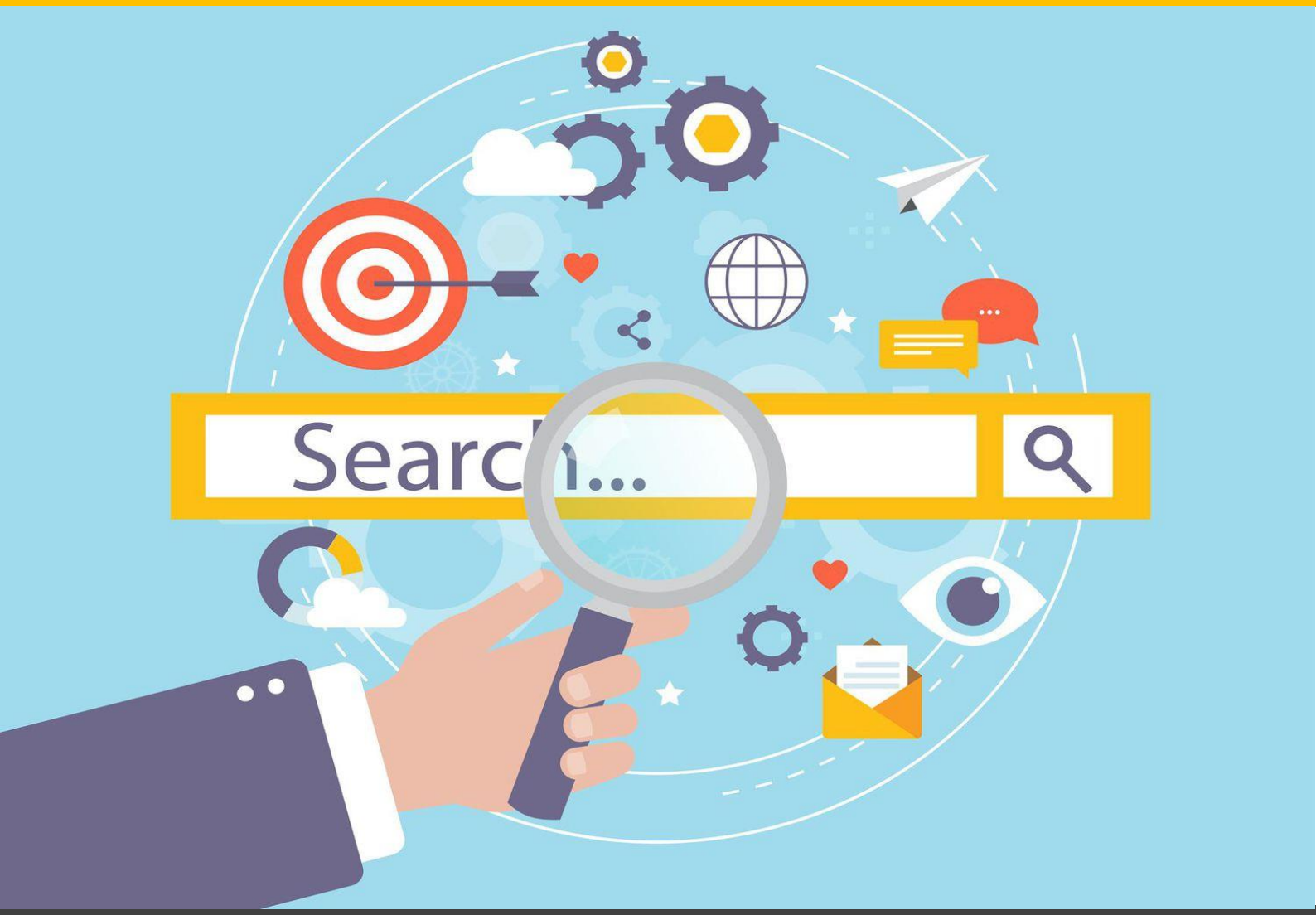
2022-2023 | Ανάκτηση Πληροφορίας



Φάση 2

GitHub repository

Μηχανή αναζήτησης τραγουδιών

11/05/2023

Φοιτητές:

Κρεββατάς Αθανάσιος, ΑΜ: 4958
Κουτσίδης Κωνσταντίνος, ΑΜ: 5114

Στόχος και λειτουργικότητα

Στόχος του συστήματος είναι η αναζήτηση τραγουδιών με βάση κάποιες λέξεις – κλειδιά που θα δίνει ο χρήστης. Η αναζήτηση αυτή μπορεί να γίνει σε διάφορα σημεία με βάση τις επιλογές του χρήστη. Πιο συγκεκριμένα, η αναζήτηση μπορεί να γίνει είτε με βάση τον τίτλο των τραγουδιών ,το νούμερο του τραγουδιού, το όνομα του καλλιτέχνη, το άλμπουμ, την χρονιά κυκλοφορίας ή την ακριβή ημερομηνία και τους στίχους του εκάστοτε τραγουδιού είτε συνδυασμούς των παραπάνω λέξεων-κλειδιών. Ο χρήστης πρέπει να επιλέξει ένα αρχείο της μορφής .csv . Το σύστημα αναζητά και επιστρέφει τα τραγούδια που περιέχουν τις λέξεις που δίνονται από τον χρήστη ως input. Τέλος ο χρήστης έχει την δυνατότητα να δει το ιστορικό των αναζητήσεων του καθώς και να ομαδοποιήσει τα αποτελέσματα της αναζήτησης όχι με βάση την συνάφεια αυτή την φορά αλλά με βάση την χρονολογία ή τον αριθμό του τραγουδιού.

Συλλογή εγγράφων

Η συλλογή των δεδομένων έγινε μέσω της πλατφόρμας Kaggle, η οποία είναι μια διαδικτυακή κοινότητα επιστημόνων που ασχολούνται με τα δεδομένα και τη μηχανική μάθηση. Μέσα από την πλατφόρμα αυτή βρήκαμε πλήθος αρχείων (.csv) με τα τραγούδια διάφορων καλλιτεχνών. Συγκεντρώσαμε τραγούδια από διαφορετικούς καλλιτέχνες ώστε να έχουμε μια ποικίλη συλλογή εγγράφων και να μην περιοριστούμε σε έναν μόνο καλλιτέχνη .Τα δεδομένα-τραγούδια που συλλέξαμε τα αποθηκεύσαμε σε ένα αρχείο τύπου .csv (songs.csv) .Το αρχείο songs.csv αποτελείται από 700 τραγούδια και στην πρώτη γραμμή δείχνει τον τρόπο με τον οποίο είναι χωρισμένη η πληροφορία. Συγκεκριμένα στην πρώτη στήλη κάθε γραμμής είναι το νούμερο του τραγουδιού, στην δεύτερη το όνομα του καλλιτέχνη (τραγουδιστή), στην τρίτη ο τίτλος του τραγουδιού, στην τέταρτη το όνομα του άλμπουμ, στην πέμπτη το έτος κυκλοφορίας του τραγουδιού, στην έκτη η ακριβής ημερομηνία κυκλοφορίας και τέλος στην έβδομη οι στίχοι. Παρακάτω παραθέτεται ένα παράδειγμα της δομής του αρχείου όπως εξηγήθηκε παραπάνω:

Num	Artist	Title	Album	Year	Date	Lyric
1	Rihanna	Love on the Brain	ANTI	2016	2016-01-28	and you got me like oh what …
2	Taylor Swift	Lover	Lover	2019	2019-08-16	we could leave the Christmas lights up…
3	Cardi B	Be Careful	Invasion of Privacy	2018	2018-03-30	yeah care for me care for me …

Ανάλυση κειμένου και κατασκευή ευρετηρίου

Για την κατασκευή του ευρετηρίου έχει γίνει μια προ επεξεργασία στο αρχείο μας. Τα αρχεία έχουν μια συγκεκριμένη δομή. Αρχικά, τα τραγούδια που έχουν επιλεχθεί είναι όλα στα αγγλικά. Αναλυτικότερα, στην πρώτη γραμμή κάθε αρχείου υπάρχουν τα πεδία (νούμερο τραγουδιού, όνομα καλλιτέχνη, τίτλος τραγουδιού, όνομα άλμπουμ, έτος κυκλοφορίας, ακριβή ημερομηνία κυκλοφορίας, στίχοι) χωρισμένα με κόμμα. Κάθε επόμενη γραμμή του αρχείου αντιστοιχεί σε ένα τραγούδι, το οποίο είναι χωρισμένο με κόμμα με βάση τα πεδία που αναφέραμε παραπάνω. Επιπλέον, επισημαίνεται ότι στο πεδίο των στίχων έχουν αφαιρεθεί όλα τα σημεία στίξης (τελεία, κόμμα, θαυμαστικό κλπ.) με στόχο να διευκολυνθεί η επεξεργασία του εγγράφου. Τα πεδία όπως προ αναφέρθηκαν είναι τα εξής: Num, Artist, Title, Album, Year, Date, Lyric . Σε όλα τα πεδία ,τα κεφάλαια γράμματα μετατρέπονται σε μικρά. Επιπλέον από το πεδίο των στίχων(Lyric) έχουν αφαιρεθεί τα stop words("a", "an", "and", "are", "as", "at", "be", "by", "for", "from", "has", "he", "in", "is", "it", "its", "of", "on", "that", "the", "to", "was", "were", "will", "with") χρησιμοποιώντας το StopAnalyze της Lucene, τα οποία εμφανίζονται πολύ συχνά στην συλλογή μας ενώ επιπλέον έχει εφαρμοστεί και stemmer στο πεδίο αυτό και στο πεδίο του τίτλου(Title) . Πιο συγκεκριμένα έχει εφαρμοστεί ο porter stemmer(PorterStemFilter) ο οποίος είναι ένας από του πιο δημοφιλείς αλγόριθμους stemming για τα Αγγλικά. Όσον αφορά τα πεδία του τίτλου(Title) , του άλμπουμ(Album) και του καλλιτέχνη(Artist), δεν έχουν αφαιρεθεί τα stop word ώστε ο χρήστης να έχει την δυνατότητα να αναζητήσει τραγούδια τα οποία πιθανότατα περιλαμβάνουν στο τίτλο τους κάποιο stop word. Ακόμα στα πεδία του καλλιτέχνη (Artist) και του άλμπουμ (Album) δεν έχει γίνει stemming. Τέλος στα υπόλοιπα πεδία(Num, Year, Date) δεν έχει γίνει κάποια ειδική προ επεξεργασία. Τα ευρετήρια που δημιουργήσαμε περιλαμβάνουν όλα τα πεδία. Δηλαδή δημιουργήσαμε τα εξής ευρετήρια(addDocument ,StandardAnalyzer):

- ένα ευρετήριο που θα περιλαμβάνει τους τίτλους των τραγουδιών
- ένα ευρετήριο που θα περιλαμβάνει τους καλλιτέχνες
- ένα ευρετήριο που θα περιλαμβάνει τα ονόματα των άλμπουμ
- ένα ευρετήριο που θα περιλαμβάνει τα έτη κυκλοφορίας των τραγουδιών
- ένα ευρετήριο που θα περιλαμβάνει την ακριβή ημερομηνία κυκλοφορίας
- ένα ευρετήριο που θα περιλαμβάνει τις λέξεις που περιέχονται στους στίχους

Αναζήτηση

Προσπαθήσαμε να δώσουμε την δυνατότητα στο χρήστη να πραγματοποιεί αναζήτηση με διάφορους τρόπους επιτρέποντας τον να αναζητεί τραγούδια με βάση τα πεδία που αναφέραμε παραπάνω αλλά και συνδυασμούς αυτών. Αναλυτικότερα δημιουργήσαμε ένα interface μέσω του οποίου αρχικά ο χρήστης θα πρέπει να φορτώσει ένα αρχείο της μορφής .csv (file -> open) π.χ σαν το songs.csv που έχει δοθεί στο κατάλογο files. Στην συνέχεια ο χρήστης έχει την δυνατότητα να προσθέσει κάποια πεδία fields π.χ Title,Artist,Year χωρισμένα με κόμμα και κάποιες τιμές (λέξεις-κλειδιά) π.χ Let,ariana,2015 και πατώντας το κουμπί search θα του εμφανιστούν bold τα αποτελέσματα στα οποία εμφανίζονται αυτά που έχει επιλέξει να αναζητήσει κάθε φορά. Ουσιαστικά παρέχουμε την δυνατότητα στον χρήστη να αναζητεί ταυτόχρονα περισσότερα από ένα πεδία κάθε φορά. Ακόμα κατά την αναζήτηση ο χρήστης μπορεί να γράψει όπως εκείνος θέλει τα πεδία και τις λέξεις-κλειδιά χωρίς να επηρεάζεται το αποτέλεσμα της αναζήτησης. Για παράδειγμα τα πεδία fields μπορούν να γραφτούν έτσι : TiTLE,ARtist,YEAR και οι λέξεις-κλειδιά έτσι : LeT,arIaNa,2015 χωρίς να αλλάζουν τα

αποτελέσματα της αναζήτησης. Επίσης παρέχουμε και ένα ιστορικό αναζητήσεων ώστε ο χρήστης να έχει την δυνατότητα να ανατρέξει σε προηγούμενες αναζητήσεις ή να μπορεί να ελέγξει και να διορθώσει τυχόν ορθογραφικά λάθη κατά την αναζήτηση που πραγματοποίησε. Σε περίπτωση που ο χρήστης δεν συμπληρώσει κάποιο πεδίο αναζήτησης(fields , value) η αναζήτηση θα ολοκληρωθεί κανονικά χωρίς κάποιο πρόβλημα. Απλά αναμενόμενα δεν θα εμφανιστεί τίποτα. Ο χρήστης έχει την δυνατότητα να αλλάξει το ήδη επιλεγμένο αρχείο και να πραγματοποιήσει αναζήτηση σε κάποιο άλλο.

Παρουσίαση Αποτελεσμάτων

Το σύστημα παρουσιάζει τα αποτελέσματα της αναζήτησης σε ένα απλό γραφικό περιβάλλον σε διάταξη με βάση τη συνάφεια τους. Πιο συγκεκριμένα θα τοποθετεί τα πεδία (fields) και τις λέξεις-κλειδιά(values) στα ειδικά κουτάκια . Μόλις έχει γράψει τα πεδία και τις τιμές που θέλει να αναζητήσει αρκεί να πατήσει το κουμπί αναζήτησης και εμείς θα κάνουμε όλη την δουλειά για αυτόν κυρίως η Lucene. Τα αποτελέσματα της αναζήτησης θα εμφανιστούν στην οθόνη χωρισμένα σε στήλες σαν exel ανάλογα με τα πεδία τους. Με βάση τις λέξεις-κλειδιά που έχει δώσει, θα τονιστούν(bold) αυτές οι λέξεις στα επιλεγμένα από τον χρήστη πεδία. Τα αποτελέσματα εμφανίζονται με βάση την συνάφεια τους δηλαδή το score όπως αυτό το καθορίζει η βιβλιοθήκη της Lucene(scoreDocs) . Εμείς όμως παρέχουμε επιπλέον την δυνατότητα στον χρήστη να ομαδοποιήσει τα αποτελέσματα όχι μόνο με την συνάφεια αλλά και με βάση την χρονολογία ή τον αριθμό του τραγουδιού. Τέλος τα αποτελέσματα παρουσιάζονται ανα δέκα. Επομένως αν τα αποτελέσματα της αναζήτησης ξεπερνούν τα δέκα τότε ο χρήστης πατώντας το κουμπί Next μπορεί να μεταφερθεί στην επόμενη σελίδα αποτελεσμάτων ενώ ταυτόχρονα μπορεί να επιστρέψει και στην προηγουμένη.