Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2007-2008 Εαρινό Εξάμηνο

2η Σειρά Ασκήσεων (Μοντέλα Ανάκτησης)
Ανάθεση: 18 Μαρτίου
Παράδοση: 1 Απριλίου (2 εβδομάδες)

Άσκηση 1 (2 βαθμοί)
(α) Δώστε τη διανυσματική παράσταση των εγγράφων $d_1, \ldots, d_5$ με βάρη TF-IDF. Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται κατά αλφαβητική σειρά.
(β) Δώστε την απάντηση που θά έχει η κάθε επερώτηση $q_1, \ldots, q_4$ βάσει του διανυσματικού μοντέλου.
(γ) Σχεδιάστε τη μορφή που θα έχει το ανεστραμμένο ευρετήριο για τη συλλογή D.

Documents D
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"

Queries Q
q_1 : "a b"
q_2 : "a"
q_3 : "c"
q_4 : "a c"

Άσκηση 2 (2 βαθμοί)
Έστω οτι έχουμε ένα μοντέλο ανάκτησης το οποίο βλέπει τα έγγραφα και τις επερωτήσεις ως σύνολα όρων. Συγκρίνετε τις ακόλουθες συναρτήσεις διαβάθμισης

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

Μπορείτε να στηρίξετε την απάντησή σας παραθέτοντας παραδείγματα.

Άσκηση 3 (2 βαθμοί)
Θεωρείστε το ακόλουθο ευρετήριο το οποίο έχει τη μορφή: term: doc1: ⟨ position1, position2, . . . ⟩; doc2: ⟨ position1, position2, . . . ⟩; κλπ.

$$\begin{aligned} angels &: 2:\langle 36,174,252,651\rangle; 4:\langle 12,22,102,432\rangle; 7:\langle 17\rangle; \\ fools &: 2:\langle 1,17,74,222\rangle; 4:\langle 8,78,108,458\rangle; 7:\langle 3,13,23,193\rangle; \\ fear &: 2:\langle 87,704,722,901\rangle; 4:\langle 13,43,113,433\rangle; 7:\langle 18,328,528\rangle; \\ in &: 2:\langle 3,37,76,444,851\rangle; 4:\langle 10,20,110,470,500\rangle; 7:\langle 5,15,25,195\rangle; \\ rush &: 2:\langle 2,66,194,321,702\rangle; 4:\langle 9,69,149,429,569\rangle; 7:\langle 4,14,404\rangle; \\ to &: 2:\langle 47,86,234,999\rangle; 4:\langle 14,24,774,944\rangle; 7:\langle 199,319,599,709\rangle; \\ tread &: 2:\langle 57,94,333\rangle; 4:\langle 15,35,155\rangle; 7:\langle 20,320\rangle; \\ where &: 2:\langle 67,124,393,1001\rangle; 4:\langle 11,41,101,421,431\rangle; 7:\langle 16,36,736\rangle; \end{aligned}$$

Ποιά έγγραφα (αν υπάρχουν) ικανοποιούν τα επόμενα ερωτήματα·
q_1: "fools rush in"
q_2: "fools rush in" AND "angels fear to tread"
Τα εισαγωγικά σηματοδοτούν phrase queries.

Άσκηση 4 (4 βαθμοί)
Θεωρείστε μια μηχανή αναζήτησης η οποία στηρίζεται σε ανεστραμμένο ευρετήριο και υποστηρίζει το Λογικό Μοντέλο (Boolean Model). Για κάθε λέξη το ευρετήριο μας δίνει μια μέθοδο που μας επιστέφει όλα τα αναγνωριστικά των εγγράφων που την περιέχουν σε αύξουσα σειρά. Θεωρείστε τις εξής επερωτήσεις:

q_1: μέλι AND ουρανός

q_2: μέλι AND ουρανός AND πορτοκάλι

q_3: (πορτοκάλι OR δέντρα) AND (μέλι OR ουρανός) AND (όραση OR μάτια)

Επίσης θεωρείστε τα εξής μεγέθη των λιστών εμφάνισης (posting list sizes):

Όρος	Μέγεθος Λίστας
δέντρα	316812
μάτια	213312
μέλι	107913
όραση	87009
ουρανός	271658
πορτοκάλι	46653

(α) Δώστε ένα γρήγορο αλγόριθμο για την αποτίμηση του q_1

(β) Δώστε ένα γρήγορο αλγόριθμο για την αποτίμηση του q_2 (εστιάστε στη σειρά με την οποία συμφέρει να γίνει η αποτίμηση)

(γ) Δώστε ένα γρήγορο αλγόριθμο για την αποτίμηση του q_3 (εστιάστε στη σειρά με την οποία συμφέρει να γίνει η αποτίμηση)

Σημείωση: Για τον υπολογισμό των παραπάνω μπορείτε να φτιάξετε σχετικό πρόγραμμα (χωρίς αυτό να είναι απαραίτητο).