6η Σειρά Ασκήσεων

Δομές Ευρετηριοποίησης
Ημερομηνία Παράδοσης: Τρίτη 8 Δεκεμβρίου 2009 στο μάθημα

Άσκηση 1 (ατομική)
Δοκιμάστε στο google:
την ερώτηση «βάσεις δεδομένων» και βάσεις δεδομένων (χωρίς εισαγωγικά) (πρόκειται για ζεύγη λέξεων που εμφανίζονται συχνά μαζί)
την ερώτηση «βάσεις κιάμος» και βάσεις κιάμος (χωρίς εισαγωγικά) (πρόκειται για ζεύγη λέξεων που δεν εμφανίζονται συχνά μαζί)
την ερώτηση «βάσεις πορτοκάλι» και βάσεις πορτοκάλι (χωρίς εισαγωγικά) (πρόκειται για ζεύγη λέξεων που δεν εμφανίζονται συχνά μαζί)

(i) Τι μπορείτε να πείτε σχετικά με τις phrase queries και τον χειρισμό τους στο google;
(ii) Σημειώστε και άλλες τυχόν παρατηρήσεις σχετικά με proximity, stemming, απομάκρυνση stop words, αυτόματη διόρθωση, κλπ. που έχουμε συζητήσει στο μάθημα (τουλάχιστον δύο).

Άσκηση 2 (ατομική)
(i) Δώστε σε ψευτοκώδικα τον αλγόριθμο difference(p1, p2) όπου p1 και p2 είναι δυο διατεταγμένες posting lists (παρόμοια με τον αλγόριθμο 1.6 για το intersect(p1, p2) του online βιβλίου).
(ii) Μπορούν οι skip pointers να βελτιώσουν τον ψευτοκώδικά σας; Αν όχι, γιατί. Αν ναι, δώστε ένα τροποποιημένο αλγόριθμο difference(p1, p2) που να τους χρησιμοποιεί (εξηγείστε σύντομα την τροποποίηση).

Άσκηση 3 (ατομική)
Άσκηση 1.6 του online βιβλίου.

Άσκηση 3 (ατομική)
Άσκηση 2.10 του online βιβλίου.

Άσκηση 4 (ατομική)
Θεωρείστε τους όρους manic, magic, moon, και mongolic.
(i) Δώστε το permuterm dictionary. Εξηγείστε ποιοι όροι του ταιριάζουν στην ερώτηση ma*ic. Εξηγείστε πως αποτιμάται η ερώτηση m*a*ic και αν υπάρχουν false positive.
(ii) Κατασκευάστε τα posting lists για 2-grams. Εξηγείστε ποιες posting lists πρέπει να κοιτάξτε για τον υπολογισμό της ma*ic.