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

3η Σειρά ασκήσεων
(Ευρετηρίαση, Αναζήτηση σε Κείμενα και Άλλα Θέματα)
(βαθμοί 12: όποιος πάρει άριστα σε όλες θα λάβει μέγιστο bonus 10%)
Ανάθεση: 24 Απριλίου
Παράδοση: 7 Μαΐου

Άσκηση 1 (2.5 βαθμοί)

Θεωρείστε ένα έγγραφο με περιεχόμενο «αυτό είναι ένα κείμενο και ένα κείμενο κανονικά έχει αρκετές λέξεις». 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 (1.0 βαθμοί)

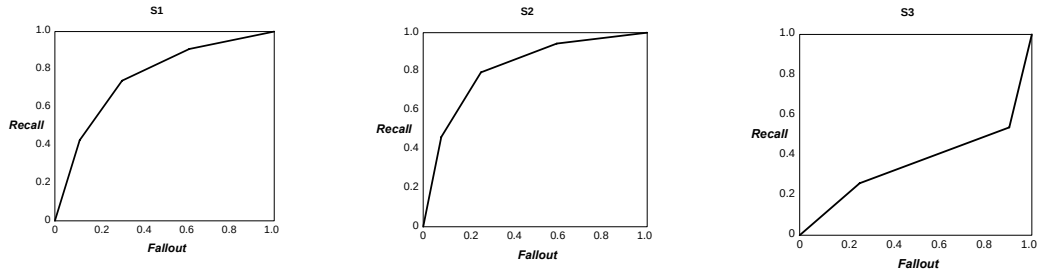
Θεωρείστε τις συμβολοσειρές π1= «to be or not to be» και π2 = «κανονικα». Δώστε τον πίνακα προεπεξεργασίας (δηλαδή τον πίνακα «next») που θα φτιάξει ο αλγόριθμος KMP (Knuth-Morris-Pratt) για κάθε μία από τις παραπάνω συμβολοσειρές.

Άσκηση 3 (1.0 βαθμοί)

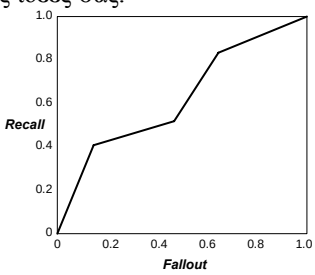
Θεωρείστε ότι θέλουμε να βρούμε αν η συμβολοσειρά «misspell» υπάρχει στη συμβολοσειρά «misspelling» με ανοχή λάθους (Edit Distance) κ=2. Σχεδιάστε τον πίνακα που απεικονίζει τον τρόπο με τον οποίο θα λειτουργήσει ο σχετικός αλγόριθμος δυναμικού προγραμματισμού.

Άσκηση 4 (6.0 βαθμοί)

Ένας εναλλακτικός τρόπος αξιολόγησης της αποτελεσματικότητας ενός ΣΑΠ, είναι οι καμπύλες Recall-Fallout. Ορίζονται ανάλογα με τις καμπύλες Precision-Recall, μόνο που τώρα ο άξονας Χ έχει τιμές του Fallout, ενώ ο Υ τιμές του Recall.
(α)[20β] Έστω ένα σύστημα Χ το οποίο για κάθε επερώτηση που λαμβάνει επιστρέφει μια τυχαία γραμμική διάταξη των κειμένων του. Τι μορφή θα είχε η καμπύλη Recall-Fallout αυτού του συστήματος; (αυτή που θα προέκυπτε από την αξιολόγηση πολλών επερωτήσεων)
(β)[20β] Θεωρείστε τρία συστήματα με τις καμπύλες Recall-Fallout που ακολουθούν. Παρατηρώντας αυτές τις καμπύλες, ποιο σύστημα θα κρίνατε ότι προσφέρει πιο αποτελεσματική ανάκτηση πληροφορίας;



(γ) [20β] Έστω ένα ΣΑΠ του οποίου η καμπύλη Recall-Fallout έχει τη μορφή που εικονίζεται παρακάτω. Θα μπορούσατε να κάνετε κάτι για να βελτιώσετε την αποτελεσματικότητα του συστήματος; Αναπτύξτε ελεύθερα τις ιδέες σας.



Άσκηση 5 (2.5 βαθμοί)

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: 2GBytes, σκληρός δίσκος: 150GB). 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.

Καλή επιτυχία