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

3η Σειρά Ασκήσεων
Ανάθεση: 12 Μαίου
Παράδοση: 30 Μαίου

Άσκηση 1 (2 βαθμοί)
Θεωρείστε ένα έγγραφο με περιεχόμενο:

> **«στο σημερινό μάθημα μάθαμε περισσότερα σε σχέση με το προηγούμενο μάθημα»**

Αγνοώντας τους τόνους, σχεδιάστε
(α) το trie του λεξιλογίου του παραπάνω εγγράφου,
(β) το δένδρο καταλήξεων του παραπάνω εγγράφου θεωρώντας ως σημεία ευρετηρίου (index points) τις αρχές των λέξεων, και
(γ) συμπτύξτε το παραπάνω δένδρο καταλήξεων στη μορφή ενός Patricia tree.

Άσκηση 2 (2 βαθμ.)
Θεωρείστε το αλφάβητο {α, β, γ, δ, ε, ζ} και την εξής πρόταση:

> **¨α α β β γ α γ α δ α α α δ β ε β ε ζ¨.**

α) Βάσει αυτής της φράσης ποια είναι η εντροπία του αλφαβήτου;
β) Δώστε τη συμπιεσμένη μορφή της φράσης χρησιμοποιώντας κανονικοποιημένους κώδικες Huffman.

Άσκηση 3 (1 βαθμός)
Υπολογίστε την Edit Distance μεταξύ των λέξεων paris και alice. Δώστε τον 5x5 πίνακα που περιγράφει τον τρόπο λειτουργίας του σχετικού αλγορίθμου δυναμικού προγραμματισμού.

Άσκηση 4 (2 βαθμ.)
Θεωρείστε μια συλλογή εγγράφων στην οποία εμφανίζονται 400 διαφορετικές λέξεις και η συχνότητα της πιο συχνά εμφανιζόμενης ισούται με 900.
(α) Εκτιμήστε τον αριθμό εμφανίσεων της 10ης πιο συχνά εμφανιζόμενης λέξης, και της 20ης.
(β) Αποφασίζετε να φτιάξετε ένα ανεστραμμένο ευρετήριο μόνο για τις 200 πιο συχνά εμφανιζόμενες λέξεις. Πόσο μικρότερο να είναι το μέγεθος του σε σχέση με εκείνο για όλες τις λέξεις; Θεωρείστε οτι δεν κάνουμε απαλοιφή λέξεων αποκλεισμού, ούτε στελέχωση.

Άσκηση 5 (3 βαθμ.)
Θεωρείστε τα ακόλουθα έγγραφα όπου τα γράμματα A-E συμβολίζουν λέξεις.

$$
\begin{aligned}
d1 &= \text{"}A\ A\ B\text{"} \\
d2 &= \text{"}B\ A\ A\text{"} \\
d3 &= \text{"}A\ B\ C\text{"} \\
d4 &= \text{"}C\ C\ B\text{"} \\
d5 &= \text{"}C\ D\ A\text{"} \\
d6 &= \text{"}D\ E\text{"} \\
d7 &= \text{"}A\ A\ B\text{"} \\
d8 &= \text{"}E\ B\text{"}
\end{aligned}
$$

Έστω ότι τα $d1, d2, d3$ ανήκουν σε ένα σύστημα $S1$, τα $d4, d5$ σε ένα σύστημα $S2$, και τα υπόλοιπα $(d6, d7, d8)$ σε ένα σύστημα $S3$. Θέλουμε να φτιάξουμε έναν μεσίτη M πάνω από αυτά τα συστήματα.

(α) Για την επιλογή πηγής ο M θέλει να περιγράψει τα περιεχόμενα της κάθε πηγής με ένα διάνυσμα. Δώστε τα διανύσματα πηγών των $S1$, $S2$ και $S3$.

(β) Έστω ότι ο M έχει ήδη τα διανύσματα πηγών των $S1, S2, S3$ και λαμβάνει την επερώτηση $q =$ "$B\ C$". Αν θέλει να προωθήσει την επερώτηση q σε μία μόνο πηγή, ποια θα επιλέξει;

(γ) Ο M λαμβάνει μια επερώτηση, την προωθεί σε όλες τις πηγές, και λαμβάνει τα εξής αποτελέσματα από την κάθε μια:
$S1$: $\langle d1, d2, d3\rangle$
$S2$: $\langle d5, d4\rangle$
$S3$: $\langle d8, d7, d6\rangle$
Δώστε την ενοποιημένη διάταξη κατά round robin interleaving

(δ) Προκειμένου ο μεσίτης να λαμβάνει από τις πηγές απαντήσεις με συγκρίσιμα σκορ, αποφασίζει να κάνει αποτίμηση επερωτήσεων σε δυο φάσεις ώστε οι πηγές να λαμβάνουν τα καθολικά στατιστικά που χρειάζονται για το σωστό υπολογισμό των σκορ. Δώστε το idf του κάθε όρου στην καθολική συλλογή εγγράφων.

(ε) Ο μεσίτης βρίσκει άλλο ένα σύστημα $S4$ το οποίο έχει την ίδια συλλογή με αυτήν του $S1$, δηλαδή και αυτό παρέχει πρόσβαση στα έγγραφα $d1, d2, d3$. Έστω ότι ο M προωθεί μια επερώτηση q στα $S1$ και $S4$ και λαμβάνει τις εξής απαντήσεις:
$S1$: $\langle d2, d1, d3\rangle$
$S4$: $\langle d2, d3, d1\rangle$
Ποιο είναι το κορυφαίο έγγραφο αν ενοποιήσουμε τις διατάξεις: (ι) κατά Borda, (ιι) κατά Condorcet;
Ο M αποφασίζει να δίνει στο χρήστη όχι μόνο την ενοποιημένη διάταξη, αλλά και την Kemeny distance μεταξύ των διατάξεων που έλαβε από τα υποσυστήματα (προκειμένου ο χρήστης να παίρνει μια γεύση για το βαθμό συμφωνίας των πηγών). Ποια είναι αυτή η απόσταση στην προκειμένη;

(στ) Τα συστήματα $S1, S2, S3$ δεν θέλουν πλέον να έχουν ανάγκη τον M και αποφασίζουν να¨ανεξαρτητοποιηθούν' φτιάχνοντας ένα σύστημα ομοτίμων (P2P), συγκεκριμένα ένα δομημένο σύστημα τύπου Chord. Προσελκύουν μάλιστα άλλα δυο συστήματα $S5$ και $S6$ (τα οποία δεν έχουν καμία συλλογή εγγράφων). Αποφασίζουν να χρησιμοποιήσουν μια συνάρτηση κατακερματισμού h των 3 bits, και έστω ότι: $h(IPaddress(S1)) = 1$, $h(IPaddress(S2)) = 2$, $h(IPaddress(S3)) = 4$, $h(IPaddress(S5)) = 5$, $h(IPaddress(S6)) = 6$. Αποφασίζουν να διανείμουν το ανεστραμμένο ευρετήριο θεωρώντας κάθε όρο σαν κλειδί και έστω ότι $h(A) = 2$, $h(B) = 3$, $h(C) = 4$ $h(D) = 4$, $h(E) = 4$.
Δώστε (ι) τους πίνακες δρομολόγησης των κόμβων $S1$ και $S3$ και (ιι) πως θα κατανεμηθεί το ανεστραμμένο ευρετήριο στους κόμβους του δικτύου (δείξτε τι ακριβώς θα έχει κάθε κόμβος).-