Πανεπιστημιο Ιωαννινων
Τμήμα Μηχανικών Η/Υ και Πληροφορικής 2023-2024
Ανάκτηση Πληροφορίας

Μηχανή αναζήτησης πληροφορίας από επιστημονικά άρθρα
1η Φαση

Εμμανουηλίδης Εμμανουήλ 4669

1. Εισαγωγή

Στόχος της άσκησης αυτής είναι ο σχεδιασμός και η υλοποίηση ενος συστήματος αναζήτησης πληροφοριών για επιστημονικά άρθρα. Η αρχιτεκτονική του συστήματος θα βασιστεί στο μοντέλο Model-View-Controller με προθεση την αξιοποίηση κατάλληλων Design Pattern που θα διευκολύνουν τη συντήρηση το testing και την επέκταση του παραγομενου κωδικα.

Για την ανάλυση του κειμένου και την κατασκευή των ευρετηρίων και για τις λειτουργίες της αναζήτησης θα χρησιμοποιήσουμε την βιβλιοθήκη ανοιχτού κώδικα της Java - όπως ζητάει και η άσκηση - Lucene.

Για την υλοποίηση του UI θα χρησιμοποιηθεί η βιβλιοθήκη γραφικών JavaFX.

2. Δημιουργία dataset - Συλλογή Εγγράφων

Για το dataset θα χρησιμοποιηθεί περιεχόμενο από την online κοινότητα data science του Kaggle. Πιο συγκεκριμένα θα χρησιμοποιήσουμε ένα dataset με 7241 επιστημονικά άρθρα που δημοσίευσε το συνέδριο του Neural Information Processing Systems απο το 1987-2017. Αυτό το dataset περιέχει την ημερομηνία δημοσίευσης, τον τίτλο, τις πληροφορίες του συγγραφέα, περιλήψεις και ολοκληρο το κείμενο.

3. Περιγραφή του Σχεδιασμού του Συστήματος

- **Στόχος του συστήματος**

Ο στόχος του συστήματος είναι με λέξεις κλειδια ο χρήστης να αναζητεί και να βρίσκει την πιο σχετική πληροφορία σύμφωνα με την αναζήτηση του.

- **Ανάλυση κειμένου και κατασκευή ευρετηρίου**

Για αρχή πρέπει να δημιουργήσουμε αυτό που η Lucene ονομάζει Index. Δηλαδή κάθε τραγούδι θα αναπαρασταθεί ως ένα Document το σύνολο των οποίων θα είναι το Index μας.

Κάθε Document θα εμπεριέχει τα προαναφερθέντα πεδία. Κάθε πεδίο μπορεί να ευρετηριοποιηθεί (indexed), δηλαδή η τιμή του αναλύεται και χωρίζεται σε λέξεις δίνοντας έτσι την δυνατότητα αναζήτησης.

Στην συνέχεια με την βοήθεια της κλάσης Analyzer της Lucene θα μετατρέψουμε το .csv αρχείο σε όρους αναζήτησης.

Στην συνέχεια μέσω της κλάσης Tokenizer εξάγεται και χωρίζεται το κείμενο σε λέξεις (tokens).

Έπειτα μέσω της Lucene StandardAnalyzer θα εφαρμόσουμε τεχνικές επεξεργασίας όπως το Stemming, το Stop Word Filtering κτλ.

Τέλος με χρήση του IndexWriter κάθε Document θα προστεθεί στο Index Directory.

- **Αναζήτηση**

Όταν ο χρήστης πληκτρολογει κάποιες λέξεις στην γραμμή αναζήτησης στόχος μας είναι να τις επεξεργαστούμε και να δημιουργήσουμε ένα Query.

Μέσω της QueryParser και της StandardAnalyzer δημιουργούμε το παραπάνω αντικείμενο.

Το Query όντας μια abstract κλάση μας δίνει την δυνατότητα να δημιουργήσουμε διαφορετικά αντικείμενα Query τα οποία αλλάζουν και το αντίστοιχο είδος αναζήτησης.

Τέλος, θα χρησιμοποιήσουμε την κλάση IndexSearcher η οποία μας δίνει την δυνατότητα να χρησιμοποιήσουμε διάφορες μεθόδους πάνω στο index μας. Η κλάση αυτή θα επιστρέφει αντικείμενα TopDocs με στόχο ο μέγιστος αριθμός επιστρεφόμενων αποτελεσμάτων να περιοριστεί. Έπειτα, θα γίνει το απαραίτητο scoring και ταξινόμηση των αποτελεσμάτων.

- **Παρουσίαση Αποτελεσμάτων**

Τα αποτελέσματα θα παρουσιάζονται στο UI (το οποίο θα υλοποιηθεί όπως αναφέρθηκε και προηγουμένως με την βιβλιοθήκη JavaFX) με βάση την σχετικότητα τους στο πεδίο αναζήτησης του χρήστη.

Το link για το database θα χρησιμοποιηθεί:
https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated/data?select=papers.csv

Το link για το GitHub repository της εργασίας:
https://github.com/emmemman/Information-Retrieval