

4^η Σειρά Ασκήσεων

Ευρετήρια. Αλγόριθμοι συμβολοσειρών
Ημερομηνία Παράδοσης: Δευτέρα 11 Μαΐου 2009
Ομάδες των 2 ατόμων

Άσκηση 1

Υπολογίστε την απόσταση (edit distance) μεταξύ των λέξεων «άνοιξη» και «κατάλυξη» χρησιμοποιώντας τον πίνακα που δίνει ο δυναμικός προγραμματισμός.

Άσκηση 2

Θεωρείστε τη συμβολοσειρά «to be or not to be».

(α) Δώστε τον πίνακα next του αλγορίθμου KNP (Knuth-Morris-Pratt).

(β) Δώστε το κατάλληλο ντετερμινιστικό αυτόματο.

Άσκηση 3

Θεωρείστε μια συλλογή εγγράφων με απλό κείμενο.

(α) Έστω ότι ο συνολικός αριθμός λέξεων σε όλα τα έγγραφα είναι 1.000.000. Ποιο είναι το εκτιμώμενο μέγεθος του λεξιλογίου (δηλαδή, ο αριθμός διαφορετικών λέξεων);

(β) Αν η πιο συχνή λέξη εμφανίζεται 100.000 φορές, πόσες φορές εκτιμάτε ότι θα εμφανίζεται η δέκατη πιο συχνή;

Άσκηση 4 (ατομική)

Θέλετε να υποθηκεύσετε ένα ανεστραμμένο ευρετήριο σε k διαφορετικούς κόμβους ενός κατανεμημένου συστήματος.

Περιγράψτε δυο διαφορετικούς τρόπους κατανομής των εγγράφων του ευρετηρίου.

Δώστε τα πλεονεκτήματα και τα μειονεκτήματα κάθε τρόπου.