Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών
2008-2009 Εαρινό Εξάμηνο

4η Σειρά Ασκήσεων

Αξία: 5% του τελικού σας βαθμού

Άσκηση 1 (3%)

Υπολογίστε την απόσταση ενημέρωσης (Edit Distance) μεταξύ των λέξεων «αυτοκινητο» και «αυτοματο». Για να δικαιολογήσετε την απάντησή σας δώστε τον πίνακα τον οποίο δημιουργεί ο αλγόριθμος δυναμικού προγραμματισμού.

Άσκηση 2 (2%)

Θεωρείστε μια συλλογή εγγράφων με απλό κείμενο.
(α) Έστω ο συνολικός αριθμός λέξεων είναι 2.000.000. Ποιο είναι το εκτιμώμενο μέγεθος λεξιλογίου (πλήθος διαφορετικών λέξεων);
(β) Έστω ότι η πιο συχνά εμφανιζόμενη λέξη εμφανίζεται 550.000 φορές. Πόσες φορές εκτιμάτε ότι θα εμφανίζεται η 20η πιο συχνά εμφανιζόμενη λέξη;

Άσκηση 3 (20%)

Θεωρείστε το κείμενο «μια φράση έχει λέξεις και αριθμούς και μια λέξη έχει γράμματα».
(α) Σχεδιάστε το δένδρο καταλήξεων του κειμένου θεωρώντας ως σημεία ευρετηρίου (index points) τις αρχές των λέξεων (μπορείτε να δώσετε κατευθείαν το PATRICIA tree).
(β) Δώστε την κωδικοποίηση του κειμένου κατά Huffman.

Άσκηση 4 (20%)

Θεωρείστε το παρακάτω τμήμα ενός ανεστραμμένου ευρετήριο όπου αποθηκεύονται και οι θέσεις εμφάνισης των λέξεων στα έγγραφα (positional index) με τη μορφή: word: document; document;

Gates:	1; 2; 3; 4;
IBM:	4; 7;
Microsoft:	1; 2; 3; 5;

Δώστε τη συμπιεσμένη μορφή που παραπάνω ευρετηρίου (συμπίεση λιστών εμφανίσεων).

Άσκηση 5 (25%)

Θεωρείστε τα έγγραφα d1…d9 και έστω ότι θέλουμε να τα ομαδοποιήσουμε. Για χάρη της άσκησης θεωρείστε ότι sim(di,dj) = (i + j) / 20.
(α) Ζωγραφίστε το γράφο των εγγράφων θεωρώντας ως κατώφλι ομοιότητας την τιμή 0.4.
(β) Δώστε το αποτέλεσμα της ιεραρχικής ομαδοποίησης κατά single link.
(γ) Δώστε το αποτέλεσμα της ιεραρχικής ομαδοποίησης κατά complete link.
(δ) Επαναλάβετε τα παραπάνω θεωρώντας ως κατώφλι ομοιότητας την τιμή 0.8.

Άσκηση 6 (30%)

Προσπαθήστε μέσω διαδικτύου να βρείτε όσο το δυνατόν περισσότερα στοιχεία σχετικά με το μέγεθος του ελληνικού παγκόσμιου ιστού (αριθμός σελίδων, μέσο μέγεθος σελίδων, πλήθος συνδέσμων, συνολικός όγκος, και όποια άλλη χρήσιμη και αξιόπιστη μέτρηση βρείτε). Σκοπός μας είναι η σχεδίαση του ευρετηρίου μιας μηχανής αναζήτησης για τον ελληνικό ιστό, το οποίο να μπορεί να φιλοξενηθεί στο μηχάνημα που έχουμε στη διάθεση μας αυτή τη στιγμή (κύρια μνήμη: 2 GBytes, σκληρός δίσκος: 150 GB). Βάσει αυτών που έχουμε δει στο μάθημα, εκτιμήστε το μέγεθος που θα έχει το λεξιλόγιο και οι λίστες εμφάνισης αν αποφασίζαμε να χρησιμοποιήσουμε μια δομή ανεστραμμένου αρχείου.
Περιγράψτε όποια άλλη εναλλακτική ή συμπληρωματική δομή ευρετηρίου κρίνετε ότι μπορεί να επιταχύνει τη λειτουργία της μηχανής αναζήτησης.
Σημείωση: Ψάξτε μήπως κάποιο πρόσφατο άρθρο περιέχει τέτοιες πληροφορίες