Προεπεξεργασία Κειμένου (Text Preprocessing)

Προεπεξεργασία

Προ-επεξεργασία Κειμένου

Κείμενο -> Όρους Ευρετηρίου

Λειτουργίες Κειμένου (Text Operations) κατασκευάζουν τις **λέξεις (όρους) ευρετηρίου** (tokens, index terms).

		Indexing Items					
		$\mathbf{k_1}$	$\mathbf{k_2}$	. . .	$\mathbf{k_j}$	. . .	$\mathbf{k_t}$
Documents	$\mathbf{d_1}$	$c_{1,1}$	$c_{2,1}$	. . .	$c_{i,1}$	. . .	$c_{t,1}$
	$\mathbf{d_2}$	$c_{1,2}$	$c_{2,2}$	. . .	$c_{i,2}$	. . .	$c_{t,2}$
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	$\mathbf{d_i}$	$c_{1,j}$	$c_{2,j}$	. . .	$c_{i,j}$	. . .	$c_{t,j}$
	. . .	. . .	. . .	. . .	. . .	. . .	. . .
	$\mathbf{d_N}$	$c_{1,N}$	$c_{2,N}$	. . .	$c_{i,N}$	. . .	$c_{t,N}$

Προεπεξεργασία

1. Collect the documents
2. Tokenized the text
3. Do linguistic processing of tokens
4. Index the documents that each term occurs in

Προεπεξεργασία Κειμένου

- Σκεπτικό
  - δεν είναι όλες οι λέξεις ενός κειμένου κατάλληλες για την παράσταση του περιεχομένου του (μερικές λέξεις φέρουν περισσότερο νόημα από άλλες)
- Στόχοι της προεπεξεργασίας κειμένου
  - βελτίωση της αποτελεσματικότητας (effectiveness)
  - βελτίωση της αποδοτικότητας (efficiency) της ανάκτησης
  - προσπάθεια ελέγχου (κυρίως μείωσης) του λεξιλογίου
    - και εκ τούτου μείωσης του μεγέθους των ευρετηρίων

Κύριες Φάσεις Προεπεξεργασίας

1. Λεξιλογική ανάλυση (lexical analysis)
   - αναγνώριση *αριθμών*, *λέξεων*, *διαχωριστικών*, *σημείων στίξεως*, κλπ **[tokens]**
2. Αποκλεισμός λέξεων (stopwords)
   - απαλοιφή λέξεων με πολύ μικρή διακριτική ικανότητα (*άρθρα*, *αντωνυμίες*, *κτητικές αντωνυμίες*, κλπ)
3. Στελέχωση (stemming) των εναπομεινάντων λέξεων
   - απαλοιφή *καταλήξεων/προθεμάτων* (αυτοκίνητο, αυτοκίνητα, αυτοκινήτων) για την ανάκτηση των κειμένων που περιέχουν μορφολογικές παραλλαγές των λέξεων της επερώτησης
4. Επιλογή των λέξεων που θα χρησιμοποιηθούν στον ευρετηριασμό
   - συχνά γίνεται βάσει του μέρους του λόγου (ουσιαστικά, επίθετα, επιρρήματα, ρήματα)
5. Κατασκευή δομών κατηγοριοποίησης

Επεξεργασία Κειμένου

Άλλες σχετικές λειτουργίες κειμένου (αργότερα)

- Συμπίεση (compression)
- Κωδικοποίηση (encryption)
- Συσταδοποίηση (cluster)

Προεπεξεργασία Κειμένου

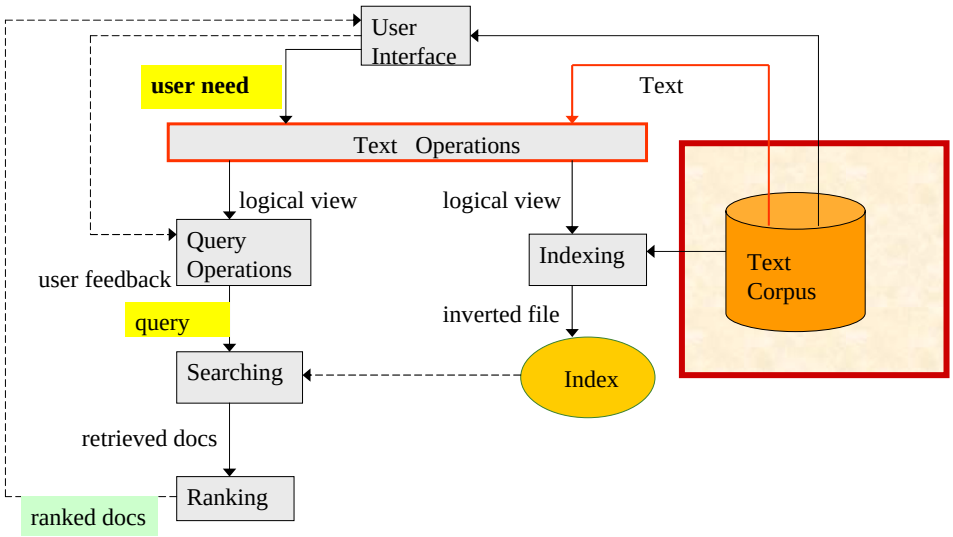
Δε βελτιώνεται πάντα η αποτελεσματικότητα:

Για παράδειγμα:

Ένας χρήστης που αναζητά ένα έγγραφο που περιέχει την έκφραση “house of the lord”

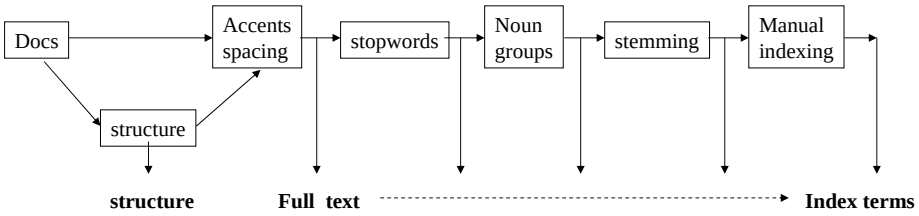
Τι γίνεται αν δεν έχει αποθηκευτεί το “of” και το “the”

Τμήματα της Αρχιτεκτονικής που Εμπλέκονται



Φάσεις Προεπεξεργασίας

Από το πλήρες κείμενο στους όρους ευρετηρίου



Tokenization: chopping character streams into tokens

Linguistic processing: building equivalence classes of tokens

Parsing a document

Obtaining the character sequence in a document

encoded -> (encoding) character sequence

Binary form (eg DOC files) and or compressed format

- What format is it in?
  - pdf/word/excel/html?
- What language is it in?
- What character set is in use? (σε κάποια δυαδική μορφή)

But these tasks are often done heuristically …

Complications: Format/language

- Documents being indexed can include docs from many different languages
  - A single index may have to contain terms of several languages.
- Sometimes a document or its components can contain multiple languages/formats
  - French email with a German pdf attachment.

Language identification: use short character subsequences as features

Complications: Index Granularity

- What is a unit document? (οι «γραμμές» στο ευρετήριο)
  - A file?
  - An email? (Perhaps one of many in an mbox.)
  - An email with 5 attachments?
  - A group of files (PPT or LaTeX in HTML)
  - A book or a book chapter?

Precision/recall trade-off

Implicit or Explicit Proximity Search

Query Pre-processing

- For either Boolean or free text queries, the same pre-processing as for the documents.

Λεξιλογική Ανάλυση (Lexical Analysis)

Λεξιλογική Ανάλυση (Lexical Analysis - Tokenization)

Σκοπός: Μετατροπή του κειμένου του εγγράφου (μιας *ροής χαρακτήρων*) σε μια *ροή λέξεων* που θα ευρετηριοποιηθούν

Παράδειγμα

Input: "***Friends, Romans and Countrymen***"

Γενικά, κόβουμε όπου κενοί χαρακτήρες;

Output: Tokens

- ***Friends***
- ***and***
- ***Romans***
- ***Countrymen***

Token: an instance of a sequence of characters in some particular document grouped together as a useful semantic unit for processing

Type: class of all tokens containing the same character sequence

Term: (perhaps normalized) type that is included in the dictionary

Each such token is now a candidate for an index entry, after further processing

But what are valid tokens to emit? (usually, noun words)

Λεξιλογική Ανάλυση (Lexical Analysis)

What are the correct tokens?

Περιπτώσεις που απαιτούν προσοχή:

- Αριθμοί
- Παύλες (hyphens)
- Σημεία στίξεως (punctuations)
- Μικρά-κεφαλαία

Εξαρτάται και από τη γλώσσα

Λεξιλογική Ανάλυση (Lexical Analysis)

Αριθμοί

- ***3/12/91*** ***Mar. 12, 1991***
- ***55 B.C.***
- ***B-52***
- ***My PGP key is 324a3df234cb23e***
- ***100.2.86.144***

Οι αριθμοί από μόνοι τους είναι πολύ ασαφής χωρίς τα συμφραζόμενα

Συχνά δεν δεικτοδοτούνται όπως το κείμενο

- Λέξεις που περιέχουν ψηφία

  O2, βιταμίνη B6, B12, Windows98, 510B.C.

- Χρήσιμοι πχ

  αριθμοί πιστωτικής κάρτας, error codes/stacktraces on the web, κλπ

- Κανονικοποίηση σε κάποια κοινή μονάδα
- Ξεχωριστή δεικτοδότηση μεταδεδομένων (πχ ημερομηνία δημιουργίας, είδος αρχείου χωριστά)

Λεξιλογική Ανάλυση (Lexical Analysis)

Παύλες (hyphens)

Used for:

- Splitting-up vowels in words (co-education) *(one token)*
- Joining nouns as names (Hewlett-Packard)
- Copy editing device to show word grouping the hold-him-back-and-drag-him-away-maneuver ? *(split up)*

"state of the art" vs "state-of-the-art", "Jean-Luc Hainaut", "Jean-Roch Meurisse", F-16, MS-DOS

- Due to inconsistency of use, breakup hyphenated words
- But, there are words which include hyphens as an integral part
- Adopt a general rule, specify exceptions on a case by case basis *(eg allow short hyphenated prefixes on words, but no longer hyphenated forms)*

Λεξιλογική Ανάλυση (Lexical Analysis)

Similar with nonseparating whitespace

- San Francisco vs San-Francisco, White space vs whitespace
- York University, New York University
- Phone numbers, dates (Nov 10, 2009), etc

Encourage users to use hyphens and cover all three cases in queries: over-eager, overeager, over eager

Λεξιλογική Ανάλυση (Lexical Analysis)

Σημεία Στίξης (punctuation)

- Συνήθως παραλείπονται – απομάκρυνση και από την ερώτηση (πχ 380 π.χ.)
- Ειδικές περιπτώσεις: OS/2, .NET, command.com, μεταβλητές x.id και xid

- **Apostrophe** for possession and contractions:
  - ***Finland's capital → Finland? Finlands? Finland's***?
  - aren't -> arent, are n't, aren t vs O'Neil -> oneill, o'neill, o' neill, o neill

Λεξιλογική Ανάλυση (Lexical Analysis)

Μικρά-κεφαλαία

- Συνήθως όλα μετατρέπονται σε μικρά
- Bank και bank, General Motors, bush και Bush
- Unix-like convention

Convert lowercase at the beginning of a sentence and all words occurring in a title that is all uppercase or in which all or most words are capitalized

Truecasing (machin learning sequence model)

Most practical solution: lowercase everything

Tokenization: Language issues

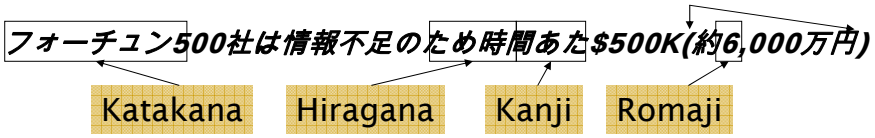
- ***L'ensemble*** → one token or two?
  - ***L*** ? ***L'*** ? ***Le*** ?
  - Want ***l'ensemble*** to match with ***un ensemble***
- German noun compounds are not segmented
  - Lebensversicherungsgesellschaftsangestellter
  - 'life insurance company employee'

  Use of a compound splitter

Tokenization: language issues

k-grams: all indexing via just short subsequences of characters, regardless if whether particular sequences cross word boundaries)

- Chinese and Japanese have *no spaces* between words:
  - 莎拉波娃现在居住在美国东南部的佛罗里达。
  - Word segmentation, Not always guaranteed a unique tokenization
  - Or, abandon word-based indexing and use *k-grams*
- Further complicated in Japanese, with multiple alphabets intermingled
  - Dates/amounts in multiple formats



End-user can express query entirely in hiragana!

Tokenization: language issues

- Arabic (or Hebrew) is basically written right to left, but with certain items like numbers written left to right (no linear order)

  Words are separated, but letter forms within a word form complex ligatures
- استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.
- ← → ← → ← start
- 'Algeria achieved its independence in 1962 after 132 years of French occupation.'
- With Unicode, the surface presentation is complex, but the stored form is straightforward (sequence of sounds – a linear structure)

Λεξιλογική Ανάλυση

- Λεξιλογική Ανάλυση για Επερωτήσεις
  - Όπως και για το κείμενο, συν αναγνώριση χαρακτήρων ελέγχου, όπως
    - λογικοί τελεστές, π.χ. AND, OR, NOT,
    - τελεστές εγγύτητας (proximity operators),
    - κανονικές εκφράσεις (regular expressions), κτλ.
- Τρόποι υλοποίησης ενός Λεξιλογικού Αναλυτή
  - (α) χρήση μιας γεννήτριας λεξιλογικών αναλυτών (lexical analyzer generator), όπως τον lex
    - η καλύτερη επιλογή αν υπάρχουν σύνθετες περιπτώσεις
  - (β) συγγραφή (προγραμματισμός) ενός λεξιλογικού αναλυτή με το χέρι
    - η χειρότερη επιλογή (επιρρεπής σε σφάλματα)
  - (γ) συγγραφή (προγραμματισμός) ενός λεξιλογικού αναλυτή ως μια μηχανή πεπερασμένων καταστάσεων (finite state machine)

Λέξεις Αποκλεισμού (Stopwords)

Λέξεις Αποκλεισμού (Stopwords)

Απαλοιφή λέξεων με πολύ μικρή διακριτική ικανότητα (little semantic context, take a lot of space) – π.χ. λέξεις που εμφανίζονται στο 80% των εγγράφων

συνήθως: *άρθρα*, *αντωνυμίες*, *κτητικές αντωνυμίες*, κλπ

- e.g. “a”, “the”, “in”, “to”; pronouns: “I”, “he”, “she”, “it”.

Επίσης κάποια ρήματα, επίθετα, επιρρήματα

- Οφέλη
  - μείωση μεγέθους ευρετηρίου **(έως και 40%)**

Λέξεις Αποκλεισμού (Stopwords)

Παρατηρήσεις

- Οι λέξεις αποκλεισμού εξαρτώνται από τη γλώσσα και τη συλλογή
- Not every frequent english word should be in the list
  - Top 200 English words include «time, war, home, life, water, world»
  - In a CS corpus we could add to the stoplist the words: «computer, program, source, machine, language»

Example: Stopwords for the English language

a be had it only she was about because has its of some we after been have last on such were all but he more one than when also by her most or that which an can his mr other the who any co if mrs out their will and corp in ms over there with are could inc mz s they would as for into no so this up at from is not says to

Example: Stopwords for the French language

a afin ah ai aie aient aies ailleurs ainsi ait alentour alias allais allaient allait allons allez alors Ap. Apr. aprs aprs demain arrire as assez attendu au aucun aucune au dedans au dehors au dela au dessous au dessus au devant audit aujourd aujourdhui auparavant auprs auquel aura aurai auraient aurais aurait auras aurez auriez aurions aurons auront aussi aussitôt autant autour autre autrefois autres autrui aux auxdites auxdits auxquelles auxquels avaient avais avait avant avant hier avec avez aviez avions avoir avons ayant ayez ayons B bah banco be beaucoup ben bien bientôt bis bon C c Ca ça ça cahin caha car ce ce ceans ceci cela celle celle ci celle la celles celles ci celles la celui celui ci celui la cent cents cependant certain certaine certaines certains certes ces cest a dire cet cette ceux ceux ci ceux la cf. cg cgr chacun chacune chaque cher chez ci ci ci aprs ci dessous ci dessus cinq cinquante cinquante cinq cinquante deux cinquante et un cinquante huit cinquante neuf cinquante quatre cinquante sept cinquante six cinquante trois cl cm cm combien comme comment contrario contre crescendo D d dabord daccord daffilee dailleurs dans daprs darrache pied davantage de debout dedans dehors deja dela demain demblee depuis derechef derrire des ds desdites desdits desormais desquelles desquels dessous dessus deux devant devers dg die differentes differents dire dis disent dit dito divers diverses dix dix huit dix neuf dix sept dl dm donc dont dorenavant douze du dû dudit duquel durant E eh elle elle elles elles en en en encore enfin ensemble ensuite entre entre temps envers environ es s est et et/ou etaient etais etait etant etc ete êtes etiez etions être eu eue eues euh eûmes eurent eus eusse eussent eusses eussiez eussions eut eût eûtes eux exprs extenso extremis F facto fallait faire fais faisais faisait faisaient faisons fait faites faudrait faut fi flac fors fort forte fortiori frais fûmes fur furent fus fusse fussent fusses fussiez fussions fut fût fûtes G GHz gr grosso gure H ha han haut he hein hem heu hg hier hl hm hm hola hop hormis hors hui huit hum I ibidem ici ici bas idem il il illico ils ils ipso item J j jadis jamais je je jusqu jusqua jusquau jusquaux jusque juste K kg km km² L l la la la la la bas la dedans la dehors la derrire la dessous la dessus la devant la haut laquelle lautre le le lequel les les ls lesquelles lesquels leur leur leurs lez loin lon longtemps lors lorsqu lorsque lui lui lun lune M m m m ma maint mainte maintenant maintes maints mais mal malgre me même mêmes mes mg mgr MHz mieux mil mille milliards millions minima ml mm mm² modo moi moi moins mon moult moyennant mt N n nagure ne neanmoins neuf ni nº non nonante nonobstant nos notre nous nous nul nulle O ô octante oh on on ont onze or ou où ouais oui outre P par parbleu parce par ci par dela par derrire par dessous par dessus par devant parfois par la parmi partout pas passe passim pendant personne petto peu peut peuvent peux peut être pis plus plusieurs plutôt point posteriori pour pourquoi pourtant prealable prs presqu presque primo priori prou pu puis puisqu puisque Q qu qua quand quarante quarante cinq quarante deux quarante et un quarante huit quarante neuf quarante quatre quarante sept quarante six quarante trois quasi quatorze quatre quatre vingt quatre vingt cinq quatre vingt deux quatre vingt dix quatre vingt dix huit quatre vingt dix neuf quatre vingt dix sept quatre vingt douze quatre vingt huit quatre vingt neuf quatre vingt onze quatre vingt quatorze quatre vingt quatre quatre vingt quinze quatre vingts quatre vingt seize quatre vingt sept quatre vingt six quatre vingt treize quatre vingt trois quatre vingt un quatre vingt une que quel quelle quelles quelqu quelque quelquefois quelques quelques unes quelques uns quelquun quelquune quels qui quiconque quinze quoi quoiqu quoique R revoici revoila rien S s sa sans sauf se secundo seize selon sensu sept septante sera serai seraient serais serait seras serez seriez serions serons seront ses si sic sine sinon sitôt situ six soi soient sois soit soixante soixante cinq soixante deux soixante dix soixante dix huit soixante dix neuf soixante dix sept soixante douze soixante et onze soixante et un soixante et une soixante huit soixante neuf soixante quatorze soixante quatre soixante quinze soixante seize soixante sept soixante six soixante treize soixante trois sommes son sont soudain sous souvent soyez soyons stricto suis sur sur le champ surtout sus T t t ta tacatac tant tantôt tard te tel telle telles tels ter tes toi toi ton tôt toujours tous tout toute toutefois toutes treize trente trente cinq trente deux trente et un trente huit trente neuf trente quatre trente sept trente six trente trois trs trois trop tu tu U un une unes uns USD V va vais vas vers veut veux via vice versa vingt vingt cinq vingt deux vingt huit vingt neuf vingt quatre vingt sept vingt six vingt trois vis a vis vite vitro vivo voici voila voire volontiers vos votre vous vous W X y y Z zero

Απαλοιφή λέξεων Αποκλεισμού: Τρόποι

Τρόποι Υλοποίησης

1/ Απαλοιφή των λέξεων αποκλεισμού μετά το τέλος της λεξιλογικής ανάλυσης

- Μπορούμε να αποθηκεύσουμε τις λέξεις αυτές σε έναν hashtable για να τις αναγνωρίζουμε γρήγορα (σε σταθερό χρόνο)

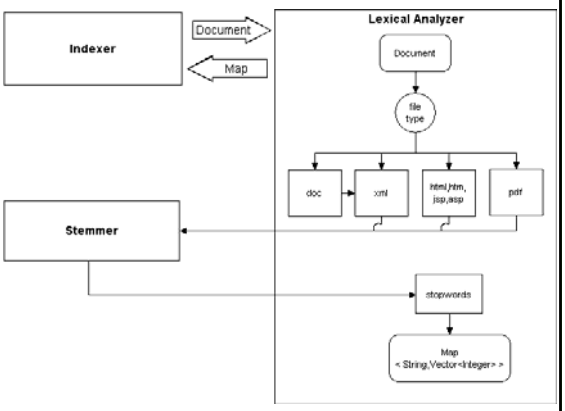
2/ Απαλοιφή των λέξεων αποκλεισμού κατά τη διάρκεια της λεξιλογικής ανάλυσης

- Πιο γρήγορη προσέγγιση αφού η λεξιλογική ανάλυση θα γίνει έτσι και αλλιώς και η αφαίρεση των λέξεων αποκλεισμού δεν απαιτεί επιπλέον χρόνο

Παρέχεται μια stop-list

Περίπτωση: Lexical Analyzer of grOOGLE

- Recognition of the document's structure.
- The lexical analyzer accepts the following file types: html (html, htm, php, jsp, asp), doc, ppt, pps, xls, rtf, txt. For the text extraction from the documents various software components are used, specifically pdfbox2 for pdf documents, Jakarta POI for doc, ppt, pps, xls, and RTFEditorKit for RDF documents.
- For more see:
  - http://groogle.csd.uoc.gr/apache2-default/index.php/Lexical_Analyzer



Stop words

You need them for:

- Phrase queries: “King of Denmark”
- Various song titles, etc.: “Let it be”, “To be or not to be”
- “Relational” queries: “flights to London”

But the trend is away from doing this – it may reduce recall

(200-300 terms, 7-12 terms, none)

- Good compression techniques means the space for including stopwords in a system is very small (reduce the cost of storing the postings of common words)
- Term-weights leads to very common words having little impact on document ranking
- Good query optimization techniques mean you pay little at query time for including stop words (impact-sorted indexes terminate scanning posting lists)

Normalization

The process of canonicalizing tokens so that matches occur despite superficial differences

- Need to “normalize” terms in indexed text as well as query terms into the same form
  - We want to match ***U.S.A.*** and ***USA, colour*** and ***color***

Most commonly implicitly define equivalence classes of terms by using **mapping rules**

e.g., by deleting periods in a term, case-folding, etc

But it is not obvious, when to add characters (turn antidiscrimating into anti-discriminating)

Normalization

Maintain relations between unnormalized tokens

Can be extended to hand-constructed lists of synomyms (e.g., car and automobile)

Method 1: Index unnormalized tokens and maintain a query expansion list of multiple vocabulary entries for each query term

A query is a disjunction of several posting lists

Method 2: Perform the expansion during index construction (e.g., we index a document containing car under automobile as well)

- Method 1: a query expansion dictionary + requires more time at query processing
- Method 2: more space

Normalization

Both methods less efficient than equivalence classes

But, more flexible, expansion lists can overlap, while not being identical, asymmetric expansion:

Enter: ***window***	Search: ***window, windows***
Enter: ***windows***	Search: ***Windows, windows***
Enter: ***Windows***	Search: ***Windows***

Open question

Can create problems: U.S.A. -> USA vs C.A.T. -> CAT

Normalization: other languages

- Accents: ***résumé*** vs. ***resume***.
- Most important criterion:
  - How are your users like to write their queries for these words?
    - Even in languages that standardly have accents, users often may not type them
- German: Tuebingen vs. Tübingen
  - Should be equivalent

Normalization: other languages

- Need to "normalize" indexed text as well as query terms into the same form

  7月30日 vs. 7/30

- Character-level alphabet detection and conversion
  - Tokenization not separable from this.
  - Sometimes ambiguous:

Morgen will ich in MIT … Is this German "mit"?

Normalization: other languages

- 60% of web pages are in English
- Less than 10% speak English
- 1/3 of blogspots (2007) in English

Στελέχωση Κειμένου (Stemming)

Στελέχωση Κειμένου (Stemming)

- Υποβίβαση λέξεων στη ρίζα τους για ανεξαρτησία από τις μορφολογικές παραλλαγές των λέξεων
  - «αυτοκίνητο», «αυτοκίνητα», «αυτοκινήτων»
  - “computer”, “computational”, “computation” all reduced to same token “compute”

Στόχοι

- Βελτίωση αποτελεσματικότητας (κυρίως της ανάκλησης)
- Μείωση του μεγέθους του ευρετηρίου
  - Συγκεκριμένα του λεξιλογίου του ευρετηρίου

Again, there is controversy in the literature for the benefit of stemming for retrieval performance

Στελέχωση Κειμένου (Stemming)

Stemming vs Lemmatization

Stemming: crude heuristics that chops off the end of words (often removal of derivational affixes)

Lemmatization: use vocabulary and morphological analysis to remove inflectional endings and return he base or dictionary form of a word (**lemma**)

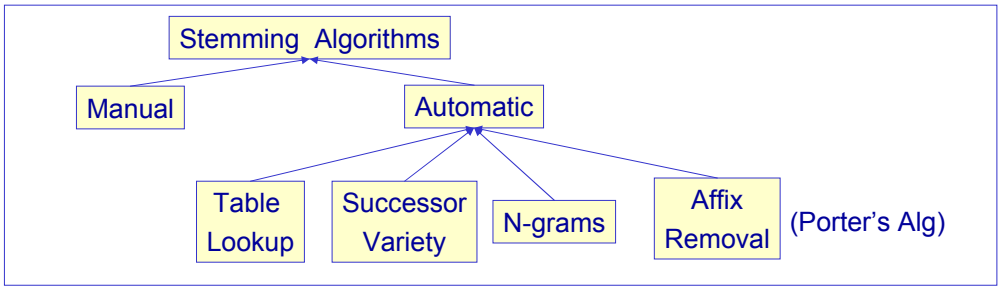
Example

saw: additional plug-in components

stemming may return just s

lemmatization return either see or saw

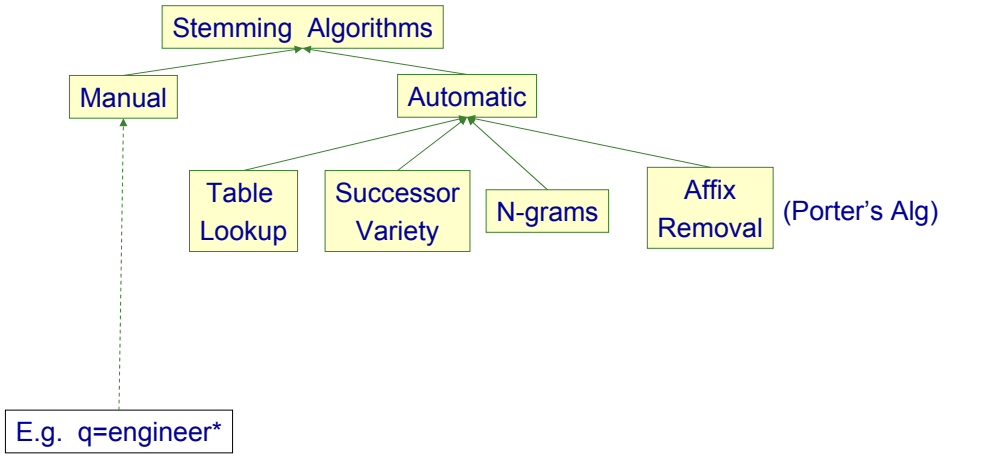
[γ] Αλγόριθμοι Στελέχωσης (Stemming Algorithms)



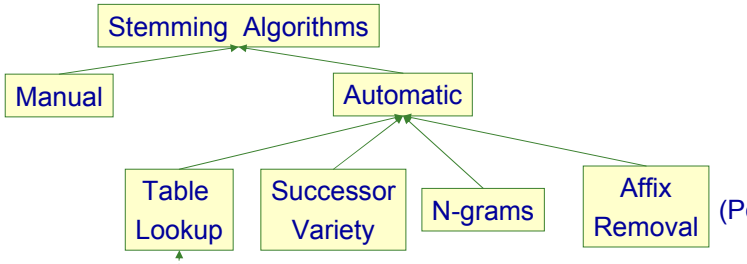
Πως αξιολογούμε έναν αλγόριθμο στελέχωσης;

- Ορθότητα (Correctness)
  - υπερστελέχωση (overstemming) έναντι υποστελέχωσης (understemming)
- Αποτελεσματικότητα ανάκτησης (Retrieval effectiveness)
- Δυνατότητα Συμπίεσης (Compression performance)

Αλγόριθμοι Στελέχωσης: Χειρονακτικός



Αλγόριθμοι Στελέχωσης: Με Πίνακα



Terms and their corresponding stems are stored in a table (stemming dictionary) ,e.g.:

Term	Stem
engineering	engineer
engineered	engineer
engineer	engineer

(such tables are not easily available)

Απλώς ψάχνουμε στον πίνακα – δύσκολο να κατασκευαστούν

Αλγόριθμοι Στελέχωσης: Successor Variety

Ιδέα: Στελέχωση βάσει των συχνοτήτων των ακολουθιών γραμμάτων σε ένα σώμα κειμένου

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)
[2] Χρήση του πίνακα για τεμαχισμό των λέξεων
[3] Επιλογή ενός τεμαχίου ως ρίζα (as stem)

Αλγόριθμοι Στελέχωσης: Successor Variety

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)

Παράδειγμα

Έστω ότι θέλουμε να βρούμε τη ρίζα της λέξης READABLE

Έστω το εξής σώμα κειμένου: ABLE, APE, BEATABLE, FIXABLE, READ, READABLE, READING, READS, RED, ROPE, RIPE

– Πρόθεμα Αριθμός Επόμενων Γραμμάτων Επόμενα Γράμματα

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	*BLANK*

Αλγόριθμοι Στελέχωσης: Successor Variety

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)

[2] Χρήση του πίνακα για τεμαχισμό των λέξεων

Πρόθεμα Αριθμός Επόμενων Γραμμάτων Επόμενα Γράμματα

Prefix	Successor Variety	Letters
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	*BLANK*

Αλγόριθμοι Στελέχωσης: Successor Variety

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)

[2] Χρήση του πίνακα για τεμαχισμό των λέξεων

Πρόθεμα	Αριθμός Επόμενων Γραμμάτων	Επόμενα Γράμματα
Prefix	**Successor Variety**	**Letters**
R	3	E,I,O
RE	2	A,D
REA	1	D
READ	3	A,I,S
READA	1	B
READAB	1	L
READABL	1	E
READABLE	1	*BLANK*

[2] Τεμαχισμός βάσει της μεθόδου «peak & plateau»:

- τεμαχισμός στο γράμμα που οι διάδοχοί του είναι **περισσότεροι** των διαδόχων του προηγούμενου γράμματος
  - REA (1), READ (3)

Άρα READABLE => READ ABLE

Αλγόριθμοι Στελέχωσης: Successor Variety

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)

[2] Χρήση του πίνακα για τεμαχισμό των λέξεων

[3] Επιλογή ενός τεμαχίου ως ρίζα (as stem)

READABLE => READ ABLE

Ένας ευρετικός κανόνας:

"if (first segment occurs in <=12 words in the corpus) select first segment, else the second"

Δικαιολόγηση: Αν εμφανίζεται πάνω από 12 φορές τότε μάλλον είναι πρόθεμα.

Αλγόριθμοι Στελέχωσης: Successor Variety

Βήματα για Στελέχωση Κειμένου

[1] Δημιουργία του πίνακα Ποικιλίας Διαδόχων (successor variety table)

[2] Χρήση του πίνακα για τεμαχισμό των λέξεων

π.χ. READABLE => READ ABLE

[3] Επιλογή ενός τεμαχίου ως ρίζα (as stem)

π.χ. READABLE => READ ABLE

Παρατήρηση:

Η τεχνική αυτή δεν απαιτεί καμία είσοδο από το σχεδιαστή. Άρα μπορεί να εφαρμοστεί αυτούσια σε πολλές διαφορετικές γλώσσες.

Αλγόριθμοι Στελέχωσης: n-grams

Ιδέα: Ομαδοποίησε λέξεις βάσει του αριθμού των κοινών διγραμμάτων ή n-γραμμάτων

Πχ: σύγκριση "**statistics**" με "**statistical**"

- "statistics":
  - digrams: st ta at ti is st ti ic cs (9)
  - unique digrams: at cs ic is st ta ti (7)
- "statistical":
  - digrams: st ta at ti is st ti ic ca al (10)
  - unique digrams: al at ca ic is st ta ti (8)

Οι λέξεις "statistics" και "statistical" έχουν 6 κοινά διγράμματα (digrams).

Αλγόριθμοι Στελέχωσης: n-grams

Οι λέξεις “statistics” και “statistical” έχουν 6 κοινά διγράμματα (digrams).

Μπορούμε να μετρήσουμε τον βαθμό ομοιότητάς τους χρησιμοποιώντας μια μετρική, όπως:

- Μέγεθος τομής: $sim(X, Y) = |X \cap Y|$
- Dice similarity: $sim(X, Y) = 2\,|X \cap Y|/(|X|+|Y|)$
  - εδώ sim(statistics,statistical) = 2*6/(7+8)=0.8

Οι λέξεις της συλλογής ομαδοποιούνται με αυτόν τον τρόπο
(όλες οι λέξεις που έχουν την ίδια ρίζα καταχωρούνται στην ίδια ομάδα)

Αλγόριθμους ομαδοποίησης θα δούμε σε επόμενο μάθημα.
Γενικά, θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε μετρική μεταξύ λέξεων (π.χ. Edit, LCS distance)

Porter’s algorithm

- Commonest algorithm for stemming English
  - Results suggest at least as good as other stemming options
- Conventions + 5 phases of reductions
  - phases applied sequentially
  - each phase consists of a set of commands
  - sample convention: *Of the rules in a compound command, select the one that applies to the longest suffix.*

The official site for the Porter Stemmer is:

http://www.tartarus.org/~martin/PorterStemmer/

Αλγόριθμοι Στελέχωσης: Affix Removal

Ιδέα: Απαλοιφή επιθεμάτων (suffixes) ή/και προθεμάτων (prefixes)

Porter's Stemmer

- Simple procedure for removing known affixes in English without using a dictionary (i.e. without a lookup table).
- Can produce unusual stems that are not English words:
  - "computer", "computational", "computation" all reduced to same token "comput"
- May conflate (reduce to the same token) words that are actually distinct.
- Not recognize all morphological derivations.

Αλγόριθμοι Στελέχωσης: Porter Stemmer

Παραδείγματα κανόνων:

- s → ∅ (for plural form)
- sses → ss (for plural form)

Εφαρμόζεται πρώτα η μακρύτερη ακολουθία

- e.g. stresses => stress,
- NOT stresses => stresse

measure of a word

A word is long enough to regard the matching portion of a rule as a suffix

Replacement -> replac

Cement -> ?

RULES

suffix	replacement	example
1a		
sses	ss	caresses->caress
ies	i	ponies->poni, ties->ti
s	NUL	cats->cat
1b		
eed	ee	agreed->agree
ed	NUL	plastered->plaster
ing	NUL	motoring->motor
2		
ational	ate	relational->relate
tional	tion	conditional->condition
izerize		digitizer->digitize
ator	ate	operator->operate
….		

Αλγόριθμοι Στελέχωσης: Porter Stemmer> Example

Original text:

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

After applying Porter's Stemmer (and eliminating stopwords):

market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

Αλγόριθμοι Στελέχωσης: Porter Stemmer >Errors

- Errors of "comission":
  - organization, organ → organ
  - police, policy → polic
  - arm, army → arm
- Errors of "omission":
  - cylinder, cylindrical
  - create, creation
  - Europe, European

Αλγόριθμοι Στελέχωσης: Porter Stemmer > Code

- See [book MIR, Appendix]
- Demo available at:
  - http://snowball.tartarus.org/demo.php
- Implementation (C, Java, …) available at:
  - http://www.tartarus.org/~martin/PorterStemmer/

Αλγόριθμοι Στελέχωσης: Για την ελληνική γλώσσα

Δυσκολίες της Ελληνικής Γλώσσας:

- Υπάρχουν πολλές διαφορετικές καταλήξεις που προκύπτουν από τον αριθμό, αλλά και από τις πτώσεις των ουσιαστικών, ανωμάλων και μη. Στα επίθετα η επιπλέον ύπαρξη του γένους, συντελεί στην περαιτέρω αύξηση των καταλήξεων. Τα ρήματα με την σειρά τους διαθέτουν δύο διαφορετικά θέματα (ενεστώτα και αορίστου), ενώ υπάρχουν αρκετές

  περιπτώσεις ανωμάλων ρημάτων.

  Παραδείγματα
  - πράττω, πράξη, πρακτικός
  - αναδιάταξη, αναδιατάσσω
  - ..

Αλγόριθμοι Στελέχωσης: Για την ελληνική γλώσσα

grOOGLE

Το grOOGLE προσφέρει ένα στελεχωτή της ελληνικής.

Η διαδικασία στελέχωσης (.. σε γενικές γραμμές):

- Η λέξη δέχεται μια αρχική επεξεργασία: μετατροπή σε **"μικρούς"** (μη κεφαλαίους) χαρακτήρες, κάθε χαρακτήρας ελέγχεται αν περιέχεται στο σύνολο των τονισμένων χαρακτήρων και αντικαθίσταται από τον αντίστοιχο **μη τονισμένο**.
- Αφαιρούνται πιθανοί επαναλαμβανόμενοι χαρακτήρες από την αρχή ή το τέλος μιας λέξης (χαρακτηριστικό που δεν εντοπίζεται στην ελληνική γλώσσα).
- Εντοπισμός και αφαίρεση **πιθανών προθεμάτων** στη λέξη.
- Παρόμοια επεξεργασία εφαρμόζεται στην χωρίς προθέματα λέξη.
- Η κατάληξη που αφαιρείται μπορεί να οδηγήσει σε θέμα ενός χαρακτήρα, οπότε απαιτείται και εφαρμόζεται μια μέθοδος αύξησης του θέματος έτσι ώστε να περιλαμβάνει τουλάχιστον μια συλλαβή.
- Η στελέχωση ολοκληρώνεται με την προσθήκη των προθεμάτων που πιθανόν να έχουν αφαιρεθεί.

Αλγόριθμοι Στελέχωσης: Για την ελληνική γλώσσα

grOOGLE

- Το groogle προσφέρει έναν στελεχωτή της ελληνικής
  - Δείτε το **http://groogle.csd.uoc.gr/apache2-default/index.php/Stemmer**

Word	Word Split	prefixes-First Stem	Increment-Alternate	Final Stem
πραττω	πραττω	πρατ	πρας	πρας
πρακτικος	πρακτικς	πρακτ	πρας	πρας
πραξη	πραξη	πρας	πρας	πρας
πραγμα	πραγμα	πραγμ	πρας	πρας
αναδιαταξη	ανα − δια − ταξη	ανα − δια − ταξ	ανα − δια − ταξ	αναδιαταξ
αναδιατασσω	ανα − δια − τασσω	ανα − δια − τασσ	ανα − δια − ταξ	αναδιαταξ
αναδιεταξα	ανα − διε − ταξα	ανα − δια − ταξ	ανα − δια − ταξ	αναδιαταξ
παω	παω	π	πηγ	πηγ
πηγαινω	πηγαινω	πηγ	πηγ	πηγ

Αλγόριθμοι Στελέχωσης

Algorithm	Language Independent
Lookup table	NO
Successor Variety	YES
N-Grams	YES
Porter's Stemmer, grOOGLE stemmer	NO

Other stemmers

- Other stemmers exist, e.g., Lovins stemmer
  http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm
  - Single-pass, longest suffix removal (about 250 rules)
  - Motivated by linguistics as well as IR

Paice/Husk stemmer (1990):

http://www.cs.waikato.ac.nz/~eibe/stemmers/
http://www.comp.lancs.ac.uk/computing/research/stemming/

Other stemmers

- Full morphological analysis – at most modest benefits for retrieval
- Do stemming and other normalizations help?
  - Often very mixed results: really help recall for some queries but harm precision on others
    - operational and research
    - operating and system
    - operative and dentistry

Works for other languages with much more morphology

Language-specificity

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

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

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

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

Επιλογή Λέξεων για την Ευρετηρίαση

Επιλογή Λέξεων για την Ευρετηρίαση

- Μια προσέγγιση είναι να θεωρήσουμε ως όρους ευρετηρίου ό,τι απέμεινε (αφαιρώντας λέξεις αποκλεισμού, κάνοντας στελέχωση)
- Μια άλλη προσέγγιση λέει ότι συνήθως τα *ουσιαστικά* είναι εκείνα που περιγράφουν κυρίως το νόημα μια πρότασης
  - Εκ τούτου θα μπορούσαμε να λάβουμε υπόψη (στην κατασκευή του ευρετηρίου) μόνο τα ουσιαστικά και άρα να παραλείψουμε τις αντωνυμίες, τα ρήματα και τα επίθετα.
  - Επίσης μπορούμε να θεωρήσουμε ομάδες ουσιαστικών που εμφανίζονται μαζί, π.χ. “computer science” , ως έναν όρο ευρετηρίου.
- Τέλος μια άλλη προσέγγιση είναι να καθορίσουμε το σύνολο των όρων ευρετηρίων από ελεγχόμενα λεξιλόγια (Θησαυρούς όρων)

Thesauri

Consists of

(1) Precompiled list of important words in a given domain of knowledge

(2) For each word in the list, a set of related words

Using a controlled vocabulary for indexing and searching