, Romans, countrymen.

Tokenizer

Token stream. Friends Romans Countrymen

Linguistic modules

Modified tokens. friend roman countryman

Indexer

Inverted index.

friend	2 → 4
roman	1 → 2
countryman	13 → 16

Parsing

Λήψη της ακολουθίας χαρακτήρων ενός εγγράφου

Ποια είναι τα θέματα;

- Σε τι format;
  - pdf/word/excel/html ή και zip

  Αν είναι σε δυαδική μορφή - χρήση αποκωδικοποιητή (decoder) ώστε να πάρουμε ακολουθία χαρακτήρων.
- Σε ποια γλώσσα;
- Σε διαφορετικές κωδικοποιήσεις (σύνολο χαρακτήρων/character set)
  - Π.χ., UTF-8.

Parsing

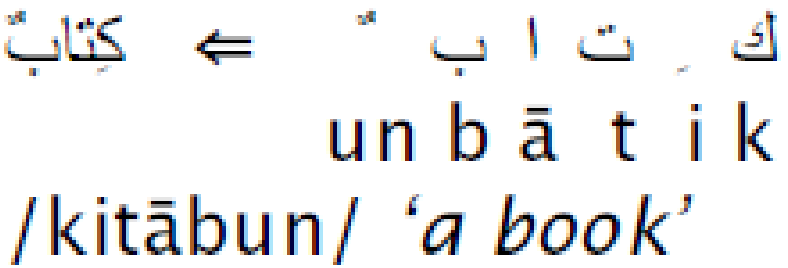
- Να αγνοήσουμε τα ειδικά σύμβολα (mark up)
  - JSON, XML
  - &amp -> & (XML).

Complications: Format/language

- Τα έγγραφα για τα οποία κατασκευάζουμε το ευρετήρια 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 κατηγοριοποίησης (classification) αλλά στην πράξη επιλογή από το χρήστη, χρήση μεταδεδομένων αρχείου κλπ.

Complications: όχι απαραίτητα σειριακά



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;

- Ενα αρχείο;
- Ενα email; (από τα πολλά στο mbox.)
- Ενα email με 5 συνημμένα έγγραφα (attachments); Αν το 1 συνημμένο σε μορφή zip;
- Ανάποδα: εργαλεία χωρίζουνε 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.

ΛΕΞΕΙΣ (TOKENS) ΚΑΙ ΟΡΟΙ (TERMS)

Tokenization – Διαίρεση σε Λέξεις

- Είσοδος: "***Friends, Romans, Countrymen***"
- Εξοδος: Tokens
  - ***Friends***
  - ***Romans***
  - ***Countrymen.***
- Μια λέξη (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 Λέξεις

Word (λέξη) μια ακολουθία χαρακτήρων στο κείμενο

Term (όρος) κανονικοποιημένη μορφή word που εισάγεται στο ευρετήριο του συστήματος. Ορίζει μια ισοδυναμία από words

Token (λεκτική μονάδα) η εμφάνιση μιας word στο κείμενο

Type (τύπος) όπως το term. Ορίζει μια ισοδυναμία από tokens.

Tokenization – Διαίρεση σε Λέξεις

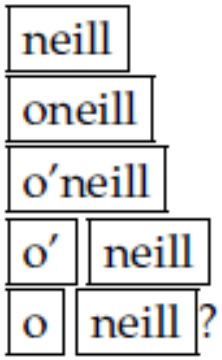
Αλλά ποια είναι τα κατάλληλα tokens;

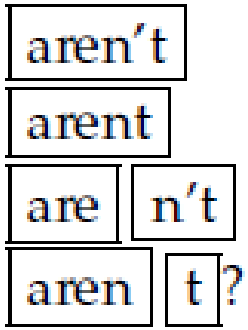
Αρκεί να χωρίσουμε το κείμενο στα κενά και στα σημεία στίξης;

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
✓ Πάντα βάζουμε το – όταν θέλουμε 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 μην 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***
      - Εως 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

- 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 ένα μοναδικό tokenization.

Tokenization: άλλες γλώσσες



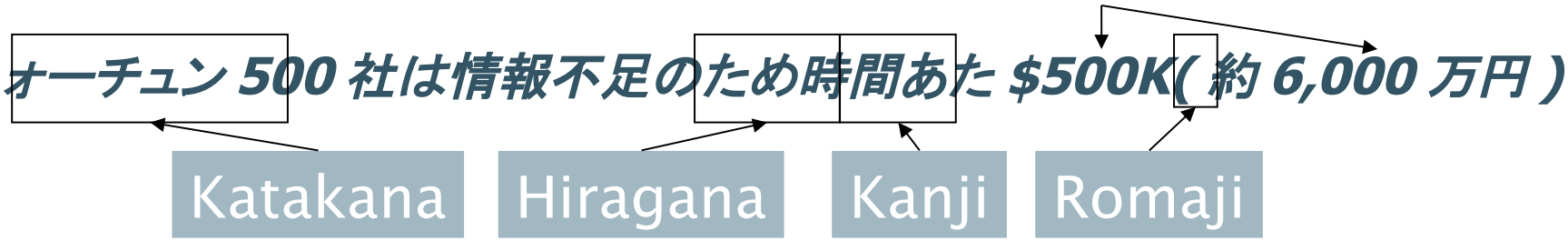
Κινέζικα: είτε ως ακολουθία δύο λέξεων “and” και “still” ή ως μια λέξη “monk”.

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”:

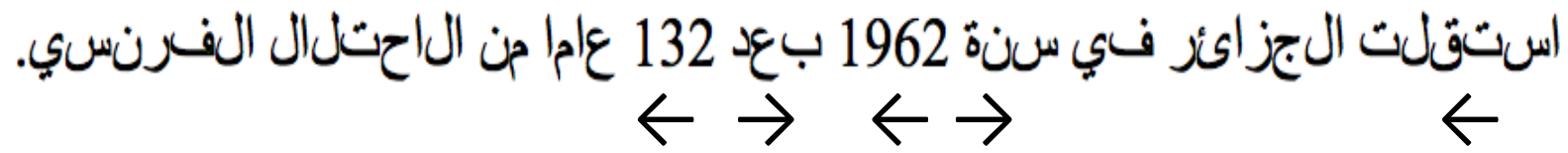
Chinese *characters,* ***hiragana*** *syllabary for inflectional endings and functional words,* ***katakana*** *syllabary for transcription of foreign words and other uses, and* ***latin****. No spaces (as in Chinese).*

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



start

- ‘Algeria achieved its independence in 1962 after 132 years of French occupation.’
- Με χρήση Unicode, η αποθηκευμένη μορφή είναι απλοποιημένη.

Stop words

- Χρήση stop list, αποκλείουμε από το λεξικό τις πιο κοινές λέξεις. Γιατί;
  - Εχουν 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 καταχωρίσεων αφορούν τις πιο συχνές 30 λέξεις.

Stop words

- Ωστόσο η τάση είναι να μη χρησιμοποιούνται τέτοιες 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 εκτέλεσης 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).
  - Ενας όρος είναι ένα (κανονικοποιημένος) τύπος λέξης που εισάγεται στο Λεξικό του Συστήματος Ανάκτησης Πληροφορίας.

Κανονικοποίηση (Token normalization)

- Συχνά ορίζουμε ισοδυναμίες (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
- Ενα 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 κανονικοποίηση μπορεί να εξαρτώνται από τη γλώσσα, όποτε εφαρμόζονται 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 κανονικοποιούνατι 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 εφαρμόζουμε rewriting 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
    - Οταν 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 ευρετικούς 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 λέξεις **(token).**

3. Γλωσσολογική προ-επεξεργασία των λέξεων.

4. Ευρετηριάζουμε τα έγγραφα στα οποία περιλαμβάνεται κάθε όρος.



Friends, Romans, countrymen.

Tokenizer

Token stream. Friends Romans Countrymen

Linguistic modules

Modified tokens. friend roman countryman

Indexer

Inverted index.

friend	2 → 4
roman	1 → 2
countryman	13 → 16

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 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) EMENT* →
  - *replacement* → *replac*
  - *cement* → *cement.*

www.tartarus.org/~martin/PorterStemmer.

Αλλοι stemmers

- Υπάρχουν και άλλοι π.χ., Lovins stemmer
  - http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm
  - Ενα πέρασμα, αφαίρεση της μεγαλύτερης κατάληξης (περίπου 250 κανόνες)
- Πλήρης 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, Φιλανδικά
    - 30% βελτίωση για τα Φινλανδικά.

Αλλοι 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 analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to pictur of express that is more biolog transpar and access to interpret.

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres.

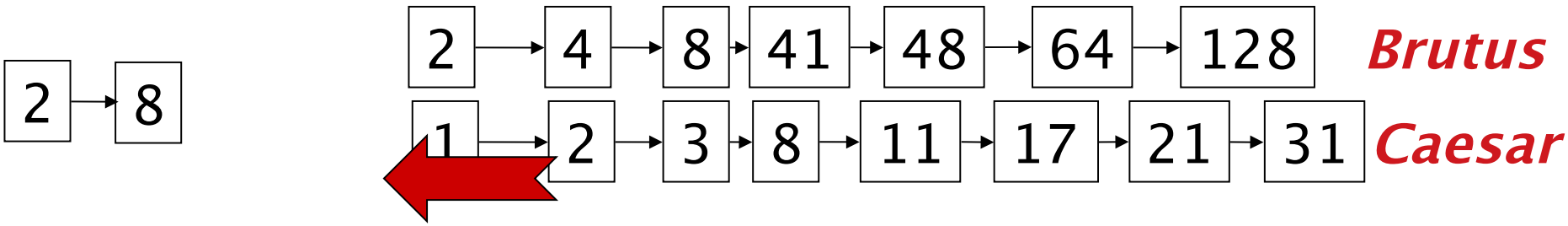
Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret.

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 ευρετηριοποίησης.
- Ελεύθερου λογισμικού και εμπορικά.

ΛΙΣΤΕΣ ΚΑΤΑΧΩΡΙΣΕΩΝ

Βασική συγχώνευση

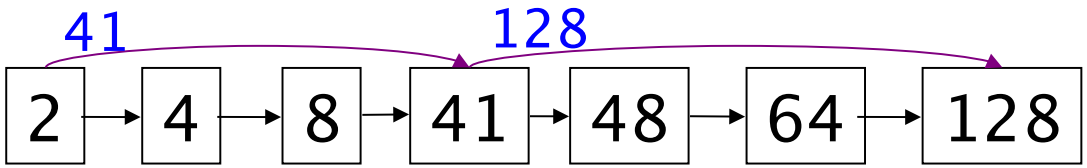


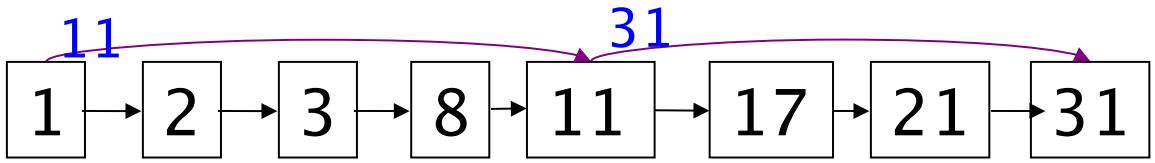
Αν τα μήκη των λιστών είναι m και n, O(*m*+*n*).

Μπορούμε και καλύτερα;
Ναι (αν το ευρετήριο είναι σχετικά στατικό).

Επέκταση των λιστών με δείκτες παράβλεψης

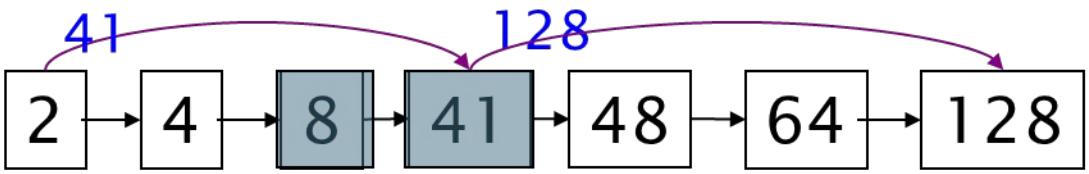
skip pointers (κατά την κατασκευή του ευρετηρίου)

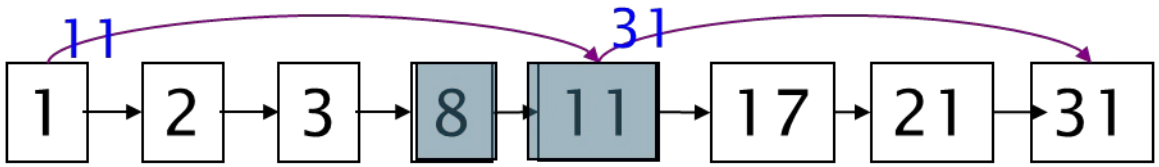




- Γιατί;
- Για να αποφύγουμε (skip) καταχωρίσεις που δεν θα εμφανιστούν στο αποτέλεσμα της αναζήτησης.
- Πώς;
- Πού να τοποθετήσουμε αυτούς τους δείκτες;

Επεξεργασία ερωτήματος με skip pointers



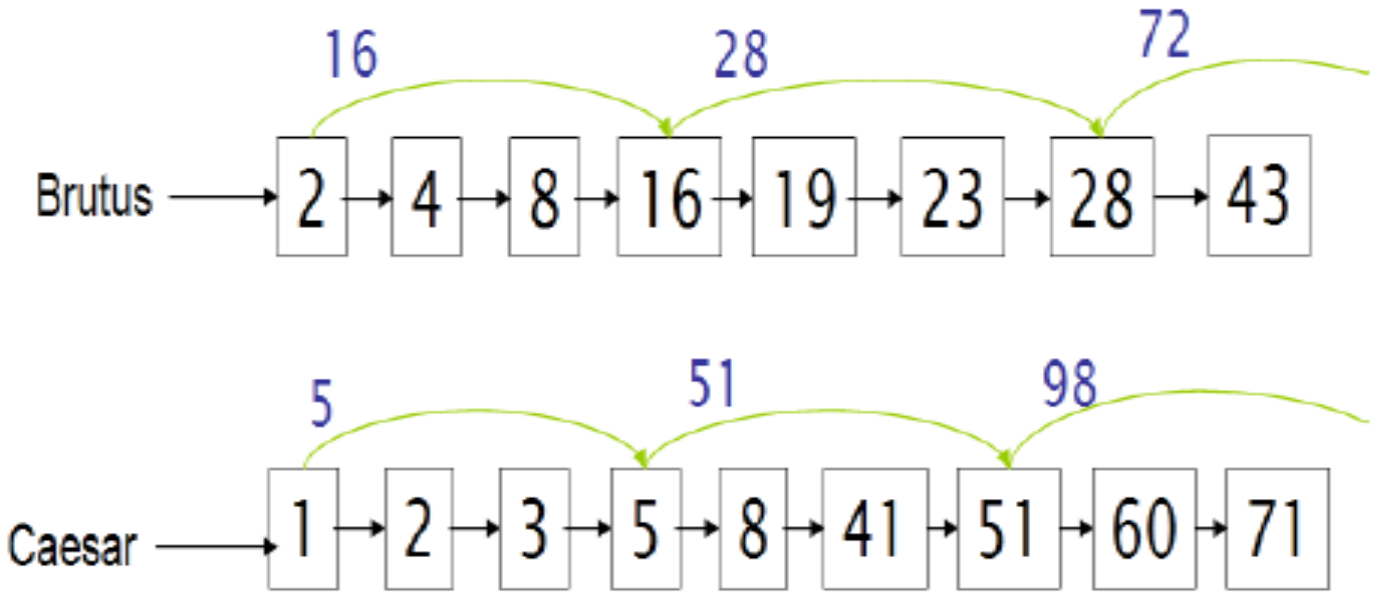


Υποθέστε ότι έχουμε διατρέξει τις λίστες και έχουμε βρει το κοινό στοιχείο **8** σε κάθε λίστα, το ταιριάζουμε και προχωράμε.

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

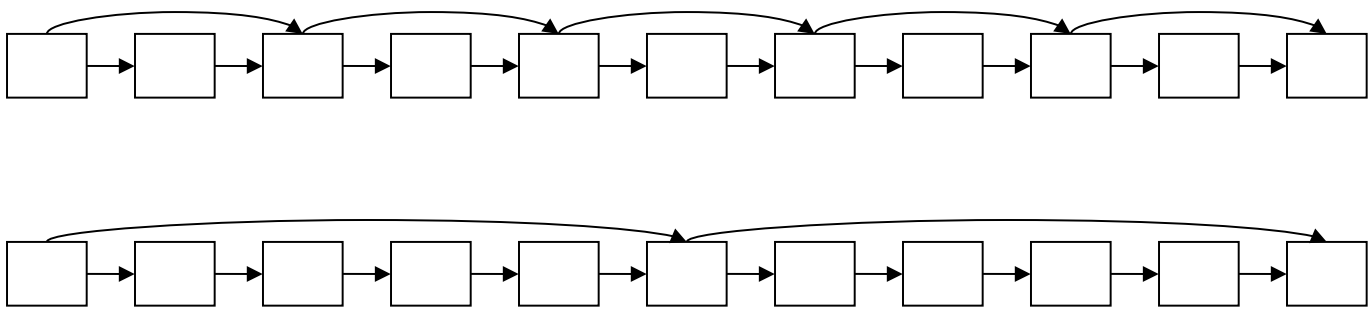
Αλλά ο δείκτης παράβλεψης του **11** είναι το **31**, οπότε μπορούμε να παραβλέψουμε τις ενδιάμεσες καταχωρίσεις.

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 παραβλέψουμε δείκτες.
  - Λιγότεροι δείκτες παράβλεψης $\rightarrow$ λιγότερες συγκρίσεις δεικτών αλλά μεγαλύτερα διαστήματα $\Rightarrow$ λίγες επιτυχημένες παραβλέψεις.



Τοποθέτηση των δεικτών

- Απλώς ευρετικός: για καταχωρίσεις μήκους for 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). Οχι τόσο με το νέο hardware.
  - Το I/O κόστος για να φορτωθεί μια μεγαλύτερη λίστα καταχωρίσεων 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 οι έμμεσες ερωτήσεις φράσεων
- 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.

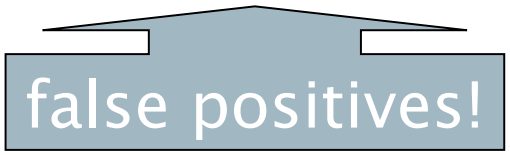
Μεγαλύτερες φράσεις

- 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.



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

Μη αποδοτικά για φράσεις μεγαλύτερες 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.

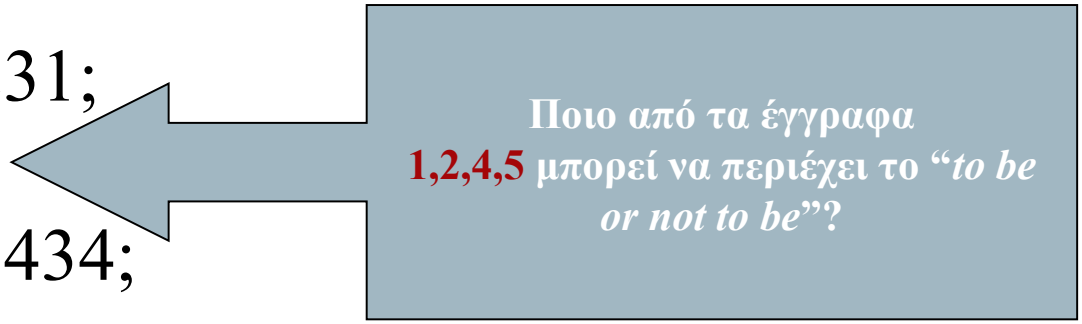
Λύση 2: Positional indexes (Ευρετήρια Θέσεων)

- Στις καταχωρίσεις, 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, …>



- 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 γειτνίασης (proximity searches).

Ερωτήματα γειτνίασης (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 συμπιέσουμε τα position values/offsets.
- Παρόλα αυτά, σημαντική αύξηση του χώρου αποθήκευση των λιστών καταχωρίσεων.
- 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

Γιατί?

- Εστω ένας όρος με συχνότητα 0.1%

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

Rules of thumb

- Ενα 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"***) η συνεχής συγχώνευση καταχωρίσεων ευρετηρίου θέσεων δεν είναι αποδοτική
  - Ακόμα περισσότερο για φράσεις όπως ***"The Who".***

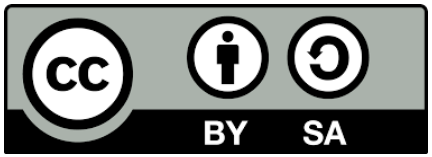
ΤΕΛΟΣ 2ου Μαθήματος

Ερωτήσεις;

Χρησιμοποιήθηκε κάποιο υλικό των:

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

Τέλος Ενότητας



Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Ιωαννίνων**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ενωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.



Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Εργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Εχουν προηγηθεί οι κάτωθι εκδόσεις:

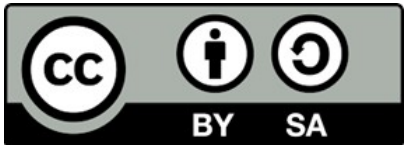
- Εκδοση 1.0 διαθέσιμη εδώ. http://ecourse.uoi.gr/course/view.php?id=1210 .

Σημείωμα Αναφοράς

Copyright Πανεπιστήμιο Ιωαννίνων, Διδάσκων: Καθ. Ευαγγελία Πιτουρά. «Ανάκτηση Πληροφορίας. Κατασκευή Ευρετηρίου και Είδη Ερωτήσεων». Εκδοση: 1.0. Ιωάννινα 2014. Διαθέσιμο από τη δικτυακή διεύθυνση: http://ecourse.uoi.gr/course/view.php?id=1210.

Σημείωμα Αδειοδότησης

- Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά Δημιουργού - Παρόμοια Διανομή, Διεθνής Εκδοση 4.0 [1] ή μεταγενέστερη.



- [1] https://creativecommons.org/licenses/by-sa/4.0/.