Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών
2009-2010 Εαρινό Εξάμηνο

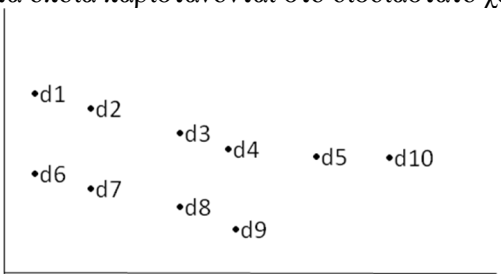
3η Σειρά Ασκήσεων

Αξία: 5% του τελικού σας βαθμού
Ημερομηνίες: 8/12 – 8/1

Άσκηση 1 (20%)
Θεωρείστε το κείμενο «one car has at least one door one wheel and one steering wheel».
(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 2 (5%)
Θεωρείστε μια συλλογή εγγράφων με απλό κείμενο.
(α) Έστω ο συνολικός αριθμός εμφανίσεων λέξεων είναι 10.000.000. Ποιο είναι το εκτιμώμενο μέγεθος λεξιλογίου (πλήθος διαφορετικών λέξεων);
(β) Έστω ότι η πιο συχνά εμφανιζόμενη λέξη εμφανίζεται 3.600.000 φορές. Πόσες φορές εκτιμάτε ότι θα εμφανίζεται η 5η πιο συχνά εμφανιζόμενη λέξη;

Άσκηση 3 (25%)
Θεωρείστε τα έγγραφα d1…d10 τα οποία παριστάνονται στο δισδιάστατο χώρο ως εξής:



(α) Δώστε το αποτέλεσμα της ιεραρχικής ομαδοποίησης κατά single link.
(β) Δώστε το αποτέλεσμα της ιεραρχικής ομαδοποίησης κατά complete link.

Άσκηση 4 (25%)
Έστω ότι έχουμε ένα ανεστραμμένο ευρετήριο στο οποίο οι posting lists αποτελούνται μόνο από αναγνωριστικά εγγράφων. Κάθε αναγνωριστικό εγγράφου καταλαμβάνει 4 bytes. Για να εξοικονομήσουμε χώρο μια επιλογή είναι να χρησιμοποιήσουμε μια διαφορετική προσέγγιση για τις συχνές λέξεις, ειδικότερα να αποθηκεύσουμε τις posting lists τους στη μορφή ενός bitmap. Συγκεκριμένα κάθε λέξη w θα σχετίζεται με μια ψηφιακή λέξη από N bits, όπου N το πλήθος των εγγράφων, και η τιμή του ψηφίου στη θέση j θα είναι 1 εάν η λέξη w εμφανίζεται στο dj, αλλιώς 0.
Για ποιες λέξεις μας συμφέρει να αποθηκεύσουμε τις εμφανίσεις τους με αυτόν τον τρόπο ώστε να εξοικονομήσουμε χώρο; Δώστε ακριβώς τη συνθήκη που πρέπει να ισχύει για να είμαστε σίγουροι ότι εξοικονομούμε χώρο.

Μια εναλλακτική προσέγγιση θα μπορούσε να είναι η εξής: αν μια λέξη είναι συχνή, αντί να κρατάμε τα αναγνωριστικά των εγγράφων στα οποία εμφανίζεται, να κρατάμε τα αναγνωριστικά των εγγράφων στα οποία δεν εμφανίζεται. Συγκρίνεται και αυτή την επιλογή με τις προηγούμενες δύο.

Συνοψίστε την παραπάνω ανάλυση συμπληρώνοντας τις συνθήκες:

```
If  (cond1)       then use ComplementPostingList for w
elseIf (cond2)    then use BitMapPostingList for w
else              use  OrdinaryPostingList for w
```

Άσκηση 5 (25%)

Θέλουμε να αναπτύξουμε ένα σύστημα ανάκτησης πληροφοριών από η-μηνύματα (e-mails). Γενικά ένα μήνυμα συγκροτείται από 6 λογικές μονάδες: From, To, Subject, Date, Content, Attachments όπου το Attachments μπορεί είναι ένα πεπερασμένο σύνολο αρχείων.

α/ Αποφασίζετε να θεωρήσετε κάθε μήνυμα ως ένα αδόμητο κείμενο και να κατασκευάσετε ένα σύστημα βασισμένο σε ένα ανεστραμμένο ευρετήριο. Θέλετε να μειώσετε το χώρο αποθήκευσης των εμφανίσεων των λέξεων υιοθετώντας μια κατάλληλη κωδικοποίηση αριθμών (π.χ. Elias-γ, Elias-δ). Τι πρέπει να κάνετε για να κερδίσετε πράγματι χώρο από αυτήν την κωδικοποίηση;

β/ (Παραλλαγή προηγούμενου ερωτήματος) Τι θα μπορούσατε να κάνετε για να κερδίσετε χώρο από αυτήν την κωδικοποίηση αν λαμβάνατε υπόψη και τη δομή των μηνυμάτων; Συγκεκριμένα μπορείτε να σκεφτείτε έναν απλό και γρήγορο (στην υλοποίηση) τρόπο;

γ/ Θέλετε να αξιοποιήσετε τις τεχνικές ανάλυσης συνδέσμων. Περιγράψτε ποιες από αυτές τις τεχνικές (και πως) θα χρησιμοποιούσατε ώστε να αναδείξετε:
(γ1) τους πιο «σημαντικούς» παραλήπτες και αποστολείς μηνυμάτων,
(γ2) τα πιο «σημαντικά» μηνύματα.