

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάριος Δ. Δικαιάκος, Αναπληρωτής Καθηγητής

ΕΠΛ660: Ανάκτηση Πληροφοριών και Μηχανές Αναζήτησης

<http://www.cs.ucy.ac.cy/courses/epl660>

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΕΡΓΑΣΙΑ 1η

Ημερομηνία ανακοίνωσης: 26/01/2010

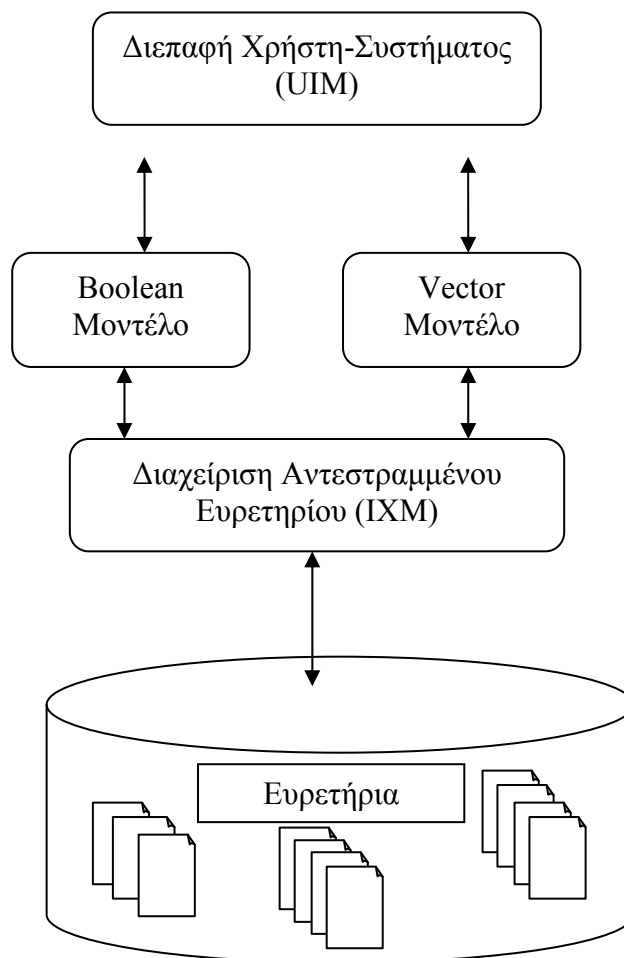
Ημερομηνία Παράδοσης: **9/02/2010**

1. Εισαγωγή

Στο πλαίσιο μιας σειράς προγραμματιστικών εργασιών θα αναπτύξετε ένα σύστημα ανάκτησης πληροφοριών με την ονομασία **Pythia**, το οποίο θα υποστηρίζει βασικές τεχνικές ανάκτησης. Το σύστημα θα έχει τη δυνατότητα να επεξεργάζεται ερωτήματα λογικών εκφράσεων και θα υποστηρίζει και το διανυσματικό μοντέλο ανάκτησης. Η διαχείριση των αρχείων κειμένου πραγματοποιείται με τη βοήθεια αντεστραμμένου ευρετηρίου. Θα χρησιμοποιήσετε το σύστημα για να πραγματοποιείτε επερωτήσεις σε διαφορετικές συλλογές αρχείων στα Αγγλικά.

2. Αρχιτεκτονική

Η **αρχιτεκτονική** του τελικού συστήματος δίνεται στο σχήμα που ακολουθεί. Όλα τα δεδομένα είναι αποθηκευμένα στη δευτερεύουσα μνήμη. Τα έγγραφα είναι οργανωμένα σε διαφορετικές συλλογές εγγράφων (document collections). Κάθε συλλογή εγγράφων αποθηκεύεται σε διαφορετικό κατάλογο στο δίσκο. Επίσης, για κάθε συλλογή υπάρχει και ένας διαφορετικός αντεστραμμένος κατάλογος (inverted index) ο οποίος χρησιμοποιείται για την οργάνωση της συλλογής. Το τμήμα του συστήματος που είναι υπεύθυνο για τη διαχείριση των αντεστραμμένων καταλόγων είναι το τμήμα IXM (indexing module). Πάνω σε αυτό στηρίζονται τα τμήματα που επεξεργάζονται ερωτήματα Boolean και ερωτήματα με βάση το διανυσματικό μοντέλο ανάκτησης (τμήματα BM και VM αντίστοιχα). Στο τελευταίο επίπεδο διακρίνουμε το τμήμα UIM, διεπαφή χρήστη-συστήματος (user interface module), που είναι υπεύθυνο για την υποστήριξη του χρήστη. Στην ουσία, η προσπέλαση στο σύστημα ανάκτησης και η επιστροφή των αποτελεσμάτων γίνεται από το τμήμα UIM.



3. Στόχος

Στόχος της παρούσας εργασίας είναι η εξοικείωσή σας με τα αντεστραμμένα ευρετήρια, τις δομές δεδομένων που χρησιμοποιούνται για την υλοποίησή τους και τη χρήση των δενδρικών δομών αναζήτησης.

4. Ζητούμενο

Καλείστε να δημιουργήσετε το λογισμικό σύστημα διαχείρισης συλλογών εγγράφων, που περιλαμβάνει τα ακόλουθα υποσυστήματα:

- a. **Υποσύστημα δημιουργίας και διαχείρισης συλλογής.** Προσφέρει τις ακόλουθες λειτουργίες:
 - i. Δημιουργία νέας συλλογής
 - ii. Εισαγωγή αρχείου στη συλλογή
 - iii. Διαγραφή αρχείου από τη συλλογή
 - iv. Δυνατότητα μαζικής εισαγωγής εγγράφων που βρίσκονται σε έναν κατάλογο
 - v. Άνοιγμα συλλογής
 - vi. Κλείσιμο συλλογής
 - vii. Εκτύπωση περιεχομένων συλλογής
 - viii. Εκτύπωση λεξικού της συλλογής
 - ix. Ανάκτηση του ευρετηρίου για μια συλλογή

b. Υποσύστημα δημιουργίας και διαχείρισης αντεστραμμένου ευρετηρίου.

Προσφέρει τις ακόλουθες λειτουργίες:

- i. Δημιουργία αντεστραμμένου ευρετηρίου μιας συλλογής, το οποίο να περιλαμβάνει πληροφορίες τοποθεσίας και συχνότητας
- ii. Δημιουργία λεξικού μιας συλλογής
- iii. Ενημέρωση λεξικού και περιεχομένων ευρετηρίου όταν προκύπτουν αλλαγές στη σύνθεση της συλλογής

Κατά τη δημιουργία του λεξικού και τους αντεστραμμένου ευρετηρίου, το υποσύστημά σας θα πρέπει να πραγματοποιεί στελέχωση κειμένου (stemming) και απαλοιφή των λέξεων αποκλεισμού (stop words).

c. Υποσύστημα αναζήτησης όρων στο αντεστραμμένο ευρετήριο.

Υποστηρίζει τη δυνατότητα εισαγωγής και αναζήτησης ενός ή περισσοτέρων όρων στο λεξικό. Σε κάθε αναζήτηση ενός όρου, το πρόγραμμα επιστρέφει, εκτός από τα αρχεία στα οποία περιλαμβάνεται ο αναζητούμενος όρος, και το χρόνο που χρειάστηκε η πραγματοποίηση της αναζήτησης.

5. Παραδοτέα

- Θα πρέπει να παραδώσετε:
 - ο πηγαίο κώδικα του συστήματός σας με σχόλια
 - ο εκτελέσιμα αρχεία
 - ο τεκμηρίωση όπου να φαίνεται ο τρόπος σκέψης σας για την ανάπτυξη του συστήματος, αλγόριθμοι και τεχνικές που χρησιμοποιήθηκαν κλπ (8-10 σελίδες Times New Roman, 11pt, all margins 1.5cm)
- Στην τεκμηρίωση θα πρέπει να υπάρχει ξεχωριστό τμήμα που να περιγράφει τις επεκτάσεις και τον τρόπο που αυτές υλοποιήθηκαν. Τα παραπάνω έγγραφα θα πρέπει να περιέχονται σε ένα αρχείο zip (ή rar) το οποίο θα αποσταλεί με e-mail στον βοηθό του μαθήματος μέχρι και την ημερομηνία παράδοσης της εργασίας.
- **Τα προγράμματα θα ελεγχθούν για αντιγραφές. Αντιγραμμένες εργασίες θα μηδενίζονται.**

6. Οδηγίες

1. Στο αντεστραμμένο ευρετήριο αποθηκεύονται εγγραφές με την ακόλουθη δομή:

term	doc. Frequency	Positional posting list
------	----------------	-------------------------

όπου το positional posting list περιέχει τις ταυτότητες των αρχείων που περιλαμβάνουν τον κάθε όρο και τα σημεία στα οποία εντοπίζεται ο κάθε όρος (term).

2. Για την αναζήτηση των λέξεων, να χρησιμοποιηθεί η δένδρική δομή αναζήτησης B-δένδρο. Θεωρείστε ότι κάθε εσωτερικός κόμβος στο B-δένδρο έχει έναν αριθμό από παιδιά στο διάστημα [2,4].
3. Για τη στελέχωση του Αγγλικού κειμένου (stemming) χρησιμοποιείστε τον αλγόριθμο Porter Stemmer. Τον κώδικα για τον αλγόριθμο Porter Stemmer μπορείτε να βρείτε στην ιστοσελίδα: <http://tartarus.org/martin/PorterStemmer/>.

4. Μία λίστα με λέξεις αποκλεισμού μπορείτε να την κατεβάσετε από τη διεύθυνση:
<http://www.cs.ucy.ac.cy/courses/EPL660/exercises/stoplist.txt>
5. Τα αρχεία τα οποία θα εντάξετε στη συλλογή του Pythia δίνονται στα ακόλουθα URL:
 - a. Δοκιμαστική συλλογή:
<http://www.cs.ucy.ac.cy/courses/EPL660/exercises/testdata.zip>
 - b. Συλλογή δημοφιλών έργων:
<http://www.cs.ucy.ac.cy/courses/EPL660/english-documents.zip>
6. Μπορείτε να υλοποιήσετε το πρόγραμμά σας με χρήση JAVA.