ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΩΝ

7ο ΕΞΑΜΗΝΟ

ΑΚ. ΕΤΟΣ 2007-2008

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 κειμένων.

1ο Μέρος (60%)

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 κείμενα της συλλογής CACM (αρχείο cacm.tar.gz). Πρόκειται για συνόψεις 3204 ερευνητικών άρθρων τα οποία έχουν δημοσιευθεί στο έγκυρο περιοδικό Communications of the ACM. Κάθε κείμενο περιλαμβάνει μία ομάδα πεδίων, μεταξύ των οποίων τα πιο σημαντικά είναι ο τίτλος του άρθρου, οι συγγραφείς καθώς και η σύνοψη/απόσπασμα από το κείμενο του άρθρου. Επίσης η συλλογή αυτή περιλαμβάνει 64 ερωτήσεις (information requests) οι οποίες αφορούν το περιεχόμενο των κειμένων της συλλογής (αρχείο query.text). Έχουν επίσης προσδιοριστεί τα σχετικά για κάθε ερώτηση κείμενα (αρχείο qrels.text).

Χρησιμοποιώντας την παραπάνω πληροφορία των σχετικών κειμένων για κάθε ερώτημα, θα εκτιμήσετε την απόδοση του συστήματος κατασκευάζοντας το διάγραμμα ακρίβειας-ανάκλησης για κάθε ερώτημα. Επίσης θα πρέπει να υπολογιστεί το διάγραμμα της μέσης ακρίβειας - ανάκλησης για όλες τις ερωτήσεις συνολικά. Επιπλέον για κάθε ερώτημα θα πρέπει να υπολογίζονται οι τιμές της “Average Precision at Seen Relevant Documents” και της R-precision. Η γραφική παρουσίαση όλων αυτών των πληροφοριών για την απόδοση του συστήματος μπορεί να γίνεται 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. Στη συγκεκριμένη συλλογή, για την εξαγωγή των όρων θα βασισθείτε αποκλειστικά στις συνόψεις, στους τίτλους καθώς και στα ονόματα των συγγραφέων των κειμένων αγνοώντας όλα τα υπόλοιπα πεδία. 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 βάση τις κλάσεις της βιβλιοθήκης Lucene. Πλήρης περιγραφή των κλάσεων αυτών θα βρείτε στη διεύθυνση http://lucene.zones.apache.org:8080/hudson/job/Lucene-Nightly/javadoc/index.html

2ο Μέρος (25%)

Επίσης το σύστημα θα δίνει τη δυνατότητα στο χρήστη να βελτιώσει την ποιότητα των αποτελεσμάτων με διαδοχικούς γύρους ανάδρασης (relevance feedback). Καταρχήν, η βελτίωση/επαναδιατύπωση του αρχικού ερωτήματος θα γίνεται αυτόματα από το σύστημα.

Πιο συγκεκριμένα, τα βασικά βήματα έχουν ως εξής:

1. Επιλέγονται τα πρώτα κ κείμενα από τη ταξινομημένη λίστα των κειμένων που προκύπτει από την εκτέλεση της αρχικής ερώτησης .
2. Για κάθε ένα από τα κ κείμενα, επιλέγονται οι λ πιο συχνοί όροι στο κείμενο.
3. Στην αρχική ερώτηση προστίθενται όλοι παραπάνω όροι δηλ. οι λ όροι από κάθε ένα από τα κ κείμενα.
4. Αν κάποιοι από τους πρόσθετους όρους υπάρχουν ήδη στο αρχικό ερώτημα αγνοούνται.
5. Η ερώτηση υποβάλλεται εκ νέου.

Οι παράμετροι κ και λ θα μπορούν να αλλάζουν δυναμικά από το χρήστη.

3ο Μέρος (15%)

Επιπλέον, το σύστημα θα δίνει τη δυνατότητα ανάδρασης από το χρήστη. Ο χρήστης μέσα από κατάλληλη διεπαφή θα μπορεί να χαρακτηρίζει ποια κείμενα είναι σχετικά και ποια άσχετα από αυτά που ανακτήθηκαν

Λαμβάνοντας υπόψη την κρίση του χρήστη, το σύστημα θα τροποποιεί τα βάρη της αρχικής ερώτησης σύμφωνα με τον τύπο Standard_Rochio:

$$\vec{q}_{new} = \alpha \vec{q}_{old} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

Επειδή η βιβλιοθήκη της Lucene δεν παρέχει απευθείας τρόπο για την υλοποίηση του παραπάνω τύπου, προσπαθήστε να προσεγγίσετε τη «λογική» του τύπου όσο πιστά μπορείτε με τη χρήση μεθόδων που παρέχονται από τη Lucene

Οι συντελεστές α, β, γ αποτελούν παράμετροι λειτουργίας του συστήματος και θα πρέπει να μπορούν να μεταβάλλονται κατά τη λειτουργία του συστήματος

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, R-precision κτλ.)
- Ημερομηνία παράδοσης: ημερομηνία εξέτασης του μαθήματος

Αριθμός ατόμων

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

Τρόπος Βαθμολόγησης

Η εργασία είναι υποχρεωτική και βαθμολογείται με άριστα το 4. Ο βαθμός αυτός προστίθεται στο βαθμό της γραπτής εξέτασης. Το άριστα στην γραπτή εξέταση είναι το 7.