ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΑΝΑΖΗΤΗΣΗ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

7ο ΕΞΑΜΗΝΟ

ΑΚ. ΕΤΟΣ 2014-2015

ΕΡΓΑΣΙΑ

Στόχος της εργασίας είναι η εξοικείωση με τις κλάσεις της βιβλιοθήκης Lucene η οποία είναι υλοποιημένη σε Java. Η βιβλιοθήκη αυτή προσφέρει βασικές λειτουργίες Ανάκτησης Πληροφοριών όπως οργάνωση συλλογής κειμένων με τη δημιουργία ευρετήριου καθώς και αναζήτηση κειμένων με βάση το ευρετήριο.

Η βιβλιοθήκη συμπεριλαμβάνεται μεταξύ των έργων (projects) της Apache Jakarta (http://lucene.apache.org/) και είναι ελεύθερη προς χρήση Η Lucene σήμερα χρησιμοποιείται ευρέως για την ανάπτυξη εφαρμογών που απαιτούν λειτουργίες ανάκτησης πληροφορίας.

1ο Μέρος (50%)

Με τη βοήθεια της βιβλιοθήκης Lucene θα υλοποιήσετε ένα σύστημα ανάκτησης πληροφοριών. Ο χρήστης του συστήματος θα μπορεί να διατυπώνει τα ερωτήματα μέσα από γραφικό περιβάλλον και το σύστημα θα προβάλλει τα κείμενα που ανακτήθηκαν κατά σειρά σχετικότητας, με πρώτο το πιο σχετικό. Θα πρέπει επίσης να προβάλλεται και το περιεχόμενο του κειμένου αν ο χρήστης το επιλέξει στη λίστα των αποτελεσμάτων.

Προκειμένου να είναι δυνατή η εκτίμηση της απόδοσης του συστήματος, θα χρησιμοποιήσετε τη συλλογή NPL (αρχείο npl.rar) η οποία περιλαμβάνει 11429 μικρά αποσπάσματα κειμένων (αρχείο doc-text) μαζί με 93 ερωτήσεις (information needs) οι οποίες αφορούν το περιεχόμενο των κειμένων της συλλογής (αρχείο query-text). Έχουν επίσης προσδιοριστεί τα σχετικά για κάθε ερώτηση κείμενα (αρχείο rlv-ass).

Χρησιμοποιώντας την παραπάνω πληροφορία των σχετικών κειμένων, θα εκτιμήσετε την απόδοση του συστήματος κατασκευάζοντας το διάγραμμα ακρίβειας-ανάκλησης για κάθε ερώτημα. Η γραφική παρουσίαση όλων αυτών των πληροφοριών για την απόδοση του συστήματος μπορεί να γίνει είτε από το ίδιο σύστημα ή εναλλακτικά με τη βοήθεια κάποιου εξωτερικού προγράμματος (π.χ. Excel).

Λεπτομέρειες Υλοποίησης

Πριν ένα σύστημα ανάκτησης πληροφορίας τεθεί σε λειτουργία θα πρέπει πρώτα να έχουν καθορισθεί οι όροι (λέξεις-κλειδιά) βάσει των οποίων θα γίνεται αναζήτηση στη συλλογή των κειμένων. Κατά τη διαδικασία αυτή, θα πρέπει να αγνοήσετε λέξεις που περιέχονται στο αρχείο common_words.txt. Το αρχείο αυτό περιέχει μία λίστα

από κοινές λέξεις (π.χ. συνδέσμους, αντωνυμίες, άρθρα, κτλ.) οι οποίες εμφανίζονται σχεδόν στα περισσότερα κείμενα και επομένως δεν έχουν κάποιο ιδιαίτερο σημασιολογικό περιεχόμενο. Επίσης για τη βελτίωση της ποιότητας των αποτελεσμάτων ανάκτησης, θα χρησιμοποιήσετε τον αλγόριθμο του Porter για stemming. Όπως είναι γνωστό, ο αλγόριθμος αυτός εξάγει τις γραμματικές ρίζες των λέξεων.

Το ευρετήριο καθώς και οι λειτουργίες αναζήτησης του συστήματος ανάκτησης πληροφοριών θα υλοποιηθούν με τη βοήθεια των κλάσεων της βιβλιοθήκης Lucene. Πλήρης περιγραφή των κλάσεων αυτών θα βρείτε στη διεύθυνση http://lucene.apache.org/.

2^{ο} Μέρος (50%)

Στο δεύτερο μέρος της εργασίας θα υλοποιήσετε μία τεχνική ανάδρασης συνάφειας και επέκτασης ερωτήματος. Έστω $q = (t_1, t_2, \dots, t_k)$ οι όροι της αρχικής ερώτησης. Μετά την εκτέλεση του ερωτήματος, ο χρήστης επιλέγει ποια κείμενα του αποτελέσματος θεωρεί σχετικά. Έστω D_l το σύνολο των αυτών των κειμένων και V_l το τοπικό λεξικό δηλ. το σύνολο των διαφορετικών όρων στο D_l. Ορίζουμε επίσης f_{ij} τη συχνότητα εμφάνισης του όρου k_i στο κείμενο $d_j \in D_l$ και τον πίνακα M διαστάσεων $|V_l| \times |D_l|$ όπου $M[i,j] = f_{ij}$. Στη συνέχεια, το γινόμενο $C = MM^T$ εκφράζει την τοπική συσχέτιση μεταξύ των όρων που βρίσκονται στα κείμενα D_l. Αναλυτικά το στοιχείο $C[u,v]$ δίνεται από τη σχέση $C[u,v] = \sum_{d_j \in D_l} f_{u,j}\, f_{j,v}$. Επίσης ορίζουμε και τον κανονικοποιημένο πίνακα C' ίδιων διαστάσεων με το C, όπου $C'[u,v] = \frac{C[u,v]}{C[u,u]+C[v,v]-C[u,v]}$.

Η επέκταση του αρχικού ερωτήματος q θα γίνει με βάση τον πίνακα C'. Ορίζουμε τη παράμετρο εισόδου n και για κάθε όρο t_i του αρχικού ερωτήματος ο οποίος εμφανίζεται στα κείμενα του D_l, βρίσκουμε τις n μεγαλύτερες τιμές στη γραμμή του πίνακα C' που αντιστοιχεί στον t_i. Στη συνέχεια, συμπεριλαμβάνουμε στη ερώτηση q τους όρους (στήλες του πίνακα C') που αντιστοιχούν στις παραπάνω μέγιστες τιμές. Έτσι η νέα ερώτηση θα περιέχει τους αρχικούς όρους όπως και τους πρόσθετους που προέκυψαν με αυτή τη διαδικασία. Ο χρήστης υποβάλλει τη νέα ερώτηση και μπορεί να προχωρήσει σε νέα επέκταση του ερωτήματος αν το επιθυμεί.

Παραδοτέα

- Περιγραφή του συστήματος με έμφαση στις κλάσεις και τις μεθόδους της Lucene που χρησιμοποιήσατε
- Κώδικας με την κατάλληλη τεκμηρίωση
- Εκτελέσιμο πρόγραμμα και οδηγίες εγκατάστασης
- Όλα τα αρχεία τα σχετικά με τις εκτιμήσεις της απόδοσης του συστήματος (διαγράμματα precision-recall)
- Ημερομηνία παράδοσης: Πρώτο δεκαήμερο Μαρτίου. Θα προσδιορισθεί με νεότερη ανακοίνωση.

Αριθμός ατόμων

Η εργασία μπορεί να παραδοθεί από ομάδες αυστηρώς μέχρι δύο ατόμων.

Τρόπος Βαθμολόγησης

Η εργασία είναι υποχρεωτική και βαθμολογείται με άριστα το 5. Η εργασία λαμβάνεται υπόψη, αν ο βαθμός του γραπτού είναι μεγαλύτερος ή ίσος του 4. Στη περίπτωση αυτή, ο βαθμός της εργασίας προστίθεται στο βαθμό της γραπτής εξέτασης. Το άριστα στην γραπτή εξέταση είναι το 7.