Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 – Συστήματα Ανάκτησης Πληροφοριών
2009-2010 Φθινοπωρινό Εξάμηνο

ΦΑΣΗ Β

Π11) (20) Θέλουμε το σύστημα μας να παρέχει και αναζήτηση εικόνων βάσει περιγραφών κειμένου. Για το λόγο αυτό πρέπει να ενσωματώσετε στο σύστημα σας κώδικα από τη μηχανή mitos ώστε να μπορείτε να διαβάσετε html σελίδες και να εξάγετε από αυτές το κείμενο και τις εικόνες.

Έτσι για κάθε html σελίδα θα πρέπει να δημιουργήσουμε ένα διάνυσμα (αντίστοιχο με αυτό που φτιάξατε για κάθε έγγραφο στην προηγούμενη φάση). **Καθοριστικής σημασίας είναι ο τρόπος κατασκευής και η βάρυνση αυτού του διανύσματος που θα περιγράφει κάθε html έγγραφο.** Περιγράψετε και σχολιάστε στην γραπτή αναφορά που θα παραδώσετε τον τρόπο βάρυνσης που επιλέξατε. Για τη βάρυνση του διανύσματος καλείστε να λάβετε υπόψη σας τις εξής πληροφορίες που μπορούμε να αντλήσουμε από μια html σελίδα :

- Τίτλο
- Επικεφαλίδες
- Keywords
- Description
- Bold
- Emphasis
- .....

Επιπλέον, το σύστημα σας όταν συναντάει μια html σελίδα πρέπει να αναγνωρίζει την εμφάνιση εικόνων (π.χ. <IMG SRC="myfoto.jpg" alt="Γιάννης Γιαννόπουλος" </IMG> ή <IMG SRC="http://www.ics.forth.gr/images/forth-Buck-building.jpg" </IMG>) ώστε να ενημερώνει ανάλογα το ευρετήριο. Εδώ τίθεται το ζήτημα του τρόπου παράστασης (και ευρετηρίασης) του περιεχομένου των εικόνων. Για μια εικόνα που είναι ενσωματωμένη σε μια ιστοσελίδα θα μπορούσαμε να αντλήσουμε τις εξής κειμενικές περιγραφές:

- όνομα εικόνας
- λεζάντα εικόνας (εάν υπάρχει)
- τίτλο σελίδας στην οποία βρίσκεται η εικόνα
- το κείμενο που «περιβάλλει» την εικόνα

Από όλα τα παραπάνω θα μπορούσαμε να κατασκευάσουμε ένα διάνυσμα, αντίστοιχο με αυτό που κρατάμε για τα αρχεία κειμένου, θεωρώντας κάθε εικόνα ως ξεχωριστό έγγραφο . Το όφελος από αυτήν την επιλογή είναι ότι η ανάκτηση εικόνων βάσει κειμένου, ανάγεται στην ανάκτηση κειμένων και άρα δεν θα χρειαστεί να γίνει καμία αλλαγή στο υποσύστημα αποτίμησης επερωτήσεων. Το μόνο που θα χρειαστεί να αλλάξετε είναι στο ευρετήριό σας να αποθηκεύετε στο "Documents File" και ένα μοναδικό αναγνωριστικό κάθε σελίδας προκειμένου να κρατάμε πληροφορία στο index σχετικά με τη σελίδα που εμφανίζεται μια εικόνα, όπως είναι για παράδειγμα το md5[1]. **Καθοριστικής σημασίας είναι ο τρόπος και η βάρυνση του διανύσματος που θα περιγράφει την εικόνα**. Περιγράψετε και σχολιάστε στην γραπτή αναφορά που θα παραδώσετε τον τρόπο βάρυνσης που επιλέξατε.

[1]Το md5 μπορείτε να το δημιουργείτε χρησιμοποιώντας το PATH και την κλάση **mitos.util.MD5**

Για να εξάγετε την πληροφορία από μια html σελίδα μπορείτε να χρησιμοποιήσετε την κλάση **mitos.lexicalAnalyzer.analyzers.HTMLAnalyzer**[2]**.** Για το σκοπό αυτό θα χρησιμοποιήσετε το **Jericho** (http://jericho.htmlparser.net/docs/index.html), ένα package που παρέχει έναν εύκολο και πλήρη τρόπο για την ανάλυση και επεξεργασία HTML σελίδων (μπορείτε να δείτε παραδείγματα χρήσης της στο παραπάνω url). Για κάθε σελίδα που θέλετε να επεξεργαστείτε, θα πρέπει να καλείτε την μέθοδο **analyze** της **HTMLAnalyzer** με όρισμα ένα **mitos.lexicalAnalyzer.analyzedFiles.AnalyzedDocument**[3]. Για κάθε αρχείο της συλλογής, θα δημιουργήσετε ένα αντικείμενο αυτού του τύπου, όπου θα πρέπει να εισάγετε πολύτιμες πληροφορίες, για το συγκεκριμένο αρχείο όπως είναι το **path, md5, encoding, type.** Καλώντας πλέον την **analyze** ο parser της HTML θα εξάγει πληροφορία από τον HTML κώδικα και θα ενημερώσει κατάλληλα τα πεδία του **AnalyzedDocument**. Τέτοια πεδία είναι ο τίτλος, το κείμενο της σελίδας, διάφορα στατιστικά σχετικά με τον αριθμό των λέξεων, καθώς και το διάνυσμα με τα βάρη των όρων, το οποίο κρατάμε στη μεταβλητή **HashMap< String, Pair< Float, ArrayList< Integer > > > wordsMap,** όπου για κάθε όρο κρατάμε ένα ζευγάρι με το tf και μια λίστα με τις εμφανίσεις τους στο κείμενο. Τέλος, πρέπει να κρατάτε για κάθε εικόνα ένα thumbnail της, το οποίο θα αποθηκεύετε στο directory που έχει η μεταβλητή **IMAGEFOLDER** στο **mitos.resources.Resources,** με όνομα "md5".jpg. Ένα παράδειγμα χρήσης της **HTMLAnalyzer** μπορείτε να δείτε στην **mitos.lexicalAnalyzer.LexExample** στο πακέτο που σας έχει δοθεί.

Επιπλέον, καλείστε να χρησιμοποιήσετε το Jericho προκειμένου να αναγνωρίζετε τις εικόνες (img tags) όπου υπάρχουν σε ένα html έγγραφο, και να μπορείτε να εξάγετε το όνομα τους (π.χ. εάν < img **src**=" http://www.ics.forth.gr/images/forth-Buck-building.jpg " /> τότε το όνομα της είναι το "forth-Buck-building") καθώς και αν έχει, την περιγραφή της (εάν < img **src**="" …**alt**='Περιστύλιο' />) τότε η περιγραφή της είναι η λέξη «Περιστύλιο»). Κάθε AnalyzedDocument κρατάει ένα ArrayList με αντικείμενα τύπου **mitos.lexicalAnalyzer.analyzedFiles.AnalyzedImages**[4]**,** τα οποία πρέπει να εμπλουτίσετε με ότι κρίνετε εσείς απαραίτητο (έτσι ώστε να κρατάτε όλες τις απαραίτητες πληροφορίες για μία εικόνα και να μπορείτε να δημιουργήσετε το κατάλληλο διάνυσμα, βάσει της βάρυνσης που έχετε προτείνει).

Έχοντας τελικά κάνει analyze ένα html κείμενο με εικόνες, στην συνέχεια μπορούμε να εισάγουμε τα διανύσματα τόσο της html σελίδας, όσο και των εικόνων της στο index που έχετε δημιουργήσει. Για παράδειγμα, αν ένα html αρχείο έχει 3 εικόνες θα σήμαινε ότι προκύπτουν 4 έγγραφα, από τα οποία 3 θα είχαν τύπο image[5] και ένα τύπο text/html και τα οποία θα έπρεπε να εισάγουμε στο index μας.

Η τεχνική που θα φανεί ότι δίνει πιο καλά αποτελέσματα, θα πάρει bonus 10% και η τεχνική θα ενσωματωθεί στη μηχανή mitos.

Θα ληφθούν υπόψη στη βαθμολογία η σχεδίαση και τεκμηρίωση του κώδικα.

---

2 http://groogle.csd.uoc.gr:8080/mitos/files/javadocs/html/d7/de0/classmitos_1_1lexicalAnalyzer_1_1analyzers_1_1htmlAnalyzer_1_1HTMLAnalyzer.html

3 http://groogle.csd.uoc.gr:8080/mitos/files/javadocs/html/d6/da5/classmitos_1_1lexicalAnalyzer_1_1analyzedFiles_1_1AnalyzedDocument.html

4 http://groogle.csd.uoc.gr:8080/mitos/files/javadocs/html/d2/d99/classmitos_1_1lexicalAnalyzer_1_1analyzedFiles_1_1AnalyzedImage.html

5 Θα χρησιμοποιήσουμε mimetypes του στυλ "image/gif", "image/jpg" για τον τύπο της εικόνας

Θα σας δοθεί μία συλλογή με html σελίδες και τις αντίστοιχες εικόνες τους.

Το πακέτο με τον κώδικα είναι διαθέσιμο στη σελίδα του μαθήματος.

Πρέπει να αλλάξετε τη μεταβλητή base στο mitos.resources.Resources, έτσι ώστε να δείχνει στο directory repository, που υπάρχει στο top-level του πακέτου.

Π12) (5) Δημιουργήστε μια γραφική διεπαφή χρήστη (GUI) για το σύστημα σας που να επιτρέπει ανάκτηση κειμένων βάσει λέξεων.

Π13) (5) Επεκτείνετε την διεπαφή ώστε να προσφέρει και καρτέλα Images για την ανάκτηση εικόνων. Καλείστε να χρησιμοποιήσετε το **Flexplorer API** που θα σας δοθεί), ώστε να υποστηρίξετε εύκολα τη δυνατότητα πολυδιάστατης πλοήγησης. Συγκεκριμένα η έδρα/διάσταση (facet) "FileType" μπορεί να έχει μια κατηγορία Images, παιδιά της οποίας θα είναι οι σχετικοί τύποι αρχείων. Η διαφορά είναι οτι στο χώρο των αποτελεσμάτων (δεξί frame) οι εικόνες είναι καλό να εμφανίζονται με διαφορετικό τρόπο απ' ότι οι ιστοσελίδες. Συγκεκριμένα για κάθε εικόνα, θα πρέπει να εμφανίζεται:
1. Μια μικρογραφία της
2. Το όνομά της
3. Σύνδεσμος προς την εικόνα
4. Σύνδεσμος (PATH) προς τη σελίδα που περιέχει την εικόνα.

Π14) (10) Σκεφτείτε πως θα αξιολογήσετε τις κειμενικές περιγραφές των εικόνων που εξάγετε. Σκεφτείτε μέτρα, τρόπους και συλλογές αξιολόγησης, και κάντε τις σχετικές μετρήσεις αποτελεσματικότητας (effectiveness). Συντάξτε σχετική γραπτή αναφορά.

Προαιρετικά
Ανάκτηση πληροφοριών και μετα-δεδομένα με χρήση του **PreScan** (http://www.ics.forth.gr/~marketak/PreScan/) το οποίο είναι ένα σύστημα το οποίο εξάγει τα ενσωματωμένα μεταδεδομένα από διάφορους τύπους αρχείων. Στην προκειμένη, θα μπορούσε να χρησιμοποιηθεί για να εξάγει πληροφορίες που είναι ενσωματωμένες στα αρχεία εικόνων (π.χ. χρόνος λήψης, διάφραγμα, κλπ) και οι οποίες θα μπορούσαν να αξιοποιηθούν ως μια επιπλέον διάσταση του FlexPlorer. Περισσότερες πληροφορίες θα δοθούν στις ομάδες που θα αποφασίσουν να το χρησιμοποιήσουν.

Χρονοδιάγραμμα

- ➔ Φάση Β: Π11-Π12(18 Νοέμβρη – 11 Δεκέμβρη)
  - Παραδοτέα: όπως στη Φάση Α
- ➔ Φάση Γ: Πειράματα και σύνταξη σχετικής αναφοράς (Π14). (12 Δεκέμβρη – 20 Δεκέμβρη)
  - Παραδοτέα: Η αναφορά και προαιρετικά ανανεωμένο .zip

Κάθε φάση θα εξεταστεί και θα βαθμολογηθεί ξεχωριστά.

Σχετικά με την παράδοση, είναι αποδεκτό η κάθε φάση να παραδοθεί μέχρι και 5 μέρες μετά από την ημερομηνία παράδοσης αλλά με 10% απώλειας του βαθμού ανά ημέρα.

Καλή εργασία