# Εργαστήριο Δομών Δεδομένων - 4η Άσκηση
## Ημερομηνία Παράδοσης : 11/01/2017 (πριν την έναρξη του εργαστηρίου - 11.00 -)

**Εργαστηριακός Διδάσκων Μαθήματος:** Δούμα Αναστασία (sia@aegean.gr)

Το ευρετήριο ενός αρχείου κειμένου είναι μία λίστα με λέξεις κλειδιά που εμφανίζονται σε ένα κείμενο μαζί με τους αριθμούς σελίδων στις οποίες εμφανίζονται. Στη συγκεκριμένη εργασία ζητείται να υλοποιήσετε κάτι ανάλογο, ένα ευρετήριο για ένα δεδομένο αρχείο κειμένου. Ζητείται δηλαδή να δημιουργήσετε έναν κατάλογο με όλες τις λέξεις που εμφανίζονται σε ένα κείμενο. Για κάθε λέξη θα καταχωρείται και μία γραμμική λίστα με όλους τους αριθμούς γραμμών στις οποίες εμπεριέχεται η λέξη. Το ευρετήριο θα πρέπει να υλοποιηθεί χρησιμοποιώντας δυαδικό δέντρο αναζήτησης ως δομή αποθήκευσης.

Πιο συγκεκριμένα, το πρόγραμμα σας αρχικά θα ζητάει το όνομα του αρχείου για το οποίο πρέπει να δημιουργηθεί το ευρετήριο και στη συνέχεια θα το ανοίγει με σκοπό την ανάγνωση του αρχείου. Στη συνέχεια θα καλείται επαναληπτικά μία συνάρτηση η οποία θα διαβάζει κάθε φορά την τρέχουσα γραμμή του αρχείου (οι γραμμές του αρχείου διαχωρίζονται με <enter>) και θα βρίσκει τις λέξεις που εμπεριέχονται. Το κείμενο που θα εμπεριέχεται στο αρχείο θα πρέπει να είναι αποκλειστικά αγγλικό και ο διαχωρισμός των λέξεων θα γίνεται με βάση τα σημεία στίξης και το κενό. Έτσι η πρόταση «*We decided to visit: Spain. More specifically the Madrid and Toledo.*» αποτελείται από 11 διαφορετικές λέξεις.

Για κάθε λέξη που εμφανίζεται στο κείμενο θα γίνεται η διαδικασία της εισαγωγής της λέξης στο δυαδικό δέντρο αναζήτησης καλώντας αντίστοιχη συνάρτηση. Πριν την εισαγωγή θα πρέπει να γίνεται μετατροπή των γραμμάτων της λέξης σε κεφαλαία. Σε περίπτωση που η λέξη δεν υπάρχει θα καταχωρείται ο νέος κόμβος που θα περιέχει την λέξη και μία γραμμική λίστα που θα περιέχει τον αριθμό γραμμής που εμφανίζεται η λέξη. Σε περίπτωση που η λέξη υπάρχει στο δυαδικό δέντρο αναζήτησης απλά θα ενημερώνεται η λίστα με την νέα γραμμή στην οποία εμφανίζεται η λέξη. Εάν η λέξη εμφανίζεται περισσότερο από μία φορά στην ίδια γραμμή, μόνο μία φορά θα καταχωρείται ο αριθμός της συγκεκριμένης γραμμής.

Ολοκληρώνοντας την εισαγωγή όλων των λέξεων του κειμένου στο ευρετήριο θα πρέπει να δίνεται η δυνατότητα στο χρήστη να εμφανίζει ολόκληρο το ευρετήριο εκτελώντας ενδοδιατεταγμένη (inorder) διάσχιση του δέντρου.

Θα πρέπει να εμφανίζονται οι λέξεις, οι αριθμοί γραμμών στις οποίες εμπεριέχονται και στατιστικά όπως το πλήθος όλων των λέξεων που διαβάστηκαν από το κείμενο, το πλήθος των διακριτών λέξεων που καταχωρήθηκαν στο ευρετήριο, και ο χρόνος που χρειάστηκε για την κατασκευή του δυαδικού δέντρου αναζήτησης.

Έτσι για παράδειγμα το ευρετήριο θα πρέπει να εμφανίζεται ως εξής:

***Ευρετήριο Λέξεων:***
***----------***
***SPAIN : 11, 20***
***TOLEDO : 12, 20, 60***
***TRIP : 15***
***…***
***Συνολικός αριθμός λέξεων κειμένου : 1835***
***Αριθμός λέξεων ευρετηρίου : 1220***

***Χρόνος που χρειάστηκε για την δημιουργία του δυαδικού δέντρου αναζήτησης είναι : 0.15 sec.***

Ολοκληρώνοντας τη διαδικασία δημιουργίας του ευρετηρίου από το αρχείο κειμένου ο χρήστης θα πρέπει να έχει τη δυνατότητα να εισάγει μία νέα λέξη καταχωρώντας ο ίδιος την λέξη και την σελίδα στην οποία εμπεριέχεται ή να ζητά εξ ολοκλήρου την διαγραφή μιας λέξης από το ευρετήριο (ολόκληρο τον κόμβο του δέντρου που περιέχει την λέξη και την γραμμική λίστα των αριθμών γραμμών).

Με την ολοκλήρωση του προγράμματος θα πρέπει να γίνεται καταχώριση ολόκληρου του ευρετηρίου (με την ίδια μορφή που εμφανίσθηκε στην οθόνη) σε αρχείο εξόδου με όνομα index_bst.txt.

Όλες οι βασικές λειτουργίες της εφαρμογής θα πρέπει να υλοποιηθούν με χρήση συναρτήσεων. Σε κάθε κόμβο του δέντρου θα πρέπει να αποθηκεύετε την πληροφορία που απαιτείται με βάση τις απαιτήσεις της εφαρμογής αλλά και 2 δείκτες **left** και **right** που θα δείχνουν στον αριστερό κόμβο και στο δεξί κόμβο του τρέχοντος κόμβου.

**Παραδοτέα:**
α) Τον κώδικα που θα έχετε υλοποιήσει μέσα στο εργαστηρίου θα πρέπει να τον παραδώσετε με την ολοκλήρωση του εργαστηρίου (ισχύει μόνο για όσους παρακολουθούν το εργαστήριο). Υπάρχει σχετικός σύνδεσμος στο eclass. Το αρχείο που θα ανεβάσετε θα πρέπει να έχει όνομα ***Exercise4_lab.cpp (ή .c - θα πρέπει να το κάνετε .zip για να το ανεβάσετε). Στην 1η γραμμή του κώδικα σας θα πρέπει υποχρεωτικά να αναγράφεται το όνομα και ο αριθμός μητρώου των μελών της ομάδας που συμμετείχε στην υλοποίηση που έγινε στο εργαστήριο.***

(β) Τελική παράδοση εργασίας :
Για την τελική παράδοση της εργασίας θα πρέπει να δημιουργήσετε **ένα αρχείο** με τον πηγαίο κώδικα με όνομα ***Exercise41.cpp (ή .c).***

***Τέλος θα δημιουργήσετε ένα zip αρχείο με όνομα αρχείου «το email σας_όνομα άσκησης» (πχ. icsd000666_Exercise4.zip) (μόνο ένα μέλος της ομάδας θα πρέπει να καταγράψει το email του). Το zip αρχείο θα πρέπει να περιέχει το παραπάνω αρχείο. ΠΡΟΣΟΧΗ: ΚΑΜΙΑ ΕΡΓΑΣΙΑ ΔΕΝ ΘΑ ΔΙΟΡΘΩΘΕΙ ΑΝ ΔΕΝ ΑΠΟΣΤΑΛΕΙ ΜΕ ΤΟΝ ΠΑΡΑΠΑΝΩ ΤΡΟΠΟ!***

Η παράδοση του αρχείου θα γίνει με τη χρήση της πλατφόρμας ηλεκτρονικής μάθησης του τμήματος (http://www.icsd.aegean.gr/eclass).

**Σε όλα τα αρχεία που θα δημιουργήσετε (πηγαίος κώδικας) θα πρέπει να αναφέρετε στην αρχή του κειμένου τα ονόματα των μελών της ομάδας εργασίας.**

**Υποδείξεις.** Η εργασία είναι ***2 ατόμων.*** Τα προγράμματα πρέπει να υλοποιηθούν σε γλώσσα C ή C++.