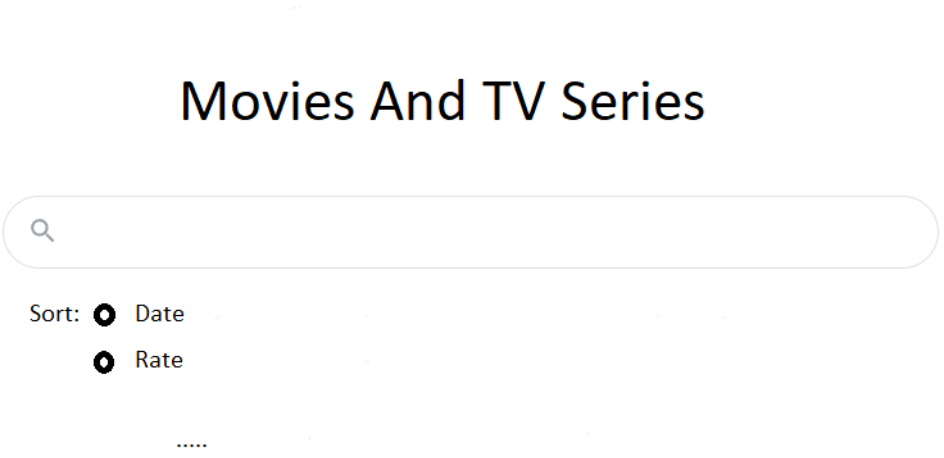
Εργασία: Μηχανή αναζήτησης ταινιών

Φάση 1



ΟΜΑΔΑ:
Φίλιππο Πρίφτης 4162
Αγκνές-Μοναλίσα Τουκαλιούκ 3346

Συλλογή Εγγράφων

Η συλλογή των δεδομένων που θα χρησιμοποιήσουμε στην εργασία είναι μία έτοιμη συλλογή που υπάρχει στο kaggle(https://www.kaggle.com/datasets/shivamb/netflix-shows) και περιέχει δεδομένα για shows του netflix σε μορφή csv. Συγκεκριμένα η συλλογή περιέχει 12 πεδία μερικά από τα οποία είναι τα εξής: showid(μοναδικό id για κάθε show), type(αν είναι ταινία ή σειρά), το σκηνοθέτη, το cast, την χώρα στην οποία γυρίστηκε, το release year, την διάρκεια, το είδος και την περιγραφή της.

Περιγραφή σχεδιασμού του συστήματος

- Στόχος του συστήματος: Ο στόχος του συστήματος(της μηχανής αναζήτησης) που θα δημιουργήσουμε είναι δοθείσης μιας ερώτησης(απο τον χρήστη) το σύστημα να επιστρέφει και να παρουσιάζει στο χρήστη τα έγγραφα τα οποία είναι σχετικά με την ερώτηση αυτή.

- Ανάλυση κειμένου και κατασκευή ευρετηρίου: Όπως αναφέρθηκε παραπάνω η συλλογή δεδομένων που θα χρησιμοποιηθεί είναι της μορφής csv και κάθε γραμμή (εκτός της πρώτης που περιγράφει τα πεδία) αποτελεί ένα έγγραφο το οποίο είναι και η μονάδα εγγράφου. Από τα 12 συνολικά πεδία της συλλογής θα κρατηθούν τα 5. Αυτά είναι τα title, director, release_year, listed_in, descripiton. Για την ανάλυση των εγγράφων και την μετατροπή των λέξεων τους σε tokens θα χρησιμοποιηθεί ο *StandardAnalyzer* ο οποίος αφαιρεί τα stopwords (the, a, an κτλ) και μετατρέπει τις λέξεις σε lowercase. Όσον αφορά τα ευρετήρια (Directory) τα οποία θα δημιουργήσουμε για να υποστηρίζει το σύστημα διάφορους τύπους ερωτήσεων θα είναι όλα της κλάσης FSDirectory προκειμένου να αποθηκεύονται στο δίσκο. Η γενική ιδέα είναι να δημιουργήσουμε ένα ευρετήριο πεδίου για κάθε πεδίο για να μπορεί το σύστημα να απαντήσει σε ερωτήσεις με βάση κάποιο πεδίο (π.χ. αναζήτηση για ταινίες που οι όροι του ερωτήματος εμφανίζονται στο πεδίο του τίτλου) και ένα ενιαίο ευρετήριο για κάθε έγγραφο. Για να γράψουμε στα ευρετήρια που θα δημιουργήσουμε θα χρησιμοποιήσουμε τους *IndexWriter* της lucene τους οποίους και θα συνδέσουμε με τον analyzer μέσω της κλάσης *IndexWriterConfig.* Αρχικά πρέπει να δημιουργήσουμε ένα document προσθέτοντας του πεδία τα οποία θα είναι stored ή οχι. Στη συνέχεια χρησιμοποιούμε τη μέθοδο *addDocument()* του *IndexWriter* για να προσθέσουμε το Document που δημιουργήσαμε στο ευρετήριο. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα document.

- Αναζήτηση: Αρχικά για να διαβάσουμε τα ευρετήρια θα χρειαστούμε έναν DirectoryReader για κάθε Directory. Στη συνέχεια θα περάσουμε κάθε DirectoryReader σε έναν IndexSearcher. Ο IndexSearcher θα χρησιμοποιηθεί για να κάνουμε αναζήτηση σε κάθε ευρετήριο.Για να κάνουμε αναζήτηση με τον IndexSearcher θα χρειαστεί να δημιουργήσουμε Query. Το Query προκύπτει από το περιεχόμενο του TextBox του γραφικού περιβάλλοντος στο οποίο ο χρήστης θα εισάγει το ερώτημα του. Συγκεκριμένα το περιεχόμενο του TextBox θα γίνει parse και θα “σπάσει” σε tokens απο τον QueryParser. Το αποτέλεσμα του QueryParser ειναι το επιθυμητό Query που χρειάζεται ο IndexSearcher το οποίο και περνάμε ως όρισμα σε αυτόν. Ακολούθως καλούμε την μέθοδο search του IndexSearcher . Όσον αφορά την αναζήτηση με βάση πεδίο ο χρήστης θα έχει την δυνατότητα να επιλέξει τα είδη των ερωτημάτων(να κάνει δηλαδή αναζήτηση με βάση πεδίου) επιλέγοντας ένα ή περισσότερα από τα CheckBox που θα αναπαριστούν πεδία μας και θα βρίσκονται κάτω από το TextBox.

- Παρουσίαση Αποτελεσμάτων: Με το πέρας της αναζήτησης θα εμφανίζονται στο χρήστη μέσω του γραφικού περιβάλλοντος τα 10 πιο συναφή αποτελέσματα,με βάση query, το ένα κάτω από το άλλο. Στα αποτελέσματα θα φαίνεται υπογραμμισμένος ο τίτλος των συναφών ταινιών. Στη συνέχεια ο χρήστης θα μπορεί να κάνει κλικ σε όποιο τίτλο της ταινίας επιθυμεί και θα εμφανίζονται τα πεδία της μαζί τις τιμές τους στο γραφικό περιβάλλον. Για να μπορέσει να δει τα υπόλοιπα αποτελέσματα μπορεί να πατήσει τα βελάκια (<-, ->) προηγούμενο, επόμενο.Τέλος ο χρήστης θα έχει την δυνατότητα να ταξινομήσει τα αποτελέσματα με βάση κάποιο πεδίο(πχ Release_year).