Πανεπιστήμιο Ιωαννίνων

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Ανάκτηση Πληροφορίας

1η Φάση

Μέλη ομάδας:

Ελευθέριος-Χρήστος Δριτσώνας, 4668
Φωτόπουλος Στέφανος, 4829

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2023

ΠΑΡΑΔΟΣΗ: Παρασκευή 7 Απριλίου

Contents

1.Περιγραφή της εργασίας

Το link για το github είναι το παρακάτω:
https://github.com/stef-fot/Anaktisi-Pliriforias.git

Ερώτημα Α

Στην συγκεκριμένη αναφορά θα περιγράψουμε τα βήματα που σχετίζονται με την υλοποίηση της 1ης φάσης στο μάθημα της Ανάκτησης Πληροφορίας.

ΣΥΛΛΟΓΗ ΕΓΓΡΑΦΩΝ

Σε αυτό το βήμα συλλέξαμε τα έγγραφα που θα χρησιμοποιήσουμε από τον παρακάτω σύνδεσμο:
https://www.kaggle.com/datasets/notshrirang/spotify-million-song-dataset
(είναι ο σύνδεσμος που υπάρχει και στις διαφάνειες του μαθήματος).Στο συγκεκριμένο λίνκ περιέχονται 643 καλλιτέχνες με συνολικά 44824 τραγούδια(έγγραφα).
Στο GitHub όταν ανεβάσαμε το dataset διαμαρτυρήθηκε ότι το αρχείο ήταν μεγαλύτερο από τα 25MB με αποτέλεσμα να ανεβάσουμε το ίδιο αρχείο αλλά συμπιεσμένο(zip).

Το format του συγκεκριμένου αρχείου είναι σε μορφή .csv (comma separated values).

Στην πρώτη στήλη του αρχείου βλέπουμε το όνομα του καλλιτέχνη ή του συγκροτήματος.
Στη δεύτερη έχουμε το όνομα του τραγουδιού.
Η Τρίτη στήλη έχει το Link του τραγουδιου και τελειώνει σε .html.
Τέλος, η τέταρτη στήλη έχει όλα τα lyrics(στίχους) των τραγουδιών.

Ερώτημα Β

Ανάλυση κειμένου και κατασκευή ευρετηρίου:
Αφού πρώτα δημιουργήσουμε έναν StandardAnalyzer και ένα ByteBuffersDirectory για να αποθηκεύουμε την πληροφορία θα χρειαστεί να φτιάξουμε και έναν IndexWriterConfig ο οποίος θα διατρέχει το ευρετήριο στην περίπτωση μας το αρχείο .csv με τα τραγούδια και θα καταγραφεί τις τιμές αυτού.
Έπειτα, θα έχουμε έναν QueryParser όπου θα εισάγουμε όλες τις πιθανές ερωτήσεις που μπορεί να κάνει ο χρήστης.Στη συνέχεια,υπάρχει ο DirectoryReader ο οποίος ανοίγει το ByteBuffersDirectory που δημιουργήσαμε και ένας IndexSearcher που έχει ως σκοπό να αναζητήσει πόσες φορές ικανοποιείται το query του χρήστη.
Αυτή την πληροφορία την κρατάμε σε έναν πίνακα hits όπου μετράμε τα scores. Επίσης, θα υπάρχει μια void συνάρτηση addDoc η οποία θα κατηγοριοποιεί τα πεδία του αρχείου .csv στα πεδία : Τίτλος τραγουδιού, Όνομα καλλιτέχνη , Ημερομηνία.

Εισαγωγή: Ποιος είναι ο στόχος και η λειτουργικότητα του συστήματος

Ο στόχος για την συγκεκριμένη εργασία είναι να δημιουργήσουμε ένα interface στο οποίο ο χρήστης θα μπορεί να θέτει ερωτήματα τα οποία θα σχετίζονται με τραγούδια που είναι αποθηκευμένα στην Βάση Δεδομένων. Αυτό επιτυγχάνεται μέσω μιας γραφικής διεπαφής (GUI) για παράδειγμα όπως το search bar της μηχανής αναζήτησης της Google.

Αναζήτηση: Πως θα γίνεται η αναζήτηση, τα είδη των ερωτημάτων
Αρχικά, η αναζήτηση θα γίνεται μέσω ενός search box το οποίο θα εμφανίζεται με σκοπό να εισάγει ο χρήστης την επιθυμητή ερώτηση που θέλει να κάνει στην βάση με τα τραγούδια.
Τα επιτρεπτά ερωτήματα που μπορεί να εισάγει είναι τα παρακάτω:

- Το όνομα του καλλιτέχνη
- Τον τίτλο κάποιου τραγουδιού
- Κάποιον συγκεκριμένο στίχο ή στίχους από τραγούδι
- Ημερομηνία Δημιουργίας του τραγουδιού

Παρουσίαση Αποτελεσμάτων: Πως θα παρουσιάζονται τα αποτελέσματα

Αρχικά, θα βρίσκουμε τα τραγούδια που ικανοποιούν το query(ερώτηση) που υπέβαλλε ο χρήστης και από αυτά θα εμφανίζονται σε ομάδες των 10 σε φθίνουσα σειρά με βάση τον αριθμό εμφανίσεων αυτού του ερωτήματος ανά τραγούδι. Έπειτα ο χρήστης θα έχει την επιλογή να δει τις επόμενες 10 πλειάδες(τραγούδια) που ικανοποιούν την ερώτηση.
Αυτό γίνεται μέχρι να τελειώσουν τα τραγούδια που έχουν αυτή την λέξη .Στο δεξί μέρος της γραμμής αναζήτησης θα υπάρχουν και suggested queries τα οποία μπορεί να επιλέξει ο χρήστης και να δει τα αποτελέσματά τους.