Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2006-2007 Εαρινό Εξάμηνο

2η Σειρά ασκήσεων
(Μοντέλα Ανάκτησης Πληροφοριών και Ευρετήρια)
Ανάθεση: 30 Μαρτίου
Παράδοση: 21 Απριλίου

Άσκηση 1 (30 βαθμοί)

Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:

Έγγραφο 1: «Computer Games»
Έγγραφο 2: «Computer Games Computer Games»
Έγγραφο 3: « Games Theory and Computer »
Έγγραφο 4: «Computer for Computer »
Έγγραφο 5: «Cheap Games Computer Games»

1) Δώστε τη διανυσματική παράσταση του κάθε εγγράφου με βάρη TF-IDF. Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται κατά αλφαβητική σειρά.
2) Θεωρείστε την επερώτηση q1=«Computer Games». Υπολογίστε το TF-IDF διάνυσμα αυτής της επερώτησης και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.

Σχεδιάστε το ανεστραμμένο ευρετήριο για αυτή τη συλλογή.

Άσκηση 2 (40 βαθμοί)

Έστω μια συλλογή από κείμενα D, και έστω A ένα διατεταγμένο υποσύνολο αυτής. Έστω ότι μας δίνουν το A και μας ζητούν να βρούμε αν υπάρχει επερώτηση q τ.ω. η απάντηση της να έχει στην αρχή της το διατεταγμένο σύνολο A. Για παράδειγμα, αν A=⟨d1, d2, d3⟩ και βρούμε μια επερώτηση q τ.ω. Answer(q)= ⟨d1, d2, d3, d8, …⟩ τότε αυτή είναι μια λύση του προβλήματος μας. Θεωρώντας ότι το σύστημα σας βασίζεται στο διανυσματικό μοντέλο, απαντήστε τα παρακάτω ερωτήματα.

(α) Πως μπορούμε να βρούμε αν υπάρχει τέτοια επερώτηση;
(β) Αν υπάρχει ποια είναι;
(γ) Αν δεν υπάρχει τέτοια επερώτηση, πως θα χαλαρώνατε το πρόβλημα και τι θα μπορούσατε να επιστρέψετε; Μπορείτε να αναπτύξετε τις σκέψεις σας όσο θέλετε.

Σημείωση: Προσέξτε ώστε το υπολογιστικό κόστος των λύσεων που θα προτείνετε για τα (α) και (β) να μην είναι απαγορευτικό.

Άσκηση 3 (30 βαθμοί)

Στο μάθημα είδαμε δύο μοντέλα ανάκτησης του βασίζονται στη Θεωρία Ασαφών Συνόλων. Το πρώτο θεωρεί βάρυνση TF*IDF, ενώ το δεύτερο είναι εκείνο που προτάθηκε από τους [Ogawa, Morita, Kobayashi, 1991]. Θεωρείστε έναν όρο t_i ενός εγγράφου d_j. Συγκρίνετε την συμπεριφορά των δύο αυτών μοντέλων για διάφορες περιπτώσεις, π.χ.: για μικρές και μεγάλες τιμές του tf_{ij}, για μικρές και μεγάλες τιμές του idf_i , για μικρές και μεγάλες τιμές του w_{ij} αν προκύπτει από tf*idf.

Καλή εργασία