Ανάκτηση Πληροφορίας 2022

Εργασία: Μηχανή αναζήτησης ταινιών

Ομάδα

Φραγκαθούλας Χρήστος ΑΜ 4196

Λιάτσος Νικόλαος ΑΜ 4101

Συλλογή των άρθρων και προ-επεξεργασία

Η συλλογή των άρθρων που χρησιμοποιήθηκαν για την εκπόνηση της άσκησης έχουν παρθεί από την ιστοσελίδα https://www.kaggle.com/datasets/shivamb/netflix-shows και βρίσκονται με την μορφή csv. Αυτήν είναι μια έτοιμη συλλογή με 8800 ταινίες και σειρές του Netflix και η επιλογή αυτής έγινε διότι, πέρα από το ποιες ταινίες και σειρές περιέχει, αναφέρει και χαρακτηριστικά αυτών, όπως είναι ο τύπος, ο σκηνοθέτης, η χρονιά που προβλήθηκε, η διάρκεια, οι ηθοποιοί, η ηλικία που προσδιορίζεται, ο τίτλος, η χώρα των γυρισμάτων, το είδος και μια μικρή περιγραφή.

Αυτά είναι όλα τα πεδία του αρχείου και για την εκπόνηση της άσκησης έχουμε κρατήσει μερικά από αυτά τα οποία είναι, title(τίτλος), director(σκηνοθέτης), release_year(χρονιά που προβλήθηκε), duration(διάρκεια), listed_in(σε τι είδος κατατάσσεται πχ κωμωδία) και description(περιγραφή).

Το φόρτωμα του αρχείου csv με τις ταινίες γίνεται μέσω της συνάρτησης csvLoader(), μίας απλής συνάρτησης για διάβασμα και χωρισμό των σειρών με βάσει το κόμμα(,) για να γίνει διαχωρισμός μεταξύ των πεδίων που αναφέρθηκαν προηγουμένως.

Η κάθε ταινία είναι ένα document στην Lucene. Για να γίνει σωστά ο διαχωρισμός σε πεδία, πρώτα μετατρέψαμε όλα τα κόμματα του αρχείου σε κάτω παύλα(_). Έτσι δίνουμε αυτό το νέο αρχείο στην συνάρτηση csvLoader() και

έπειτα με την συνάρτηση restorelines() μετατρέπουμε τις κάτω παύλες σε κόμματα. Τελικά, έχουμε χωρίσει το αρχείο στα πεδία που χρειαζόμαστε και χωρίς να χαλάσουμε την αρχική μορφή του με τα κόμματα.

Κατά τη διαδικασία του φορτώματος είχε μερικά λάθη, συγκεκριμένα στις γραμμές 8204 και 8422, στις οποίες το κείμενο που αναγραφόταν δεν ήταν σε μια γραμμή και την μεταφέραμε από πάνω ώστε να αποτελεί μια συνέχεια και συνάφεια με το προηγούμενο.

Τέλος, για την προ-επεξεργασία παρουσιάζονται παρακάτω σε εικόνες η διαδικασία που έγινε με το εργαλείο jupyter

Ευρετηριοποίηση

Η μονάδα εγγράφου έχει μια ταινία και αποτελείται από τα εξής πεδία: title,director, release_year, duration, type και description. Από αυτά τα πεδία στο document για την ευρετηριοποίηση έχουμε προσθέσει μόνο title, release_year, type, και description. Κάθε έγγραφο που δημιουργείται, τοποθετείται σε ένα directory με τα πεδία που αναφέρθηκαν παραπάνω, ώστε στην συνέχεια να χρησιμοποιηθεί για την αναζήτηση που θα γίνεται από τον χρήστη. Την δημιουργία του directory την αναλαμβάνει η συνάρτηση createIndex() και ο analyzer ορίζεται ως ο StandardAnalyzer(). Το directory είναι τύπου FSDirectory, δημιουργείται αν δεν υπάρχει από πριν και αποθηκεύεται ώστε να μπορεί να ξαναχρησιμοποιηθεί, χωρίς να γίνει από την αρχή η δημιουργία του, με αποτέλεσμα το γρηγορότερο φόρτωμα την εφαρμογής κατά το άνοιγμα.

Την τοποθέτηση των documents στο directory την αναλαμβάνει η συνάρτηση initalizeDocuments() και σε συνδυασμό με την συνάρτηση createDocument(), η οποία είναι υπεύθυνη για την τοποθέτηση των πεδίων, σχηματίζουν τελικά το ολοκληρωμένο ευρετήριο που θα χρησιμοποιηθεί για την αναζήτηση των ταινιών.

Αναζήτηση

Η αναζήτηση των ταινιών από την εφαρμογή γίνεται με ερωτήσεις και κάθε ερώτηση ανάγεται σε διαφορετικό τρόπο αναζήτησης. Πιο συγκεκριμένα ο χρήστης μπορεί να κάνει αναζήτηση με βάση τα πεδία που έχουμε ορίσει παραπάνω δηλαδή, title, release_year, description, type, ή και χωρίς κάποιο συγκεκριμένο πεδίο, ελεύθερα σε όλα τις ταινίες. Για την επίτευξη των ερωτημάτων έχει γίνει χρήση του QueryParser, ο οποίος παίρνει το ερώτημα του χρήστη και το αντίστοιχο πεδίο στο οποίο θα γίνει η αναζήτηση. Για την αναζήτηση χωρίς συγκεκριμένο πεδίο, γίνεται αναζήτηση σε όλα τα πεδία με βάση την ερώτηση του χρήστη

Όλη αυτήν η διαδικασία των ερωτήσεων και των επιλογών που έχει ο χρήστης γίνεται μέσω ενός απλού γραφικού περιβάλλοντος (gui), το οποίο έχει δημιουργηθεί με την Swing. Σε αυτό το γραφικό περιβάλλον υπάρχουν τα κουμπιά για τα πεδία της ερώτησης Title, Release Year, Description και Type, ένα πεδίο(search box) για την ερώτηση του χρήστη και το κουμπί search. Για την επίτευξη της αναζήτησης έχει δημιουργηθεί η συνάρτηση Search() και οι αντίστοιχες συναρτήσεις SearchByTitle(), SearchByReleaseYear(), SearchByDescription(). Στις συναρτήσεις SearchByX(όπου X είναι το πεδίο όπως αναφέρεται παραπάνω) έχει ανατεθεί σε έναν πίνακα, ο αριθμός του κουμπιού, αν είναι πατημένο ή όχι και το χρώμα του στην αντίστοιχη κατάσταση.

Ο χρήστης κατά την διαδικασία της ερώτησης μπορεί να πατήσει τα κουμπιά με τα αντίστοιχα πεδία που θέλει να αναζητήσει. Για να γνωρίζει το σύστημα ποιο ή ποια πεδία πάτησε ο χρήστης, τοποθετούμε σε ένα arraylist getIndexes() το όνομα του κουμπιού/πεδίου, ώστε αυτό να διαβάζεται πριν γίνει το search. Για να γίνει η αναζήτηση σε κάποιο/α πεδίο/α, ο χρήστης γράφει την ερώτηση του, έπειτα πατάει το κουμπί με το/α πεδίο/α που επιθυμεί και στην συνέχεια το κουμπί search ή το Enter.

Το κουμπί του πεδίου type είναι τύπου dropdown menu, όπου δίνονται οι επιλογές στον χρήστη να επιλέξει το είδος των ταινιών για αναζήτηση αποτελεσμάτων και μεταξύ αυτών είναι τα εξής είδη: Drama, Comedies, Sprots, Documentaries, Thrillers, Fantasy, Romantic, Independent Movies και Anime Features. Για να γνωρίζει το σύστημα ποιο από τα διαθέσιμα είδη έχει επιλέξει ο χρήστης, αποθηκεύουμε σε ένα πίνακα με βάση το index το είδος. Μπορεί να επιλέξει ένα είδος την φορά χωρίς να χρειαστεί να γράψει κάτι στην περιοχή αναζήτησης και εμφανίζονται οι αντίστοιχες ταινίες. Τα υπόλοιπα κουμπιά που απομένουν είναι απλά κουμπιά, τα οποία τα πατάει ο χρήστης ανάλογα με την επιθυμία του για αναζήτηση σε κάποιο πεδίο. Επιπρόσθετα, υπάρχει ένα κουμπί δίπλα και αριστερά από το search box το οποίο έχει την ιδιότητα να καθαρίζει την περιοχή αναζήτησης.

Επίσης, με κάθε νέα αναζήτηση που εκτελεί ο χρήστης η περιοχή των αποτελεσμάτων διαγράφεται, αν υπάρχουν αποτελέσματα από αναζητήσεις που ενδεχομένως να έκανε πριν.

Παρακάτω αναλύονται οι διάφοροι τρόποι αναζήτησης όταν ο χρήστης:

- Δεν επιλέξει κάποιο πεδίο για αναζήτηση και έχει τοποθετήσει την ερώτηση στο πεδίο αναζήτησης, τότε το σύστημα δημιουργεί 4 ερωτήσεις τύπου QueryParser για κάθε πεδίο και εκτελεί αναζήτηση σε όλα τα πεδία
- Επιλέξει ένα πεδίο για αναζήτηση. Αν αυτό είναι το type, δεν χρειάζεται να πληκτρολογήσει κάτι στην περιοχή αναζήτησης, μιας και το σύστημα αναζητεί της ταινίες με βάση τον τύπο που καταχωρήθηκε. Αν είναι κάποιο άλλο πεδίο, τότε δίνει την ερώτηση που επιθυμεί μαζί με το αντίστοιχο πεδίο που θέλει να αναζητήσει.
- Έχει επιλέξει δύο ή περισσότερα πεδία για αναζήτηση

  i)Αν έχει πατηθεί το κουμπί type: Θα αναζητήσει πρώτα τις ταινίες με βάσει το συγκεκριμένο είδος και στην συνέχεια με βάση τα πεδία που έχει επιλέξει, εκτελείται η τομή τους

  και παρουσιάζονται τα πρώτα K αποτελέσματα με βάση το score

  ii)Αν έχει πατηθεί το κουμπί Release year: Το ίδιο με πριν, απλά πρώτα θα αναζητεί τις ταινίες στην χρονολογία που έχει γράψει στην ερώτηση κοκ

Παρουσίαση αποτελεσμάτων

Για την εμφάνιση των αποτελεσμάτων στην περιοχή εμφάνισής τους έχει επιτευχθεί με την χρήση της συνάρτησης TopDocs() της Lucene , η οποία επιστρέφει με βάσει το score, τα έγγραφα εκείνα που είναι πιο συναφή με την ερώτηση που έχει γίνει προηγουμένως.

Εμφάνιση αποτελεσμάτων ανάλογα με τις περιπτώσεις αναζήτησης:

- Δεν επιλεχθεί κάποιο πεδίο για αναζήτηση, τότε παρουσιάζει τα πρώτα 5 title, 3 description, 3 description, 2 release year, πιο συναφή έγγραφα. Αν τα αποτελέσματα είναι λιγότερα από τον ελάχιστο αριθμό που δόθηκε για προβολή τότε εμφανίζονται μόνο αυτά. Αυτήν η επιλογή έγινε για να δοθεί ένα «βάρος», στα πεδία και ο αριθμός των εμφανιζόμενων αποτελεσμάτων δείχνει το «βάρος» που επιλέχθηκε.
- Στην περίπτωση επιλογής ενός μόνου πεδίου η εμφάνιση των αποτελεσμάτων είναι με μέγιστο όριο τα 10,αν αυτά υπάρχουν, αλλιώς εμφανίζονται όσα αποτελέσματα ικανοποιούν την ερώτηση.
- Τελευταία περίπτωση είναι αυτήν όπου ο χρήστης επιλέγει δύο ή παραπάνω πεδία για αναζήτηση η εμφάνιση γίνεται ανάλογα με το πόσες ταινίες βρέθηκαν στην τομή των ερωτημάτων.
- Σε όλες αυτές τις διαφορετικές περιπτώσεις, η απεικόνιση των αποτελεσμάτων γίνεται με έναν συγκεκριμένο τρόπο. Για κάθε ταινία που εμφανίζεται στο αποτέλεσμα, υπάρχει ο τίτλος της και από κάτω βρίσκεται η σύντομη περιγραφή της, έτσι όπως υπάρχει στο αρχείο με τις ταινίες.

Παρατηρήσεις

1) Στο GitHub θα δοθεί και το αρχείο με τις ταινίες και θα πρέπει να γίνει αλλαγή των paths στον κώδικα για το διάβασμα του αρχείου καθώς και η δημιουργία ενός path για το directory
2) Εγκατάσταση Swing package