ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ 2021-2022

ΕΡΓΑΣΙΑ: Μηχανή αναζήτησης ταινιών

ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ:27/5/2022

ΟΜΑΔΑ

Ονοματεπώνυμα:

1. ΝΕΚΤΑΡΙΟΣ ΒΙΔΑΛΑΚΗΣ, Α.Μ. 4033
2. ΚΩΝΣΤΑΝΤΙΝΟΣ ΤΡΙΤΣΩΝΗΣ Α.Μ. 4185

Πίνακας περιεχομένων

Η Περιγραφή του dataset

Στην παρούσα εργασία ο τρόπος υλοποίησης της, που χρησιμοποιήθηκε για την διεξαγωγής της, ήταν μια εκ των έτοιμων συλλογών που παρουσιάστηκαν κατά την διάρκεια διεξαγωγής των μαθημάτων. Πιο συγκεκριμένα επιλέξαμε να εκμεταλλευτούμε την συλλογή tmdb Top 10000 Popular movies dataset.

(https://www.kaggle.com/datasets/sankha1998/tmdb-top-10000-popular-movies-dataset).

Αρχικά, το **format** του **dataset** είναι τύπου **csv** και αυτό αποτελείται από 6 πεδία:

- Το 1ο πεδίο αφορά την αρίθμηση των ταινιών.
- Το 2ο πεδίο αφορά τον τίτλο της εκάστοτε ταινίας.
- Το 3ο πεδίο αφορά μία σύντομη περιγραφή της ταινίας.
- Το 4ο πεδίο αφορά την γλώσσα ομιλίας της ταινίας.
- Το 5ο πεδίο αφορά το συνολικό αριθμών αξιολογήσεων της ταινίας.
- Το 6ο πεδίο αφορά το μέσο όρο των αξιολογήσεων.

Η Λειτουργικότητα Συστήματος

Η συγκεκριμένη εργασία έχει ως στόχο την υλοποίησή ενός συστήματος που αφορά την αναζήτησης ταινιών ή κριτικών για ταινίες. Στην συνέχεια, ακολουθεί μια συνοπτική αναφορά που παρουσιάζει την λειτουργικότητα του παρόντος συστήματος στα πεδία που θα επισημανθούν παρακάτω:

- **1 Πεδίο: Η Ανάλυση κειμένου και η κατασκευή ευρετηρίου.**
  Σε αυτό το στάδιο ξεκινά η Προ επεξεργασία των δεδομένων και η δημιουργία του εγγράφου και των ευρετηρίων που θα χρησιμοποιηθούν για την αναζήτηση. Αξίζει να σημειωθεί ότι κατά την διάρκεια της αναζήτησης μπορούν να χρησιμοποιηθούν διάφορες λειτουργίες , όπως για παράδειγμα είναι η διόρθωση τυπογραφικών λαθών, η επέκταση ακρωνύμων με σκοπό τα βέλτιστα αποτελέσματα.
- **2 Πεδίο: Η Αναζήτηση λέξεων.**
  Σε αυτό το σημείο παρέχετε στον χρήστη η δυνατότητα να αναζητήσει μέσα στο dataset. Η αναζήτηση αυτή γίνεται χρησιμοποιώντας λέξεις κλειδιά.
- **3 Πεδίο: Η Αναζήτηση σε πεδίο.**
  Στο επόμενο στάδιο ο χρήστης έχει την επιλογή να αναζητήσει σε ποια από τα πεδία του αρχείου να πραγματοποιηθεί η αναζήτηση (πχ. τίτλος, πλοκή).
- **4 Πεδίο: Το Ιστορικό αναζητήσεων.**
  Στη συνέχεια, η επόμενη φάση αναφέρεται στο σύστημα όπου διατηρεί ένα ιστορικό αναζητήσεων ώστε να έχει την δυνατότητα να μπορεί να προτείνει εναλλακτικά ερωτήματα στον χρήστη.
- **5 Πεδίο: Η Παρουσίαση των Αποτελεσμάτων.**
  Το τελευταίο στάδιο αφορά τα αποτελέσματα όπου το σύστημα τα παρουσιάζει σε διάταξη με βάση τη συνάφεια τους σχετικά με το ερώτημα.

Η Συλλογή Δεδομένων

Αρχικά, τα δεδομένα της συλλογής αποτελούνται από 10.000 ταινίες. Στην συνέχεια, επιλέξαμε να κρατήσουμε τα πεδία που αποτελούνται από τον τίτλο, την πλοκή και μέσο όρο αξιολογήσεων από το σύνολο των 6 πεδίων καθώς τα υπόλοιπα από αυτά δεν είναι χρήσιμη πληροφορία για την αναζήτηση μας.

Η Προ επεξεργασία Δεδομένων

Η Διαδικασία σχετικά με την προ επεξεργασία των δεδομένων ακολουθεί τα παρακάτω στάδια. Αρχικά, διαβάζουμε το αρχείο μέσω ενός **csvReader** που δημιουργήσαμε και αποθηκεύουμε τα δεδομένα σε ένα **Arraylist**. Επιπλέον, δημιουργούμε το **Document** και προσθέτουμε τα πεδία που θέλουμε να χρησιμοποιήσουμε δηλαδή τον τίτλο, την πλοκή, τον μέσο όρο αξιολογήσεων. Τονίζουμε ότι τα πεδία που θέλουμε να γίνουν **tokenized** είναι ο τίτλος και η πλοκή, και γι' αυτό τον λόγο χρησιμοποιούμε **new TextField** ενώ για την συνολική αξιολόγηση που δεν χρειάζεται **tokenized** επιλέγουμε **new StringField**. Τέλος, προσθέτουμε τα **fields** στο εκάστοτε **document**.

Η Δημιουργία Ευρετηρίων

Η δημιουργία του ευρετηρίου ακολουθεί αφού γίνει η προ επεξεργασία. Πιο συγκεκριμένα, χρησιμοποιούμε το **StandardAnalyzer** για την ανάλυση μας καθώς αυτή θεωρείται μία πιο εκλεπτυσμένη μορφή των υπόλοιπων analyzers. Στην συνέχεια η αποθήκευση του **index** πραγματοποιείται γίνεται σε ένα ξεχωριστό αρχείο μέσα στον δίσκο μέσω του **FSDirectory.open**. Επιπρόσθετα θέτουμε **setOpenMode(Create)** στον **IndexWriterConfig** έτσι ώστε να δημιουργεί το **index**. Παράλληλα προσθέτουμε τα **Documents** στον **IndexWriter.** Μονάδα του εγγράφου αποτελεί ο τίτλος ή η πλοκή. Τέλος δημιουργούμε ευρετήρια για τα πεδία(τον τίτλο και την πλοκή).

Η Αναζήτηση

Πρώτη κίνηση στην αναζήτηση είναι να λαμβάνουμε την λέξη κλειδί ώστε να χρησιμοποιηθεί για την αναζήτηση και να διαβάζουμε το **directory** μέσω του **DirectoryReader**. Στη συνέχεια, δημιουργούμε έναν **MultiQueryParser** καθώς έχουμε 2 πεδία αναζήτησης(**title,overview**) και κάνουμε parse την λέξη κλειδί. Έπειτα κάνουμε **search** και αποθηκεύουμε τα αποτελέσματα της αναζήτησης σε ένα **TopDocs**. Επίσης εκτελούμε την εντολή **scoreDocs** για να πάρουμε τα **hits** τα οποία διατρέχουμε και δίνουμε στο σύστημα για να τα εμφανίσει κατάλληλα στον χρήστη. Επιπλέον υποστηρίζονται αναζητήσεις για συγκεκριμένα πεδία μεμονωμένα, η χρήση **wildcard(*,?)**, το **fuzzy search** σε περίπτωση λαθών**(~)** και οι λογικοί όροι (**AND**,**OR**,**NOT**). Τέλος διατηρούμε το ιστορικό αναζητήσεων σε ένα αρχείο με την δυνατότητα επαναχρησιμοποίησης κάποιας αναζήτησης αλλά δεν υλοποιήθηκε η χρήση του για να προτείνουμε εναλλακτικά ερωτήματα.

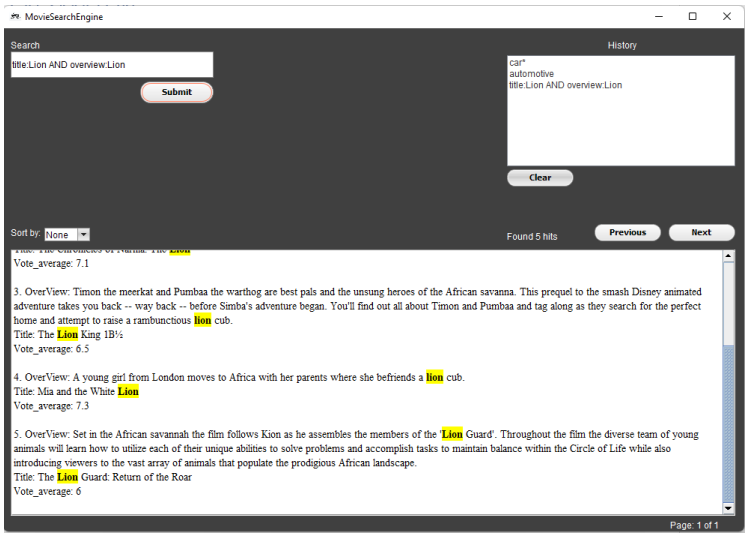
Η Παρουσίαση Αποτελεσμάτων

Όσον αφορά την παρουσίαση των αποτελεσμάτων ,το σύστημα παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους σε σχέση με το ερώτημα. Πιο συγκεκριμένα τα αποτελέσματα θα εμφανίζονται στον χρήστη ανά 10 με την δυνατότητα αλλαγής σελίδας μπρος ή πίσω ενώ ταυτόχρονα η λέξη κλειδί θα εμφανίζεται σε όλο το έγγραφο τονισμένη με την χρήση του **highlighter** το οποίο τονίζει τα αποτελέσματα σε μορφή **html**, συνεπώς όλο το κείμενο έγινε σε μορφή **html** για να μπορέσει να υποστηριχθεί η λειτουργία κατάλληλα. Ακόμη υπάρχει η δυνατότητα να μπορεί να γίνει αναδιάταξη των αποτελεσμάτων με βάση τον τίτλο ή ακόμη και την συνολική αξιολόγηση μέσω της μεθόδου **sort** της **lucene**.

Τα Ενδεικτικά Παραδείγματα Αποτελεσμάτων

Παράδειγμα 1ο

Στο παρακάτω παράδειγμα παρατηρούμε ότι δίνουμε στο πεδίο αναζήτησης την λέξη **Lion**, αλλά να εμφανιστούν τα αποτελέσματα που έχουν την λέξη κλειδί και στον τίτλο και στην πλοκή. Δεξιά βλέπουμε το ιστορικό των αναζητήσεων ενώ στο κάτω μέρος του προγράμματος βρίσκονται τα αποτελέσματα στα οποία βλέπουμε πόσα είναι τα **hits** που είχαμε και σε ποια σελίδα βρισκόμαστε και βλέπουμε την λέξη τονισμένη όπου αυτή εμφανίζεται.



Παράδειγμα 2ο

Στα παρακάτω παραδείγματα παρατηρούμε ότι έχουμε βρει για τις λέξεις **bad boys** 238 αποτελέσματα χωρισμένα ανά 10 καθώς βρισκόμαστε στην σελίδα 6 και είμαστε στο νούμερο 60. Επίσης έχουμε την να δυνατότητα να επιστρέφουμε και προς τα πίσω στα αποτελέσματα. Στην δεύτερη εικόνα έχουμε διατάξει τα δεδομένα με βάση την βαθμολογία(**sort by:Rating**) με φθίνουσα κατάταξη.