Ανάκληση Πληροφορίας

Information Retrieval

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 6η

Διόρθωση πληκτρολόγησης

Διόρθωση πληκτρολόγησης

- Δυο κύριες χρήσεις
  - Διόρθωση εγγράφων που θα εισαχθούν στον inverted index
  - Διόρθωση ερωτημάτων χρηστών για ν' ανακτηθούν οι "σωστές" απαντήσεις
- Σε δυο γεύσεις:
  - Μεμονωμένη λέξη
    - Έλγχος κάθε λέξης για σφάλματα
    - Δεν θα εντοπίσει σφάλματα που εμπλέκουν σωστές λέξεις
    - Π.χ., ***from → form***
  - Ευαισθησία στο "περιβάλλον" της λέξεως
    - Έλεγχος και των περιβαλλόντων λέξεων
    - Π.χ., ***I flew <u>form</u> Heathrow to Narita***

Διόρθωση εγγράφου

- Ειδικά χρήσιμο για έγγραφα που προέκυψαν από διαδικασίες OCR
  - Ειδικοί αλγόριθμοι διόρθωσης
  - Μπορεί να χρησιμοποιηθεί γνώση ειδική σε κάθε πεδίο
    - Π.χ., το OCR μπορεί να μπερδέψει το O και D συχνότερα από ότι το O με το I (πλαϊνά στο QWERTY πληκτρολόγιο)
- Web pages και τυπωμένο κείμενο έχουν λάθη
- Στόχος: το dictionary να περιέχει λιγότρεα misspellings
- Αλλά συχνά δεν αλλάζουμε τα έγγραφα, αλλά στοχεύσουμε να διορθώσουμε την αντιστοίχιση ερωτήματος-εγγράφου

Mis-spellings σε ερωτήματα

- Το κύριο ενδιαφέρον μας
  - Π.χ., το ερώτημα ***dimitros katsaros***
- Μπορούμε είτε να
  - Ανακτήστουμε τα έγγραφα που θα προέκυπταν από το σωστό spelling, ή
  - Επιστρέψουμε συστάσεις εναλλακτικών ερωτημάτων με το ορθό spelling
    - *Did you mean* ***dimitrios*** *katsaros?*

Διόρθωση μεμονωμένης λέξεως

- Θεμελιώδης προϋπόθεση – υπάρχει ένα lexicon από το οποίο προκύπτουν τα ορθά spellings
- Δυο βασικές επιλογές γι' αυτό
  - Ένα τυπικό lexicon όπως το:
    - Webster's English Dictionary
    - Ένα "ειδικο βιομηχανικό" lexicon – συντηρούμενο χειρωνακτικά
  - Το lexicon της indexed συλλογής εγγράφων
    - Π.χ., όλες οι λέξεις του Web
    - Όλα τα ονόματα, ακρωνύμια, κ.τ.λ.
    - (Περιλαμβανομένων των mis-spellings)

Διόρθωση μεμονωμένης λέξεως

- Δεδομένου ενός lexicon και μιας ακολουθίας χαρακτήρων Q, επίστρεψε τις λέξεις του lexicon που ειναι "κοντινότερες" στην Q
- Τι σημαίνει "κοντινότερη";
- Θα μελετήσουμε διάφορες εναλλακτικές
  - Edit distance (Levenshtein distance)
  - Weighted edit distance
  - Επικάλυψη *n*-gram

Edit distance

- Δεδομένων δυο strings S_1 και S_2, ο ελέχιστος αριθμός λειτουργιών που απαιτείται για να μετατραπεί το ένα στο άλλο
- Οι λειτουργίες είναι συνήθως σε επίπεδο χαρακτήρα
  - Insert, Delete, Replace, (Transposition)
- Π.χ., η edit distance από το ***dof*** στο ***dog*** είναι 1
  - Από το ***cat*** στο ***act*** είναι 2 (Μόνο 1 με transpose)
  - Από το ***cat*** στο ***dog*** είναι 3.
- Υπολογίζεται με δυναμικό προγραμματισμό
- Δείτε
  - http://www.let.rug.nl/kleiweg/lev/
  - http://www.ripelacunae.net/projects/levenshtein/

Weighted edit distance

- Όπως προηγουμένως, αλλά το βάρος μιας λειτουργίας εξαρτάται από τον/τους χαρακτήρα/ρες που εμπλέκονται
  - Για να "πιάσει" λάθη OCR ή πληκτρολόγησης, π.χ., το ***m*** είναι πιο πιθανό να γίνει mis-typed ως ***n*** παρά ως ***q***
  - Επομένως, αντικαθιστώντας το ***m*** με το ***n*** έχει ως συνέπεια μικρότερη edit distance από ότι το ***q***
- Απαιτεί πίνακα βαρών ως είσοδο
- Τροποποίηση του δυναμικού προγραμματισμού για τον χειρισμό βαρών

Χρήση της edit distances

- Δεδομένου ενός ερωτήματος, πρώτα απαριθμούμε όλες τις ακολουθίες χαρακτήρων που απέχουν μια προκαθορισμένη (weighted) edit distance (π.χ., 2)
- Τέμνουμε αυτό το σύνολο με την λίστα των "σωστών" λέξεων
- Δείχνουμε στον χρήστη τους όρους που βρέθηκαν
- Εναλλακτικά,
  - Βρίσκουμε όλα τα έγγραφα που ταιριάζουν με όλες τις πιθανές διορθώσεις (πολύ αργή επιλογή)
  - Εκτελούμε την πιο πιθανό διόρωση
- Οι εναλλακτικές στερούν δύναμη από τον χρήστη, αλλά γλιτώνουν επιπλέον αλληλεπίδραση

Edit distance προς όλους τους όρους του dictionary;

- Δεδομένου ενός (mis-spelled) ερωτήματος – πρέπει να υπολογίσουμε την edit distance του προς κάθε όρου του dictionary;
  - Ακριβό και αργό
  - Εναλλακτικές;
- Πώς ελαττώνουμε το σύνολο των υποψηφίων όρων του dictionary;
- Μια δυνατότητα είναι να χρησιμοποιήσουμε την επικάλυψη των *n*-gram

n-gram overlap

- Απαριθμούμε όλα τα *n*-grams του υποβληθέντος ερωτήματος, καθώς επίσης και στο lexicon
- Χρησιμοποιούμε τον *n*-gram index (θυμηθείτε την αναζήτηση με wild-cards) για να ανακτήσουμε όλους τους όρους του lexicon που ταιριάζουν με οποιοδήποτε από τα *n*-grams του ερωτήματος
- Θέτουμε ένα κατώφλι στον αριθμό των *n*-grams που ταιριάζουν

Παράδειγμα με trigrams

- Υποθέστε ότι το κείμενο είναι το ***november***
  - Τα trigrams είναι τα: *nov, ove, vem, emb, mbe, ber*
- Το ερώτημα είναι το ***december***
  - Τα trigrams είναι τα: *dec, ece, cem, emb, mbe, ber*
- Συνεπώς, 3 trigrams επικαλύπτονται (από το σύνολο των 6 σε κάθε όρο)
- Πώς μπορούμε να μετατρέχουμε αυτό σε ένα κανονικοποιημένο μέτρο επικάλυψης;

Μια επιλογή – Συντελεστής Jaccard

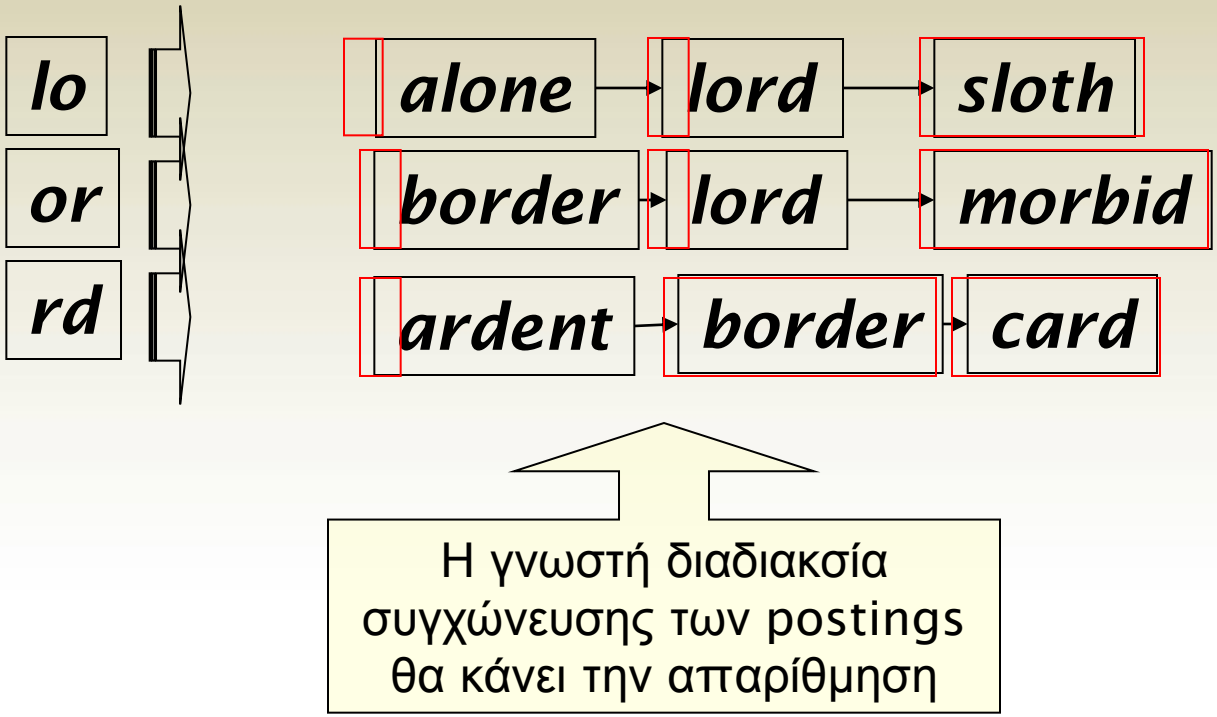
- Ένα κοινώς χρησιμοποιούμενο μέτρο επικάλυψης
- Έστωσαν X και Y δυο σύνολα, τότε ο J.C. ορίζεται ως:

$$|X \cap Y| / |X \cup Y|$$

- Ισούται με 1, όταν τα X και Y έχουν τα ίδια στοιχεία, και με 0 όταν είναι διακριτά
- Τα X και Y δεν είναι ανάγκη να είναι ίδιου μεγέθους
- Δίνει ως αποτέλεσμα έναν αριθμό μεταξύ 0 και 1
  - Θέτουμε ένα κατώφλι για να αποφασίσουμε εάν υπάρχει "ταίριασμα"
    - Π.χ., εάν J.C. > 0.8, δηλώνουμε "ταίριασμα"

Matching trigrams

- Θεωρήστε το ερώτημα ***lord*** – επιθυμούμε να αναγνωρίσουμε λέξεις που ταιριάζουν με 2 από τα 3 bigrams της (***lo, or, rd***)



Διόρθωση εξαρτώμενη από το περιβάλλον

- Κείμενο: ***I flew <u>from</u> Heathrow to Narita.***
- Θεωρήστε το ερώτημα φράσεως ***"flew <u>form</u> Heathrow"***
- Θα επιθυμούσαμε να αποκριθούμε ως εξής:

  Did you mean "***flew from Heathrow***"?

  επειδή κανένα έγγραφο δεν ταίριαξε με το ερώτημα φράσεως

Διόρθωση εξαρτώμενη από το περιβάλλον

- Χρειαζόμαστε γνώση του “περιβάλλοντος” για να το κάνουμε
- Πρώτη ιδέα: ανάκτηση των όρων του dictionary που είναι κοντινοί (με όρους της weighted edit distance) σε κάθε όρο του ερωτήματος
- Δοκιμάζουμε όλες τις δυνατές προκύπτουσες φράσεις “διορθώνοντας” έναν όρο κάθε φορά
  - ***flew from heathrow***
  - ***fled form heathrow***
  - ***flea form heathrow***
- **Hit-based spelling correction:** Συστήνουμε την εναλλακτική που έχει πολλά hits (δηλαδή επιστρέφει τα πιο πολλά έγγραφα)

Μια άλλη προσέγγιση

- Διάσπαση του ερωτήματος φράσεως σε σύζευξη biwords (προηγούμενες διαλέξεις)
- Αναζήτηση για biwords που απαιτούν διόρθωση μόνο ενός όρου
- Απαρίθμηση ταιριασμάτων φράσεων και … διάταξη αυτών!

Γενικά ζητήματα στην διόρθωση

- Απαριθμούμε πολλαπλές εναλλακτικές για το “Did you mean?”
- Πρέπει ν’ αποφασίσουμε ποια εναλλακτική θα παρουσιάσουμε στον χρήστη
- Χρήστη ευριστικών
  - Η εναλλακτική που επιστρέφει τα πιο πολλά έγγραφα
  - Query log analysis + tweaking
    - Για εξαιρετικά δημοφιλή, topical ερωτήματα
- Η διόρθωση είναι εξαιρετικά ακριβή υπολογιστικά
  - Ν’ αποφεύγεται η εκτέλεσή της για κάθε ερώτημα
  - Να εκτελείται μόνο για τα ερωτήματα που επέστρεψαν πολύ λίγα έγγραφα

ΣΥΓΚΕΝΤΡΩΤΙΚΑ: Ποια ερωτήματα μπορούμε να επεξεργαστούμε;

- Έχουμε λοιπόν
  - Positional inverted index με skip δείκτες
  - Wild-card ευρετήριο
  - Διόρθωση πληκτρολόγησης
  - Soundex
- Ερωτήματα όπως:

(SPELL(moriset) /3 toron*to) OR SOUNDEX(chaikofski)