1η Σειρά Ασκήσεων

Εισαγωγή. Μοντέλα Ανάκτησης
Ημερομηνία Παράδοσης: Τρίτη 27 Οκτωβρίου2009 στο μάθημα
Ομάδες των 2 ατόμων
Άριστα το 10

Άσκηση 1 [3]

Προσπαθήστε να βρείτε μια ερώτηση της μορφής [query-term-1 query-term-2] (χωρίς εισαγωγικά) που όταν τρέξει στο Google παράγει τουλάχιστον ένα αποτέλεσμα που δεν περιέχει τον ένα από τους δύο όρους. Δηλαδή, βρείτε μια ερώτηση που δεν ερμηνεύεται ως AND.
Δώστε την ερώτηση και σημειώστε για καθένα από τα πρώτα 20 αποτελέσματα της 2 αν περιέχει και τους δυο όρους της ερώτησης, 1 αν περιέχει το πολύ έναν από αυτούς και 0 αν δεν περιέχει κανένα (δηλαδή, δώστε μια λίστα της μορφής: 1^o αποτέλεσμα χ, 2^o αποτέλεσμα χ, …20^o αποτέλεσμα χ (όπου χ = 0, 1 ή 2).
Με βάση τα παραπάνω, σχολιάστε το μοντέλο ανάκτησης που πιθανώς χρησιμοποιεί η Google.

Άσκηση 2 [προαιρετική +3]

Δοθείσας μιας ερώτησης *q* και εγγράφων d_1, d_2, … τα οποία θεωρούμε ως διανύσματα μπορούμε να διατάξουμε τα έγγραφα με βάση την αύξουσα Ευκλείδεια απόστασή τους από την *q*.
(α) Δείξτε ότι αν η ερώτηση *q* και τα διανύσματα d_i είναι κανονικοποιημένα, τότε η διάταξη που παράγεται είναι ίδια με αυτήν αν χρησιμοποιήσουμε ομοιότητα συνημίτονου.
(β) Συγκρίνετε την καταλληλότητα των δύο αποστάσεων για το χαρακτηρισμό της ομοιότητας μεταξύ ενός εγγράφου και μιας ερώτησης.

Άσκηση 3 [7]

Θεωρείστε τα έγγραφα:
d_1 : *a b*
d_2 : *a b a b*
d_3 : *a b a b c*
d_4 : *a b c*
d_5 : *a a c*
και την ερώτηση
q : *a b*

(α) Διατάξατε τα έγγραφα με βάση το ποιο σας φαίνεται διαισθητικά πιο σχετικό με την ερώτηση.
(β) Θεωρείστε τις ερωτήσεις και τα έγγραφα ως (πολύ) σύνολα όρων (bag) και τις ακόλουθες συναρτήσεις διαβάθμισης (ranking functions):

$$R_1(d, q) = \frac{|d \cap q|}{|q|} \quad R_2(d, q) = \frac{|d \cap q|}{|d|} \quad R_3(d, q) = \frac{|d \cap q|}{|d| + |q|} \quad R_4(d, q) = \frac{|d \cap q|}{|d \cup q|} \quad R_5(d, q) = |d \cup q|$$

Δώστε τη διάταξη των εγγράφων με βάση αυτές
(γ) Κατασκευάστε τη διανυσματική αναπαράσταση των εγγράφων με βάρη TF-IDF. Διατάξτε τα έγγραφα με βάση αυτό το μοντέλο.