ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

HY463 – Συστήματα Ανάκτησης Πληροφοριών

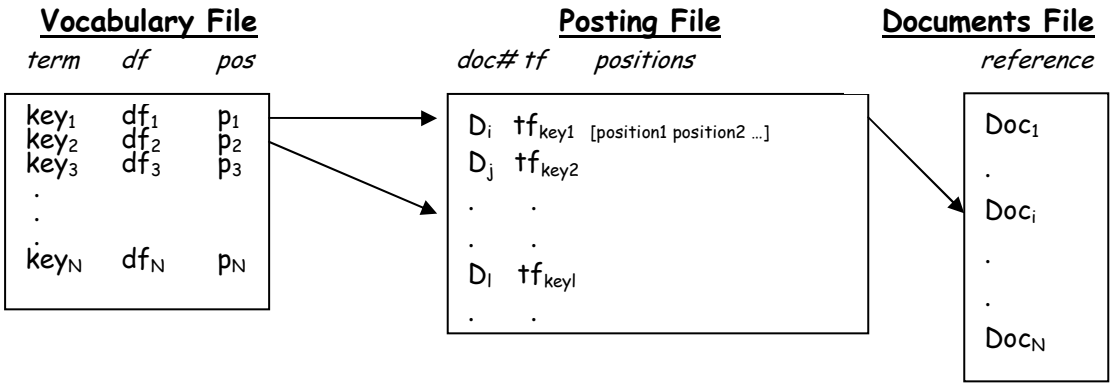
Εργασία: Ανεστραμμένο Ευρετήριο

Εισαγωγή

Σκοπός της εργασίας είναι η δημιουργία ενός ανεστραμμένου ευρετηρίου για τη μηχανή αναζήτησης **Μίτος**, το οποίο θα είναι εντελώς ανεξάρτητο από το σημερινό ευρετήριο που βασίζεται σε Σχεσιακό Σύστημα Βάσεων Δεδομένων (postgreSQL). Το ευρετήριο πρέπει να περιέχει τουλάχιστον τις πληροφορίες που περιέχονται στο DBMS-based ευρετήριο του **Μίτου**. Η δημιουργία ενός ανεστραμμένου ευρετηρίου (και η χρήση συνεπτυγμένων κωδικών) αναμένεται να μειώσει κατά πολύ το μέγεθος του ευρετηρίου, σε σχέση με το υπάρχον ευρετήριο. Συνάμα καλείστε να κάνετε τις απαραίτητες αλλαγές ώστε ο Μίτος να μπορεί να λειτουργήσει και με το νέο ευρετήριο. Αυτό περιλαμβάνει την επέκταση της διεπαφής (interface) *Index* και διάφορες άλλες συμπληρωματικές εργασίες.

Α.1. Δημιουργία DBMS-Free Ευρετηρίου.

Για τα ανεστραμμένα ευρετήρια έχουμε μιλήσει αναλυτικά στο μάθημα και σχετικό υλικό υπάρχει στις διαφάνειες και στα βιβλία του μαθήματος. Ένα παράδειγμα ενός ανεστραμμένου αρχείου είναι το παρακάτω:



Vocabulary File: Το λεξιλόγιο της συλλογής. Ιδανικά πρέπει να βρίσκεται στην μνήμη (εφόσον χωράει). Δεδομένου οτι η αναζήτηση γίνεται βάσει των όρων, χρησιμοποιήστε μια κατάλληλη δομή. Για κάθε όρο κρατάμε το πλήθος των εγγράφων στα οποία εμφανίζεται (df) καθώς επίσης και ένα δείκτη προς το posting file στο σημείο που καταγράφονται οι πληροφορίες σχετικές με τις εμφανίσεις του όρου (αναγνωριστικά εγγράφων, θέσεις, κλπ).

Posting File: Περιέχει τα αναγνωριστικά των εγγράφων στα οποία εμφανίζονται οι όροι του λεξιλογίου. Περιέχει επίσης το tf για κάθε όρο στο αντίστοιχο έγγραφο. Μπορεί επίσης να περιέχει και άλλες πληροφορίες όπως οι θέσεις εμφάνισης του κάθε όρου στο έγγραφο ή εναλλακτικά το block (εφόσον χρησιμοποιείται block addressing). Το Posting File είναι αρκετά μεγάλο και επομένως δεν μπορεί να βρίσκεται στην μνήμη.

(Documents File: Περιέχει αναφορές προς τα έγγραφα. Συγκεκριμένα κάθε εγγραφή που υπάρχει στο *Posting file* θα "δείχνει" στο σημείο του *Documents File* στο οποίο υπάρχει το

αναγνωριστικό του εγγράφου στο οποίο αναφέρεται η συγκεκριμένη εγγραφή. Τα περιεχόμενα του *Documents File* θα είναι ο τίτλος των εγγράφων (π.χ. documentMD5.txt) και/είτε οτιδήποτε άλλο κρίνεται σκόπιμο/χρήσιμο.

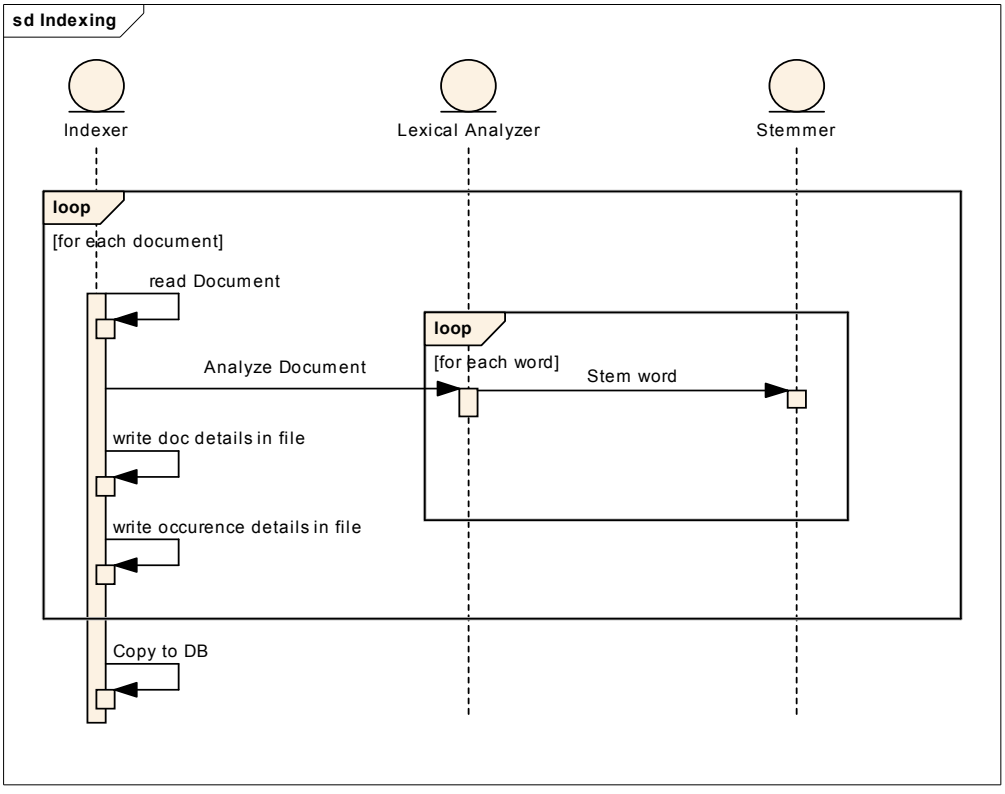
Αρχικά καλείστε να εξάγετε τα απαραίτητα δεδομένα από την βάση δεδομένων και να τα κρατήσετε στις DBMS-Free δομές που έχετε δημιουργήσει. Θα πρέπει η πληροφορία που κρατάτε στο ευρετήριο να είναι τέτοια ώστε :

- Να είναι η ελάχιστη δυνατή ώστε να μην υπάρχει πλεονάζουσα (και επομένως άχρηστη πληροφορία στο ευρετήριο) και συνεπώς το ευρετήριο να έχει το ελάχιστο δυνατό μέγεθος.
- Να περιέχει όλα τα απαραίτητα στοιχεία που χρειάζονται για να γίνει ανάκτηση της πληροφορίας.

Μπορείτε να ενημερωθείτε για το υπάρχον (DBMS-based) ευρετήριο από το ακόλουθο link http://groogle.csd.uoc.gr/apache2-default/index.php/Indexer_DBMS.

Α.2. Ευρετηρίαση εγγράφων με το νέο ευρετήριο

Ευρετηριάστε μία συλλογή απευθείας στις νέες δομές που δημιουργήσατε χωρίς να παρεμβάλλεται καθόλου η βάση δεδομένων πλέον. Παρακάτω μπορείτε να δείτε την διαδικασία που ακολουθείται για να γίνει ευρετηρίαση μιας συλλογής. Συγκεκριμένα ο ***ευρετηριαστής*** διαβάζει κάθε έγγραφο από την συλλογή την οποία έχει προηγουμένως κατεβάσει ο ***ερπυστής.*** Για κάθε έγγραφο γίνεται η ανάλυση του από τον ***λεξικογραφικό αναλυτή*** και τον ***στελεχωτή*** και κατόπιν αποθηκεύονται οι πληροφορίες για το εκάστοτε έγγραφο καθώς και για τις εμφανίσεις των λέξεων σε αυτό σε αρχεία. Το λεξιλόγιο (βλ. **Vocabulary File** παραπάνω) κρατείται στην μνήμη. Μόλις ολοκληρωθεί η παραπάνω διαδικασία γίνεται αντιγραφή των δεδομένων που βρίσκονται στην μνήμη (λεξιλόγιο), και των δεδομένων που βρίσκονται σε αρχεία (έγγραφα, εμφανίσεις όρων σε κάθε έγγραφο) στο δίσκο.



Κατά τη διάρκεια δημιουργίας του ανεστραμμένου ευρετηρίου η τρέχουσα posting list αποθηκεύεται στη μνήμη. Για κάθε νέα εγγραφή σε κάποιο posting list δεσμεύεται δυναμικά μνήμη. Όταν γίνεται δέσμευση μνήμης πρέπει να γίνεται έλεγχος για το εάν έχει εξαντληθεί η μνήμη. Εάν έχει εξαντληθεί, τότε η διαδικασία σταματά και το μερικό ευρετήριο (partial index) αποθηκεύεται στο δίσκο. Συνεχίζετε με αυτόν τον τρόπο έως ότου να ολοκληρώσετε την ευρετηρίαση όλων των εγγράφων της αποθήκης του ερπυστή. Στο τέλος πρέπει να ενοποιήσετε να μερικά ευρετήριο για να φτιάξετε το τελικό. Διαβάστε τη σχετική ύλη[1].

Στο παράρτημα 1 μπορείτε να βρείτε μία περιγραφή για σχετικά με το αρχείο εισόδου που παράγεται από τον ερπυστή και χρησιμοποιείται (εκτός των άλλων) για τον ευρετηριασμό νέων εγγράφων.

Β. Κωδικοποίηση – Συμπίεση Δεδομένων

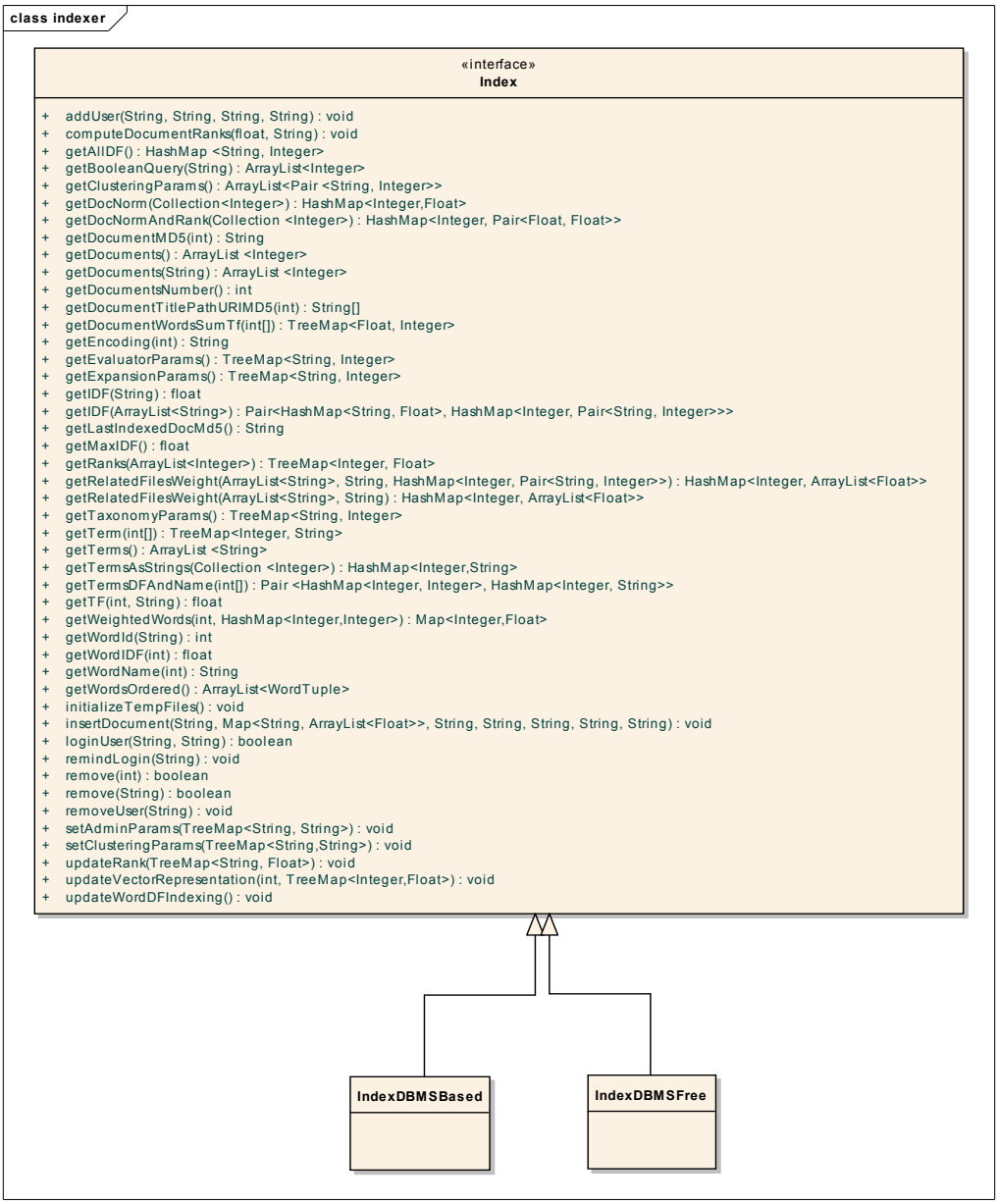
Αφού κατασκευάσετε τις κατάλληλες δομές πρέπει κατόπιν να εφαρμόσετε κάποια κωδικοποίηση στα δεδομένα ώστε να καταλαμβάνουν τον ελάχιστο δυνατό χώρο στον δίσκο. Μπορείτε να κάνετε χρήση συνεπτυγμένων κωδικών (βλ. Προγραμματιστική άσκηση 2). Για να επιτύχουμε πράγματι μείωση του αποθηκευτικού χώρου ενδείκνυται η ομαδοποίηση των εγγράφων και η εκχώρηση αναγνωριστικών βάσει του αποτελέσματος της ομαδοποίησης. Μια πιο γρήγορη λύση περιγράφεται στο άρθρο:

- http://soave.isti.cnr.it/~silvestr/wp-content/uploads/2007/04/paper.pdf

Γ. Επέκταση του Index, και χρήση του ευρετηρίου στο mitos

Επεκτείνετε το component ***Index*** ώστε το **mitos** να μπορεί να λειτουργεί με το νέο ευρετήριο που μόλις κατασκευάσατε. Συγκεκριμένα δημιουργήστε ένα interface το οποίο να περιέχει την συμπεριφορά που πρέπει να έχει το συστατικό (για DBMS-based index και για DBMS-Free index). Σκοπός είναι η δημιουργία του παρακάτω :

[1] Μπορείτε να συμβουλευτείτε και άλλες πηγές, π.χ. http://jodi.tamu.edu/Articles/v01/i05/Frieder/frieder3.pdf



Πληροφορίες για τον Query Evaluator και για τις μεθόδους του Index που χρησιμοποιούνται για την αποτίμηση των επερωτήσεων μπορείτε να βρείτε στο link http://groogle.csd.uoc.gr/apache2-default/index.php/Edit_Distance.

Δ. Μετρήσεις και Συγκρίσεις

Αφού ολοκληρωθεί η υλοποίηση σας θα πρέπει να συλλέξετε πληροφορίες σχετικές με το χρόνο που χρειάζεται για να γίνει η ευρετηρίαση των δεδομένων αλλά και την απόκριση του Μίτος, όπως το χρόνο που χρειάζεται για την αποτίμηση μιας επερώτησης. Οι μετρήσεις θα πρέπει να γίνουν και για τις δύο υλοποιήσεις του ευρετηριαστή(indexer), με το ανεστραμμένο ευρετήριο και με την Σχεσιακή Βάση Δεδομένων, ώστε να μπορεί να γίνει η σύγκριση των δύο υλοποιήσεων. Συγκεκριμένα οι μετρήσεις που πρέπει να κάνετε περιελαμβάνουν:

- Ευρετηριασμός της συλλογής με χρήση λίστας λέξεων αποκλεισμού (stopwords) και χρήση στελεχωτή (stemming). Συγκεκριμένα πρέπει να υπολογίσετε:
    - Χρόνος ευρετηριασμού (με το DBMS-free ευρετήριο και με το DBMS-based)

    - Μέγεθος λεξιλογίου
    - Μέγεθος posting file
- Ευρετηριασμός της συλλογής με παραλλαγές (χωρίς χρήση λέξεων αποκλεισμού, χωρίς χρήση στελεχωτή)
- Απόκριση του νέου ευρετηρίου για την αποτίμηση επερωτήσεων.

Οποιαδήποτε άλλη μέτρηση, σύγκριση κρίνεται απαραίτητη.

Καλή εργασία και διασκέδαση

Παράρτημα 1 (Εξοδος Ερπυστή / Είσοδος αρχείων στον Ευρετηριαστή)

Για να ευρετηριαστεί μία συλλογή πρέπει ο ***ευρετηριαστής*** να εντοπίσει τα αρχεία τα οποία έχει κατεβάσει ο ***ερπυστής***. Όταν ο ερπυστής "κατεβάζει" σελίδες τις αποθηκεύει τοπικά στον δίσκο και παράλληλα ενημερώνει ένα αρχείο με τις σελίδες τις οποίες έχει μέχρι στιγμής κατεβάσει καθώς και άλλες πληροφορίες σχετικά με την εκάστοτε σελίδα. Συγκεκριμένα ενημερώνει ένα αρχείο index το οποίο έχει την παρακάτω μορφή:

UrlMD5checksum	**normalizedURL originalURL**	**"Title" Encoding type lastModified serverDate**
<tab>	@<normalized 1st link retrieved from this page>	"<anchor text for this 1st link>"
<tab>	@<normalized 2st link retrieved from this page>	"<anchor text for this 2st link>"
	…	…
<tab>	@<normalized Nst link retrieved from this page>	"<anchor text for this Nst link>"

Επεξήγηση πεδίων:

- **UrlMD5CheckSum:** Μια χαρακτηριστική τιμή κατακερματισμού(MD5) του Url.
- **NormalizedURL:** Το κανονικοποιημένο Url.
- **OriginalURL:** Το URL πριν την κανονικοποίηση.
- **"Title":** Ο τίτλος του ανακτημένου εγγράφου
- **Encoding:** Η κωδικοποίηση της σελίδας (utf-8 κλπ...).
- **Type:** Ο τύπος (type) της σελίδας (html, pdf, κλπ).
- **lastModifiedDate:** Η ημερομηνία και ο χρόνος στον οποίο ο διακομιστής πιστεύει ότι τροποποιήθηκε το αρχείο στο οποίο αναφέρθηκε η αίτηση που έλαβε.
- **serverDate:** Η ημερομηνία και ο χρόνος στον οποίο δημιουργήθηκε το μήνυμα-απάντηση του διακομιστή. Για παράδειγμα: Date: Tue, 15 Nov 1994 08:12:31 GMT
- **Λίστα αποδεκτών συνδέσμων που ανακτήθηκαν από την σελίδα:** Μετά από τη γραμμή που θα περιέχει όλα τα παραπάνω πεδία ακολουθεί μία λίστα στοιχισμένη με στηλογνώμονες (tabs) που αναφέρει όλα τα αποδεκτά links που ανακτήθηκαν από την εν λόγω σελίδα (ένα link ανά γραμμή μαζί με το anchor text του). Anchor Text είναι το κείμενο που έχει ο εν λόγω σύνδεσμος (π.χ. this is anchor text)

Η κάθε σελίδα έχει και τοπικό αντίγραφο το οποίο μπορεί να το βρει κανείς ακολουθώντας το path του URL αρχίζοντας από το βασικό κατάλογο αποθήκευσης σελίδων. Για παράδειγμα το αντίγραφο της σελίδας index.html που ανήκει στο domain `**www.ics.forth.gr**' θα πρέπει να βρίσκετε στο **./www.ics.forth.gr/index.html**.

Περισσότερες λεπτομέρειες σχετικά με τον αρχείο index μπορείτε να βρείτε στην εκφώνηση του project σχετικά με τον ερπυστή (*Project Ταραντούλα*).