Ανάκτηση Πληροφορίας

Ράμμος Θωμάς ΑΜ: 4583
Μητρόπουλος Γεώργιος ΑΜ:4733

Link στο GitHub repository: https://github.com/Thomas-Rammos/InformationRetrieval

Εισαγωγή

Lucene πρόγραμμα ειδικά σχεδιασμένο για αναζήτηση πληροφορίας από επιστημονικά άρθρα από CSV αρχείο. Το πρόγραμμα διαβάζει δεδομένα από επιστημονικά άρθρα από ένα αρχείο CSV, ευρετηριάζει τα δεδομένα αυτά με τη χρήση της Lucene βιβλιοθήκης και δίνει τη δυνατότητα στους χρήστες να αναζητήσουν τα ευρετηριασμένα αυτά δεδομένα χρησιμοποιώντας λέξεις-κλειδιά, αναζήτηση πεδίου δηλαδή μπορούν να φιλτράρουν την αναζήτησή τους με βάση πεδία όπως ο τίτλος, abstract και full text. Το σύστημα διατηρεί επίσης ένα ιστορικό των ερωτημάτων των χρηστών και παρέχει προτάσεις για μελλοντικά ερωτήματα. Επίσης υποστηρίζει αναζήτηση πεδίου και με το όνομα του συγγραφέα.

Συλλογή

Για την δημιουργία της συλλογής (*corpus*) κατεβάσαμε μια έτοιμη συλλογή από το Kaggle: **All NeurIPS (NIPS) Papers.**

Από όλα τα δεδομένα της παραπάνω συλλογής, επιλέχθηκαν τυχαία 250 γραμμές που να έχουν συμπληρωμένα τα πεδία ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’. Αυτά τα δεδομένα τα αποθηκεύσαμε σε ένα νέο αρχείο csv που θα είναι αυτό που θα επεξεργαστούμε.
Οι συγκεκριμένες ενέργειες έγιναν με τη χρήση ενός script σε python.

Ανάλυση κειμένου και κατασκευή ευρετηρίου

Το αρχείο CSV, το οποίο αποτελεί την πηγή δεδομένων, αναμένεται να έχει πέντε στήλες: ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’. Το σύστημα διαβάζει το αρχείο CSV γραμμή προς γραμμή και αναλύει αυτά τα πέντε πεδία. Κάθε άρθρο αντιμετωπίζεται ως έγγραφο και το αντίστοιχο ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’ ευρετηριάζονται στο Lucene χρησιμοποιώντας έναν StandardAnalyzer. Για την ανάλυση των πεδίων θα χρησιμοποιηθεί ο StandardAnalyzer(). Το ευρετήριο κατασκευάζεται χρησιμοποιώντας πέντε πεδία (‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’), καθένα από τα οποία είναι ένα TextField.

Αναζήτηση

Για τα ερωτήματα αναζήτησης, το σύστημα θα υποστηρίζει αναζήτηση με λέξεις-κλειδιά και αναζήτηση με βάση τα πεδία. Ο χρήστης θα μπορεί να κάνει αναζήτηση με λέξεις-κλειδιά σε όλα τα πεδία ή να περιορίσει την αναζήτηση σε ένα συγκεκριμένο πεδίο (π.χ. ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’). Ακόμα, το σύστημα αναζήτησης θα διατηρεί ιστορικό των ερωτημάτων του χρήστη. Κάθε νέο ερώτημα προστίθεται σε ένα ευρετήριο προηγούμενων ερωτημάτων. Αυτά τα προηγούμενα ερωτήματα θα επιστρέφονται ως προτάσεις στον χρήστη. Έτσι, ο χρήστης θα μπορεί να αναζητήσει παλαιότερα queries.

Παρουσίαση αποτελεσμάτων

Για την παρουσίαση των αποτελεσμάτων θα χρησιμοποιήσουμε JavaFX. Συγκεκριμένα τα αποτελέσματα θα παρουσιάζονται σε ένα περιβάλλον που θα μοιάζει με ιστοσελίδα. Κάθε αποτέλεσμα αναζήτησης θα εμφανίζεται ως συνδυασμός του ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’. Ο χρήστης θα μπορεί να επιλέξει να κάνει αναζήτηση σε συγκεκριμένα πεδία - ‘source_id’, ’year’, ’title’, ’abstract’ και ‘full_text’ - χρησιμοποιώντας αναζήτηση. Όταν θα εκτελείται μια αναζήτηση, το ερώτημα αναζήτησης θα προστίθεται στο ιστορικό αναζήτησης.