ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΟΜΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΡΧΕΙΩΝ

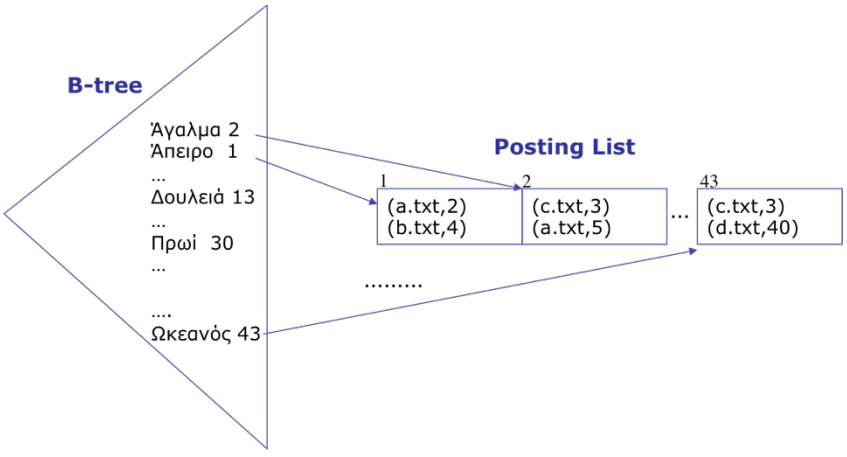
2η άσκηση

Ημερομηνία παράδοσης: 28 Απριλίου 2017

Αναζήτηση σε Αρχεία με χρήση Δεικτοδότησης (indexing)

Κατασκευάστε ένα πρόγραμμα που δέχεται ένα ή περισσότερα αρχεία κειμένου (όπως αυτά που σας δίνονται ενδεικτικά) και δημιουργεί μια δομή δεδομένων στο δίσκο που απαντά σε ερωτήσεις όπως «**βρες κείμενα που περιέχουν την λέξη 'άγαλμα'**» ή οποιαδήποτε άλλη λέξη. Η Αναζήτηση γίνεται με την βοήθεια «δεικτοδότησης» (indexing) δηλαδή όχι απευθείας πάνω στα κείμενα αλλά μέσω μιας δομής αρχείου που σκοπό έχει να επιταχύνει την αναζήτηση στην περίπτωση που τα αρχεία είναι πάρα πολλά. Η δομή αρχείου αποτελείται από το «**Λεξικό**» και το «**Ευρετήριο**». Η δομή αρχείου στο σχήμα δηλώνει ότι η λέξη «Άπειρο» του Λεξικού υπάρχει στην 1η σελίδα του Ευρετηρίου. Εκεί βρίσκεται η πληροφορία ότι ή λέξη υπάρχει στο αρχείο a.txt στην θέση 2bytes από την αρχή του αρχείου και στο αρχείο b.txt στην θέση 4bytes από την αρχή του αρχείου.

Το Λεξικό θα έχει την δομή αρχείου B-tree και το Ευρετήριο θα έχει την δομή αρχείου Posting List. Δηλαδή, θα δημιουργηθούν δύο δομές αρχείου και αντίστοιχα δύο αρχεία στον δίσκο ένα για την κάθε δομή αρχείου.



A. B-Tree στον Δίσκο (4 μονάδες)

Υλοποιείστε στον δίσκο ένα B-tree τάξης με μέγεθος σελίδας N = 128 bytes. Η υλοποίηση πρέπει να λειτουργεί για οποιοδήποτε N. Κάθε σελίδα του B-tree αποθηκεύει

εγγραφές που περιέχουν ένα κλειδί (string μεγέθους το πολύ 12 χαρακτήρων) και ένα πεδίο info (ακέραιος 4 bytes) που δείχνει σε ποια σελίδα του ευρετηρίου υπάρχει πληροφορία για την λέξη (θα εξηγηθεί παρακάτω), δείκτες στις σελίδες-παιδιά του κόμβου, ένα δείκτη στον κόμβο πατέρα και τον αριθμό των εγγραφών που αποθηκεύεται στην σελίδα. Τα κλειδιά αποθηκεύονται σε αύξουσα σειρά. Υπολογίστε τον βαθμό του δένδρου n.

Το B-tree θα χρησιμοποιηθεί για να δεικτοδοτήσει τις λέξεις που υπάρχουν σε αρχεία κειμένου. Κάθε λέξη αρχείου είναι ένα κλειδί. Κάθε κλειδί (λέξη) μπορεί να υπάρχει σε ένα ή περισσότερα αρχεία και σε μία ή περισσότερες θέσεις ενός αρχείου. Αυτή η πληροφορία αποθηκεύεται στο Ευρετήριο. Στην πράξη το πεδίο info περιέχει το αριθμό της σελίδας στο Ευρετήριο στην οποία θα βρείτε πληροφορία που δηλώνει σε ποια αρχεία κειμένου και σε ποιες θέσεις (αριθμός χαρακτήρων από την αρχή) υπάρχει η λέξη.

Υλοποιείστε τις πράξεις εισαγωγής και αναζήτησης κλειδιών. Μέρος της υλοποίησης εισαγωγής είναι η συνάρτηση “split" που διασπά ένα κόμβο σε δύο (ταυτόχρονα με την δημιουργία ενός νέου κόμβου). Η υλοποίηση πρέπει να λειτουργεί και για μεγάλο αριθμό κλειδιών (λέξεων στην περίπτωση της άσκησης).

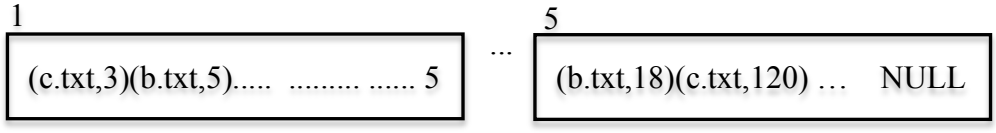
Αν προτιμάτε μπορείτε να βρείτε μια έτοιμη υλοποίηση από το διαδίκτυο και να την προσαρμόσετε έτσι ώστε να λειτουργεί για τον δίσκο (υπάρχουν πολλές όπως http://sourcecode4all.com/b-tree/).

Β. Ευρετήριο στον Δίσκο (4 μονάδες με προϋπόθεση να έχει γίνει το Α)

Το Ευρετήριο είναι ένα αρχείο του οποίου κάθε σελίδα αποθηκεύει ζεύγη της μορφής (**όνομα αρχείου – θέσεις bytes από την αρχή του κειμένου)** π.χ. η λέξη «Άπειρο» στο σχήμα, συνοδεύεται στο Λεξικό από τον ακέραιο 2 που σημαίνει ότι δείχνει στην $2^{η}$ σελίδα του Ευρετηρίου. Η σελίδα 2 στο Ευρετήριο έχει τις εγγραφές (b.txt,4) (a.txt,2) που δηλώνουν ότι η λέξη «Άπειρο» υπάρχει στο αρχείο “a.txt” στην θέση 2bytes από την αρχή του και στο αρχείο “b.txt” στη θέση 4bytes από την αρχή του. Αν μία λέξη υπάρχει στο Λεξικό τότε υπάρχει γι αυτήν τουλάχιστον μία σελίδα στο Ευρετήριο.

Για κάθε δήλωση π.χ. (c.txt,3) στο Ευρετήριο χρειάζονται 12bytes δηλαδή 8 για το όνομα του αρχείου και 4bytes για την θέση της λέξης στο αρχείο. Το Ευρετήριο αποτελείται από σελίδες μεγέθους 128bytes. Άρα κάθε σελίδα χωράει 10 εγγραφές και περισσεύουν 8bytes (4 από αυτά ίσως χρειαστούν για να συνδέσετε μεταξύ τους δύο σελίδες όπως περιγράφεται αμέσως παρακάτω).

Αν μία λέξη επαναλαμβάνεται συχνά στα κείμενα τότε μπορεί να έχει πολλές εγγραφές στο Ευρετήριο που δεν χωράνε όλες σε μία σελίδα. Μόλις γεμίσει μια σελίδα στο Ευρετήριο, δημιουργείται μία νέα στο τέλος του αρχείου και συνδέεται με την πρώτη όπως στο σχήμα: η σελίδα 1 συνδέεται με την σελίδα 5 (μπορεί όμως να τύχει να είναι συνεχόμενες σελίδες στο αρχείο δηλαδή η 1 να δείχνει στην 2 κλπ).



Β. Ανάλυση Απόδοσης (2 μονάδες, εφόσον έχουν υλοποιηθεί σωστά τα Α και Β)

Γράψτε ένα πρόγραμμα που στην είσοδό του δέχεται μία λέξη και ως απάντηση επιστρέφει όλα τα κείμενα στα οποία υπάρχει η λέξη και την θέση στην οποία βρίσκεται. Το αποτέλεσμα που εμφανίζεται στην οθόνη έχει την μορφή

«κείμενο 'c.txt' περιέχει την λέξη 'άγαλμα' στην θέση 3»
«κείμενο 'a.txt' περιέχει την λέξη 'άγαλμα' στην θέση 5»
«κείμενο 'b.txt' περιέχει την λέξη 'άγαλμα' στην θέση 18»
«κείμενο 'c.txt' περιέχει την λέξη 'άγαλμα' στην θέση 120»

Κάθε διάβασμα σελίδας στο B-tree κοστίζει μία πρόσβαση στο δίσκο. Κάθε διάβασμα σελίδας στο Ευρετήριο κοστίζει επίσης μία πρόσβαση στο δίσκο. Η αναζήτηση δέχεται ένα κλειδί και επιστρέφει το αριθμό προσβάσεων στον B-tree αλλά και τον αριθμό των προσβάσεων στο αρχείο της posting list.

Εισάγετε στο B-tree τις λέξεις που θα βρείτε σε τρία αρχεία που δίνονται. Υπολογίστε τον μέσο αριθμό προσβάσεων σε σελίδες δίσκου για 100 αναζητήσεις λέξεων από αυτές που υπάρχουν στο B-tree και σε 100 τυχαίες λέξεις από τις οποίες κάποιες μπορεί να μην υπάρχουν στο B-tree.

Συγκεντρώστε τα αποτελέσματα στον παρακάτω και προσπαθήστε να δικαιολογήστε την απόδοση κάθε μεθόδου.

	Α. Εισαγωγή	**Β. Επιτυχής Αναζήτηση**		**Γ. Τυχαία αναζήτηση**	
Μέθοδος	**Μέσος αριθμός προσβάσεων στο δίσκου κατά την δημιουργία του B-tree**	**Μέσος αριθμός προσβάσεων δίσκου στο B-tree για 100 αναζητήσεις λέξεων που υπάρχουν**	**Μέσος αριθμός προσβάσεων σε σελίδες δίσκου στο Ευρετήριο**	**Μέσος αριθμός προσβάσεων δίσκου στο B-tree για 100 αναζητήσεις τυχαίων λέξεων**	**Μέσος αριθμός προσβάσεων σε σελίδες δίσκου στο Ευρετήριο**
Μέτρηση					

<u>Παραδοτέα:</u> Ένα συμπιεσμένο zip αρχείο που περιέχει ότι ζητείται παρακάτω:

- Ο κώδικας περιέχει συνοπτικά σχόλια που εξηγούν την υλοποίηση.
- Μία έκθεση που περιγράφει σε 1-2 σελίδες πως φτιάχτηκε ο κώδικας (δηλαδή για κάθε ερώτημα ποια είναι η γενική ιδέα της λύσης σε 3-4 προτάσεις), υπάρχουν σαφείς οδηγίες μετάφρασης από compiler και εκτέλεσης, τι λάθη έχει (αν έχει, περιπτώσεις που δεν δουλεύει το πρόγραμμα, ή περιπτώσεις που κάνει

περισσότερα από όσα σας ζητεί η άσκηση, τι χρησιμοποιήσατε από έτοιμα προγράμματα ή πηγές πληροφόρησης. Υποδείξετε ακόμα και πηγές στο WWW όπως Wikipedia ή ακόμα και συναδέλφους που σας βοήθησαν στην άσκηση.

- Για το ερώτημα Ε πρέπει να υπάρχει τεκμηρίωση των αποτελεσμάτων με σαφήνεια.
- Εκτός των παραπάνω, οι ασκήσεις βαθμολογούνται με άριστα εφόσον:
    - Το zip είναι πλήρες
    - Οι κώδικες περνούν από compiler και εκτελούνται κανονικά και σωστά σε windows ή Linux περιβάλλον