Πανεπιστήμιο Ιωαννίνων Ανάκτηση Πληροφορίας

Τμήμα Μηχανικών Η/Υ & Πληροφορικής Ακαδημαϊκό Έτος 2022-2023

Μηχανή αναζήτησης τραγουδιών - Φάση 1η

Κωνσταντίνος Δεδικούσης - 2962
Δημητρόπουλος Δημήτριος - 4352

Εισαγωγή

Στόχος της παρούσας άσκησης είναι ο σχεδιασμός και η υλοποίησης ενός συστήματος αναζήτησης πληροφοριών για τραγούδια. Η αρχιτεκτονική του συστήματος θα βασιστεί στο μοντέλο Model-View-Controller με πρόθεση την αξιοποίηση κατάλληλων Design Pattern που θα διευκολύνουν τη συντήρηση το testing και την επέκταση του παραγώμενου κώδικα.

Για την ανάλυση του κειμένου και την κατασκευή των ευρετηρίων και για τις λειτουργίες της αναζήτησης θα χρησιμοποιήσουμε τη βιβλιοθήκη ανοιχτού κώδικα της Java - που μας υποδείχθηκε - Lucene.

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

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 τα κορυφαία 500 τραγούδια όλων των εποχών το οποίο θα περιέχει τον τίτλο, το άλμπουμ, την ημερομηνία κυκλοφορίας, τον τραγουδιστή, τον παραγωγό, τον συνθέτη καθώς και μια περιγραφή των στίχων του τραγουδιού.

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.

Τέλο, 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 Swing της Java) με βάση την σχετικότητα τους στο πεδίο αναζήτησης του χρήστη.

Το dataset που θα χρησιμοποιηθεί μπορείτε να το βρείτε στον ακόλουθο σύνδεσμο https://www.kaggle.com/code/mpwolke/500-greatest-songs σε csv format. Να σημειωθεί, πως αποφασίσαμε να κάνουμε drop τα cols streak και position καθώς θεωρήσαμε πως δεν είναι σχετικά/χρήσιμα για αναζήτηση.

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