Πήττας Νικόλαος, ΑΜ:5007

Ανάκτηση πληροφορίας 1η φάση εργασίας

GitHub repository link: https://github.com/Nik-Pt/InformationRetrievalProject.git

README file copy:

Συλλογή εγγράφων:

Η συλλογή των εγγράφων που θα χρησιμοποιηθεί θα περιέχει έγγραφα τύπου(format) CSV. Συγκεκριμένα υπάρχει διαθέσιμος ένας φάκελος που περιέχει 21 διαφορετικά έγγραφα οπού το κάθε ένα έχει ως όνομα το όνομα του τραγουδιστή σε format CSV.Κάθε αρχείο περιέχει περίπου 200 με 300 διαφορετικά τραγούδια στα οποία δεν αναφέρονται μόνο οι στοίχοι αλλά θα περιέχουν επίσης όνομα τραγουδιστή, τίτλο τραγουδιού, έτος κυκλοφορίας και ακριβής ημερομηνία κυκλοφορίας τραγουδιού, πεδία τα οποία μπορούν να χρησιμοποιηθούν κατά την αναζήτηση. Η συγκεκριμένη συλλογή βρίσκεται στο παρακάτω URL: https://www.kaggle.com/datasets/deepshah16/song-lyrics-dataset ωστόσο κατά την διάρκεια της υλοποίησης υπάρχει η πιθανότητα να μεγαλώσει η συλλογή ώστε να υπάρχει μεγαλύτερη ποικιλία. Δηλαδή θα γίνει προσπάθεια συλλογής περισσοτέρων lyrics από διάφορες σελίδες όπως το Wikipedia ή από διαφορά social media όπως το twitter ή το reddit.

Περιγραφή σχεδιασμού:

Το σύστημα θα έχει την δυνατότητα να δέχεται input από τον χρήστη λέξεις κλειδιά ή ακόμα και πολλαπλές λέξεις – φράσεις τις οποίες θα ψάχνει στην συλλογή και θα επιστρέφει τα αρχεία στα οποία εμφανίζονται οι λέξεις ή οι φράσεις αλλά οι ίδιες θα εμφανίζονται τονισμένες και τα αρχεία θα είναι ταξινομημένα ανάλογα με την συχνότητα εμφάνισης των παραπάνω. Συγκεκριμένα αν ο χρήστης έχει δώσει ως input μια φραση το σύστημα θα εμφανίζει πρώτα τα αρχεία που περιέχουν αυτήν την φράση και έπειτα θα εμφανίζει τα αρχεία όπου περιέχονται μεμονωμένες οι λέξεις που έδωσε ο χρήστης. Επίσης το σύστημα θα είναι ικανό να διατηρήσει ένα ιστορικό αναζήτησης ώστε να δίνει προτεινόμενα αποτελέσματα στον χρήστη. Τέλος θα γίνει προσπάθεια για πολύ γρήγορη εύρεση και επιστροφή των λέξεων που έδωσε ο χρήστης αλλά λόγω του μεγάλου όγκου πληροφορίας ίσως να μην είναι εφικτό ανάλογα την υλοποίηση που θα ακολουθήσω.

Ως προς την ανάλυση κειμένου δεν είναι σίγουρο ότι θα χρειαστεί κάποια προεπεξεργασία αλλά ανάλογα με την υλοποίηση που θα ακολουθήσω ίσως χρειαστεί κάποια αλλαγή(π.χ να αφαιρεθεί από όλα τα αρχεία το πεδιο με τις ακριβείς ημερομηνίες). Για να έχω την δυνατότητα να μπορώ να ψάξω στα διαφορά αρχεία για συγκεκριμένες λέξεις ή φράσεις θα πρέπει να αποθηκεύω σαν αντικείμενο κάθε γραμμή του αρχείου(η κάθε γραμμή περιέχει όλα τα χαρακτηριστικά που αναφέρθηκαν στην συλλογή εγγράφων) ώστε να έχω πρόσβαση σε όλες τις πληροφορίες του τραγουδιού. Θα δημιουργηθούν διαφορά πεδία τα οποία θα είναι ζεύγη κλειδιών – τιμών που θα περιέχουν κλειδιά όπως όνομα, ημερομηνίες, κλπ ως περιεχόμενα προς αναζήτηση. Το κάθε πεδίο θα έχει οριστεί ως προς ανάλυση, αποθήκευση ή κανένα από τα 2.

Για την αναζήτηση θα ανοίγει ένα ξεχωριστό παράθυρο(popup) οπού ο χρήστης θα έχει την δυνατότητα να εισάγει λέξεις κλειδιά και στην συνέχεια θα διαλέγει αν θέλει να εμφανίσει τους τραγουδιστές, τους τίτλους ή τα lyrics. Θα μπορεί επίσης να εισάγει πολλαπλές λέξεις και το σύστημα θα επιστρέφει τα αρχεία ταξινομημένα ανάλογα με την συχνότητα εμφάνισης ,την ακριβή εμφάνιση των λέξεων μέσα στα αρχεία και σε περίπτωση φράσεων θα επιστρέφονται και τα αρχεία όπου βρίσκεται μεμονωμένη κάθε λέξη της φράσης. Ωστόσο αν επιθυμεί να κάνει την αναζήτηση μιας λέξεις ή φράσεις γενικότερα δηλαδή

να γίνει αναζήτηση και σε τίτλους ,τραγουδιστές ,στοίχους τότε αρκεί να μην επιλέξει κανένα φίλτρο από αυτά που ανέφερα παραπάνω(με λίγα λογία κάτι σαν default option). Θα γίνει προσπάθεια ώστε το σύστημα να αναγνωρίζει κάποιους ειδικούς χαρακτήρες ή γράμματα από διαφορετικές γλώσσες όπως το é ή æ. Για να πέτυχουμε την αναζήτηση θα γίνει χρήση της μεθόδου IndexSearcher της Lucene η οποία παρέχει δυνατότητες υπολογισμού του score κάθε λέξης ώστε να μπορούμε να επιστρέψουμε τα αρχεία ταξινομημένα αλλά και της Analyzer ώστε να μπορούμε να έχουμε την δυνατότητα να αφαιρούμε stop words όπως το the,he,she,it... και να κάνουμε stemming.

Για τα αποτελέσματα θα ανοίγει επίσης ένα ξεχωριστό παράθυρο οπού θα εμφανίζει τα αποτελέσματα της αναζήτησης ταξινομημένα ανάλογα με την εμφάνιση της λέξης ή φράσης ,την ακριβή εμφάνιση των λέξεων μέσα στα αρχεία και σε περίπτωση φράσεων θα επιστρέφονται και τα αρχεία όπου βρίσκεται μεμονωμένη κάθε λέξη της φράσης. Μια βελτιωμένη εκδοχή θα εμφάνιζε έναν συγκεκριμένο αριθμό αποτελεσμάτων στον χρήστη και πατώντας ένα βελάκι θα πήγαινε στην επόμενη σελίδα αποτελεσμάτων οπού ξανά θα εμφάνιζε έναν περιορισμένο αριθμό από αρχεία.Εδώ θα γίνει χρήση της IndexReader ώστε να μπορούμε να ανακτήσουμε τα αρχεία αυτά στα οποία εμφανίζονται οι λέξεις που αναζητά ο χρήστης

Μερικά από τα έγγραφα της συλλογής:

Τα έγγραφα είναι διαθέσιμα στο GitHub