Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
ΗΥ463 - Συστήματα Ανάκτησης Πληροφοριών
2007-2008 Εαρινό Εξάμηνο

Προγραμματιστική Άσκηση

MyIR-B

Ανάθεση: 18 Μαρτίου
Παράδοση: 8 Απριλίου (3 εβδομάδες)

Επεκτείνετε το σύστημα που φτιάξατε στην προηγούμενη προγραμματιστική άσκηση (MyIR-A) έτσι ώστε:

Π1) Να διαβάζει μια λέξη από την κονσόλα και να τυπώνει τα αναγνωριστικά των εγγράφων στα οποία εμφανίζεται.

Π2) Να διαβάζει πολλές λέξεις από την κονσόλα (δηλαδή μια επερώτηση σε φυσική γλώσσα) και να τυπώνει την απάντηση ως προς το διανυσματικό μοντέλο.

Π3) Να κάνει ό,τι το (Π2) αλλά η απάντηση να υπολογίζεται βάσει του Okapi BM25[1].

Π4) Να μπορεί να τυπώσει το διάνυσμα ενός εγγράφου με βάρυνση TF-IDF. Δοκιμάστε διαφορετικούς τρόπους και συγκρίνετε το χρονικό τους κόστος. Ένας τρόπος είναι για κάθε λέξη του εν λόγω εγγράφου να ανοίξετε το αντίστοιχο αρχείο εμφανίσεων απ' όπου θα πάρετε το TF της λέξης. Ένας άλλος είναι ο επανυπολογισμός του TF της κάθε λέξης καθώς σαρώνετε το αρχείο.

Σημείωση: Αν είχατε υλοποιήσει στελέχωση και αποκλεισμό λέξεων, τότε εφαρμόστε αυτές τις λειτουργίες και στις επερωτήσεις (λέξεις εισόδου) του χρήστη.

MyIR++

Δημιουργείστε μια παραλλαγή (MyIR++) του συστήματος που έχετε φτιάξει μέχρι τώρα η οποία:

Π5) Να κάνει ομαδοποίηση εγγράφων και εκ νέου εκχώρηση αναγνωριστικών στα έγγραφα βάσει του αποτελέσματος της ομαδοποίησης. Επαναδημιουργία του αρχείου DocumentIDs.txt. Αναφέρετε χρόνους ομαδοποίησης για διάφορες.

Π6) Να επαναδημιουργεί τα αρχεία εμφανίσεων (δηλαδή τα αρχεία που βρίσκονται στο φάκελο CollectionIndex/InvertedLists) βάσει των νέων αναγνωριστικών. Πιο συγκεκριμένα τα αναγνωριστικά που θα εμφανίζονται σε αυτά τα αρχεία πρέπει να καταγράφονται με σχετικό/αυξητικό τρόπο. Ενημερώστε κατάλληλα τον κώδικα που αποτιμά επερωτήσεις. Βεβαιωθείτε ότι το κάνατε σωστά δοκιμάζοντας το σύστημα.
Ανάλογα θα πρέπει να πράξετε στην περίπτωση όπου δεν έχετε δημιουργήσει ένα αρχείο εμφανίσεων για κάθε λέξη αλλά ένα (θυμηθείτε InvFile) για τις εμφανίσεις όλων.

Π7) Κάντε μια παραλλαγή του (Π6) έτσι ώστε τα αναγνωριστικά που καταγράφονται με σχετικό/αυξητικό τρόπο να παριστάνονται με συνεπτυγμένους κωδικούς (π.χ. Elias-δ). Ενημερώστε κατάλληλα τον κώδικα που αποτιμά επερωτήσεις. Βεβαιωθείτε ότι το κάνατε σωστά δοκιμάζοντας το σύστημα.
Σας δίδετε έτοιμος κώδικας από το project του Terrier[2] το οποίο χρησιμοποιεί Elias gamma coding[3]. Στον κώδικα που σας δίδετε περιέχετε το πακέτο compression από το οποίο θα χρησιμοποιήσετε τις

[1] http://en.wikipedia.org/wiki/Probabilistic_relevance_model_(BM25)

κλάσεις BitOut.java και BitOutputStream.java καθώς και η κλάση DirectIndexBuilder.java που τις χρησιμοποιεί.

Π8) Για την ίδια συλλογή (/συλλογές), συγκρίνετε το χώρο που καταλαμβάνει το ευρετήριο του **MyIR** με το χώρο που καταλαμβάνει αυτό του **MyIR++**.

Π9 - Bonus) Δημιουργήστε μια γραφική διεπαφή χρήστη (GUI) του συστήματός σας.

Π10 – Bonus) Κάντε χρήση JUnit για τον έλεγχο ορθότητας του MyIR και του MyIR++ συστήματός σας.

Σχετικοί Σύνδεσμοι/Βοηθητικό Υλικό

- Elias: http://aliasi.com/lingpipe/docs/api/com/aliasi/io/BitInput.html

[2] http://ir.dcs.gla.ac.uk/terrier/publications/ounis06terrier-osir.pdf
[3] http://en.wikipedia.org/wiki/Elias_gamma_coding