ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΑΝΑΖΗΤΗΣΗ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ

7ο ΕΞΑΜΗΝΟ

ΑΚ. ΕΤΟΣ 2023-2024

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

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. Για κάθε ανακτηθέν κείμενο, θα πρέπει να προβάλλεται επίσης και κάποιο απόσπασμά του με επισημασμένους τους όρους της ερώτησης που εντοπίστηκαν στο συγκεκριμένο απόσπασμα. Μία επιπλέον δυνατότητα που θα πρέπει να προσφέρει η διεπαφή χρήστη είναι η επιλογή «Όμοιο με αυτό» δίπλα σε κάθε κείμενο της λίστας των αποτελεσμάτων προκειμένου ο χρήστης να μπορεί, αν το επιθυμεί, να βρει περισσότερα κείμενα με περιεχόμενο όμοιο με το επιλεγμένο κείμενο.

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. 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 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/.

2ο Μέρος

Στο 2ο μέρος της εργασίας θα τροποποιήσετε και θα επεκτείνετε τον κώδικα που αναπτύξατε στο 1ο μέρος. Συγκεκριμένα:

- Θα τροποποιήσετε τον αλγόριθμο που χρησιμοποιεί η βιβλιοθήκη Lucene για την κατάταξη των εγγράφων, ώστε να χρησιμοποιείται το μοντέλο βαθμολόγησης **apn.apn** (δείτε την ενότητα 6.4 του [1]). Θα χρειαστεί να υπερκαλύψετε (override) τις κατάλληλες μεθόδους της Lucene που ορίζουν το μοντέλο βαθμολόγησης.
- Αντί ενός, θα δημιουργήσετε δύο λεξικά στο ίδιο πρόγραμμα. Στο ένα θα συμπεριλάβετε μόνο τους όρους των τίτλων των άρθρων (αγνοώντας πάλι τις λέξεις που περιέχονται στο αρχείο common_words.txt). Στο δεύτερο θα συμπεριλάβετε μόνο τους όρους των συνόψεων (abstracts) των άρθρων (αγνοώντας πάλι τις λέξεις που περιέχονται στο αρχείο common_words.txt). Τα ερωτήματα του αρχείου query.text θα υποβάλλονται και στα δύο λεξικά. Σκοπός είναι να αξιολογήσετε κατά πόσο οι τίτλοι των άρθρων είναι αντιπροσωπευτικοί των συνόψεων. Για να γίνει αυτό θα πρέπει να συγκρίνετε την κατάταξη των εγγράφων που θα επιστρέφεται για κάθε ερώτημα. Ως μέτρο ομοιότητας δύο κατατάξεων μπορεί να χρησιμοποιηθεί το πλήθος των αναστροφών (inversions) των δύο κατατάξεων. Το πλήθος των αναστροφών μπορεί να υπολογιστεί σε χρόνο O(n·logn), με χρήση της μεθόδου Διαίρει και Βασίλευε. Μπορείτε να βρείτε περισσότερες λεπτομέρειες στον παρακάτω υπερσύνδεσμο:
  https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjQ4qjh9Jn7AhXlRPEDHRPXAVoQFnoECA4QAQ&url=https%3A%2F%2Fhomes.cs.washington.edu%2F~jrl%2Fteaching%2Fcse312au10%2Flec24.pdf&usg=AOvVaw0XkX_rwm2DxYrjv-rMnq0x

Προτεινόμενη Βιβλιογραφία

[1] Manning, Christopher D., Hinrich Schütze, and Prabhakar Raghavan, «Εισαγωγή στην Ανάκτηση Πληροφορίας», Εκδόσεις Κλειδάριθμος, 2012.

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: Κυριακή 14/01/2024**

Αριθμός ατόμων

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

Τρόπος Βαθμολόγησης

Η εργασία είναι υποχρεωτική και βαθμολογείται με άριστα το 3. Ο βαθμός αυτός προστίθεται στο βαθμό της γραπτής εξέτασης. Το άριστα στην γραπτή εξέταση είναι το 7. Για να συνεκτιμηθεί ο βαθμός της εργασίας, θα πρέπει στη γραπτή εξέταση ο βαθμός να είναι τουλάχιστον 2,8 (=4*7/10).