

Μηχανή Αναζήτησης Τραγουδιών

GitHub repository link: <https://github.com/MinasE98/song-search-engine.git>

Εισαγωγή:

Στόχος αυτού του project είναι η ανάπτυξη μιας μηχανής αναζήτησης τραγουδιών χρησιμοποιώντας τη βιβλιοθήκη Lucene. Η μηχανή αναζήτησης θα πρέπει να είναι σε θέση να επεξεργάζεται και να αναζητά τραγούδια από το σύνολο δεδομένων Spotify Million Song Dataset.

Σύνολο Δεδομένων:

Το σύνολο δεδομένων περιέχει πληροφορίες για περισσότερα από 44000 τραγούδια, πάνω από 600 καλλιτέχνες και περιλαμβάνει τα εξής πεδία:

1. artist: Ο καλλιτέχνης που ερμηνεύει το τραγούδι.
2. song: Ο τίτλος του τραγουδιού.
3. link: Ο σύνδεσμος προς το τραγούδι στο διαδίκτυο.
4. text: Οι στίχοι του τραγουδιού.

Το αρχείο είναι σε μορφή CSV.

Για την πρώτη φάση του project έχουμε επιλέξει τουλάχιστον 20 έγγραφα από τη συλλογή τραγουδιών που θα χρησιμοποιήσουμε για την εργασία μας. Τα έγγραφα προέρχονται από τη βάση δεδομένων του Spotify (πηγή: <https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset>) και έχουν εισαχθεί σε ένα αρχείο CSV με την ονομασία first_20_songs.csv.

Ανάλυση κειμένου και κατασκευή ευρετηρίου:

Για την προεπεξεργασία των δεδομένων, θα πρέπει να φορτώσουμε το αρχείο CSV και να εξάγουμε τις απαραίτητες πληροφορίες για κάθε τραγούδι. Στη συνέχεια, θα πρέπει να εφαρμόσουμε τεχνικές προεπεξεργασίας κειμένου όπως stemming, απαλοιφή stop words, επέκταση συνωνύμων, κλπ χρησιμοποιώντας το Lucene Analyzer. Επίσης, κάποιες λειτουργίες, όπως η διόρθωση τυπογραφικών λαθών, ή η επέκταση ακρωνύμων, μπορούν να γίνουν εναλλακτικά κατά τη διάρκεια της αναζήτησης.

Με τη χρήση της βιβλιοθήκης Lucene IndexWriter, θα δημιουργήσουμε ευρετήρια για τα πεδία που επιλέξαμε. Τα ευρετήρια θα βελτιστοποιηθούν για την υποστήριξη διάφορων τρόπων αναζήτησης.

Αναζήτηση:

Η αναζήτηση θα βασίζεται στα ευρετήρια που έχουν δημιουργηθεί. Οι χρήστες θα μπορούν να αναζητήσουν τραγούδια με λέξεις κλειδιά και επίσης να διατηρεί πληροφορία για την ιστορία των αναζητήσεων. Μπορούμε να χρησιμοποιήσουμε το Lucene QueryParser της Lucene για να δημιουργήσουμε ερωτήματα αναζήτησης και να λάβουμε τα αποτελέσματα. Στη συνέχεια, θα

χρησιμοποιήσουμε το Lucene IndexSearcher για να αναζητήσουμε τα ευρετήρια και να επιστρέψουμε τα σχετικά αποτελέσματα στους χρήστες .

Παρουσίαση Αποτελεσμάτων:

Για να καταστήσουμε την εφαρμογή φιλική προς τον χρήστη, θα χρειαστεί να δημιουργήσουμε μια διεπαφή χρήστη που να επιτρέπει εύκολη πλοήγηση και αναζήτηση. Για να το επιτύχουμε αυτό το σύστημα μας θα παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους με το ερώτημα. Ακόμα θα παρουσιάζει τα αποτελέσματα ανά 10, με δυνατότητα στον χρήστη να προχωρήσει στα επόμενα. Επίσης οι λέξεις κλειδιά θα παρουσιάζονται τονισμένες στο αποτέλεσμα και θα παρέχει δυνατότητα ομαδοποίησης των αποτελεσμάτων με κριτήρια που θα ορίσουμε.

Συνοψίζοντας, η εργασία μας θα αναπτύξει μια μηχανή αναζήτησης τραγουδιών που βασίζεται στη συλλογή δεδομένων του Spotify Million Song Dataset, παρέχοντας στους χρήστες μια ισχυρή και εύχρηστη εργαλειοθήκη για την ανακάλυψη και εξερεύνηση μουσικής. Η εφαρμογή μας θα χρησιμοποιεί την τεχνολογία της Apache Lucene για την κατασκευή ευρετηρίων και την αναζήτηση, ενώ θα προσαρμόσουμε τις λειτουργίες της για να καλύψουμε τις απαιτήσεις του συγκεκριμένου προβλήματος.

Κατά την ανάπτυξη της εφαρμογής, θα δοκιμάζουμε και θα αξιολογούμε διάφορες προσεγγίσεις και τεχνικές, προκειμένου να επιτύχουμε τη βέλτιστη απόδοση και λειτουργικότητα. Τέλος, θα επικεντρωθούμε στην παρουσίαση των αποτελεσμάτων με έναν καθαρό και φιλικό προς το χρήστη τρόπο, για να καταστήσουμε την εμπειρία αναζήτησης όσο το δυνατόν πιο ευχάριστη και παραγωγική.