Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2004-2005 Εαρινό Εξάμηνο

2η Σειρά ασκήσεων
(Μοντέλα Ανάκτησης Πληροφοριών και Ευρετήρια)
Ανάθεση: 16 Μαρτίου
Παράδοση: 29 Μαρτίου

Άσκηση 1 (40 βαθμοί) (Διανυσματικό Μοντέλο)

Θεωρείστε μια συλλογή κειμένων που περιέχει τα ακόλουθα 5 έγγραφα:

Έγγραφο 1: «New Year»
Έγγραφο 2: « New Year New Year »
Έγγραφο 3: «Financial New Times»
Έγγραφο 4: «Financial Year»

1) Δώστε τη διανυσματική παράσταση του κάθε εγγράφου με βάρη TF-IDF (για ευκολία θεωρήστε ότι IDF=N/DF και όχι IDF=log(N/DF)). Θεωρείστε ότι η θέση της κάθε λέξης στα διανύσματα γίνεται κατά αλφαβητική σειρά.
2) Θεωρείστε την επερώτηση q1=«new financial». Υπολογίστε το TF-IDF διάνυσμα αυτής της επερώτησης και δώστε την διάταξη των εγγράφων που θα επιστρέψει ένα σύστημα που βασίζεται στο διανυσματικό μοντέλο.
3) Σχεδιάστε το ανεστραμμένο αρχείο για αυτή τη συλλογή.

Άσκηση 2 (40 βαθμοί) (συνάρτηση διαβάθμισης)

Θεωρείστε ένα Σύστημα Ανάκτησης Πληροφοριών (ΣΑΠ) από μια μεγάλη συλλογή κειμένων. Θέλουμε να δώσουμε τη δυνατότητα χρήσης του ΣΑΠ μέσω κινητού τηλεφώνου. Για το λόγο αυτό θέλουμε να ορίσουμε μια συνάρτηση διαβάθμισης (ranking function) η οποία να ευνοεί τα μικρά κείμενα, αφενός για να κρατήσουμε σε χαμηλά επίπεδα τον όγκο δεδομένων που θα μεταφέρονται και αφετέρου διότι οι χρήστες κινητών τηλεφώνων προτιμούν τα μικρά κείμενα (ένεκα του μικρού μεγέθους της οθόνης). Θεωρείστε ότι οι επερωτήσεις των χρηστών είναι σάκοι λέξεων (bag of words). Σχεδιάστε μια συνάρτηση διαβάθμισης για το σκοπό αυτό για κάθε μια από τις παρακάτω περιπτώσεις

(α) Το ευρετήριο του ΣΑΠ έχει δυαδικά (0,1) βάρη (όπως για παράδειγμα το ευρετήριο του Boolean μοντέλου)
(β) Το ευρετήριο έχει βάρη TF-IDF.

Τεκμηριώστε τις προτάσεις σας (με αποδείξεις ή παραδείγματα).

Άσκηση 3 (20 βαθμοί)

Έστω ένα ΣΑΠ που βασίζεται στο διανυσματικό μοντέλο, το οποίο υποστηρίζει τον κλασσικό τρόπο αλληλεπίδρασης (ο χρήστης διατυπώνει επερώτηση και το σύστημα επιστρέφει ένα διατεταγμένο σύνολο εγγράφων) συν μια λειτουργία *προσαρμογής ευρετηρίου*. Συγκεκριμένα ο χρήστης μπορεί να αλλάξει τη θέση ενός εγγράφου που εμφανίζεται σε μια απάντηση (π.χ. από τη $2^η$ θέση να το πάει στην $1^η$ ή στην $18^η$). Μετά από μια τέτοια εντολή, το σύστημα πρέπει να «προσαρμοστεί» στην απαίτηση του χρήστη, τροποποιώντας κατάλληλα το διάνυσμα του εγγράφου που μετακινήθηκε. Η τροποποίηση πρέπει να είναι τέτοια ώστε αν ο χρήστης επανυποβάλει την επερώτηση, τότε το εν λόγω έγγραφο να τοποθετηθεί στη θέση που όρισε ο χρήστης.

Περιγράψτε με ποιο τρόπο θα τροποποιούσατε το διάνυσμα του εγγράφου προκειμένου να επιτύχετε την παραπάνω λειτουργικότητα.

Λάβετε υπόψη ότι αν υπάρχουν πολλοί τρόποι υλοποίησης μιας λειτουργίας προσαρμογής, τότε ως κριτήριο για την επιλογή του καταλληλότερου τρόπου συχνά θεωρείται η αρχή της ελάχιστης αλλαγής.

Υπόδειξη: Θεωρείστε αρχικά ότι το πλήθος των όρων είναι 1, κατόπιν 2 και εν συνεχεία γενικεύστε.

Άσκηση 4 (20 βαθμοί) (Προαιρετική)

Να επεκτείνετε το σύστημα που παρουσιάστηκε στο δεύτερο φροντιστήριο (T2) έτσι ώστε να υποστηρίζει και το Extended Boolean Model. Συγκεκριμένα θα πρέπει να υλοποιήσετε τα εξής:

1. Προσθήκη μιας επιπλέον στήλης στο ResultPanel η οποία θα περιέχει τα αποτελέσματα με βάση το Ranking που προέκυψε από το Extended Boolean Model
2. Προσθήκη μιας επιπλέον συνάρτησης στο SearchEngine η οποία θα συλλέγει το Ranking από το DocumentTreeSet. Η συνάρτηση αυτή θα είναι παρόμοια με την getDocumentRankings και ο σκοπός της θα είναι να στέλνει το Query σε κάθε κείμενο που βρίσκεται στο DocumentTreeSet.
3. Στο DocumentTreeSet θα πρέπει να προσθέσετε τη συνάρτηση που θα υπολογίζει το Similarity μεταξύ του Query και του Document. Στην υλοποίηση αυτής της συνάρτησης θα πρέπει να λάβετε υπόψη, πως πρέπει να γνωρίζεται το Max Inverse Document Frequency για να κανονικοποιήσετε το διάνυσμα με τα βάρη των κειμένων. Επίσης θα πρέπει να ανιχνεύσετε το τελεστή που χρησιμοποίησε ο χρήστης στο Query του ώστε να επιλέξετε το κατάλληλο τύπο για να βρείτε το Similarity (AND/OR). Προαιρετικά μπορείτε να υλοποιήσετε την υποστήριξη πολλών τελεστών με βάση την προτεραιότητα τους.