Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής
Πολυτεχνική Σχολή - Πανεπιστήμιο Ιωαννίνων

Εργασία: Μηχανή αναζήτησης ταινιών

Φάση 1: Αρχικός σχεδιασμός και συλλογή δεδομένων

Αθανάσιος Τόμπρας ΑΜ: 3345
Βασίλειος Βογιάννου ΑΜ: 3193

Github Link:
https://github.com/vasilisvog/info_retrv_prjct.git

Περιεχόμενα

Συλλογή εγγράφων

Για τη συλλογή εγγράφων θα χρησιμοποιηθεί το προτεινόμενο dataset του Kaggle με τις 10000 πιο δημοφιλείς ταινίες. Παρατίθεται ο σύνδεσμος*:* https://www.kaggle.com/datasets/sankha1998/tmdb-top-10000-popular-movies-dataset

Τα πεδία που θα ληφθούν υπ΄όψιν είναι και τα 6 πεδία που περιέχει το dataset για κάθε ταινία ενώ το format των δεδομένων θα είναι της μορφής .csv. Ως πεδία του dataset ορίζονται τα «index, title, overview, original language, vote count, vote average».

Πιο αναλυτικά:

- Index: Παίζει το ρόλο του αναγνωριστικού (id) της κάθε ταινίας στη λίστα των δεδομένων (Integer- Unstored)
- Title: Αποτελεί τον τίτλο της κάθε ταινίας (String - Text)
- Overview: Μία μικρή περίληψη της κάθε ταινίας (String - Text)
- Original language: Η γλώσσα στην οποία γυρίστηκε η ταινία (String - Unstored)
- Vote count: Το σύνολο των χρηστών που ψήφισαν στο imdb για την ταινία (Integer - Unstored)
- Vote average: Ο μέσος όρος των ψήφων για την κάθε ταινία (Float - Numeric)

Τέλος, παρατηρήθηκαν διπλότυποι τίτλοι και missing values στο πεδίο overview του dataset με αποτέλεσμα την εξαίρεση των συγκεκριμένων γραμμών από τα δεδομένα μας για μεγαλύτερη ευκολία στο χειρισμό τους.

Στόχος και λειτουργικότητα του συστήματος

Ο στόχος του συστήματος είναι η υλοποίηση μιας αποδοτικής μηχανής αναζήτησης αξιοποιώντας την βιβλιοθήκη Apache Lucene. Το σύστημα θα επεξεργάζεται το αρχικό CSV αρχείου και θα δημιουργεί μια βιβλιοθήκη αρχείων αναγνώσιμων από το βασικό πρόγραμμα, θα δημιουργεί Index μέσω αυτών των απλών αρχείων, και τέλος θα βρίσκει την κατάλληλη ταινία βάσει της λέξης/φράσης αναζήτησης του χρήστη. Τέλος θα υλοποιηθεί ένα φιλικό προς τον χρήστη σύστημα γραφικών GUI με σκοπό την εύκολη περιήγηση του στην εφαρμογή.

Δημιουργία των Αρχείων

Τα έγγραφα τα οποία έχουν επιλεγεί είναι συγκεντρωμένα σε ένα αρχείο CSV.
Η βιβλιοθήκη Lucene δεν αναγνωρίζει αρχεία CSV, για αυτό το λόγο θα πραγματοποιηθεί μια διαδικασία μετατροπής του πηγαίου αρχείου σε αρχείο Excel. Με τη χρήση ενός CSV Parser πραγματοποιείται η μετατροπή του CSV αρχείου σε αρχείο excel, επιτρέποντας έτσι στο σύστημα να διαβάσει και να δημιουργήσει τα Documents. Το παραγόμενο excel αρχείο θα αποτελεί το back-up του dataset.

Ανάλυση κειμένου και κατασκευή ευρετηρίου

Ο έλεγχος θα βασίζεται στον Standard Analyzer της Lucene. Με αυτόν τον τρόπο επιτυγχάνεται ο χωρισμός (split) του κειμένου στην κάθε λέξη που το αποτελεί(StandardTokenizer), αφαιρούνται σημεία στίξης και stopwords ενώ ταυτόχρονα μετατρέπει όλα τα γράμματα των παραγόμενων tokens σε πεζά.Όσον αφορά το Documentation phase της Lucene θα δημιουργηθούν πολλαπλά Documents με πεδία : “Title” , “Overview”, “Vote average” που θα αξιοποιούνται στην αναζήτησή. Το είδος των πεδίων θα είναι τύπου Text για τα δυο πρώτα και Numeric για το τελευταίο αντίστοιχα. Η δημιουργία αυτών των Documents μέσω του Standard Analyzer ,θα δημιουργεί τα Indexes με τη χρήση ενός Index Writer και θα τα τοποθετεί στο τοπικό Directory.

Αναζήτηση

Η διαδικασία αναζήτησης γίνεται μέσω ενός Index Searcher. Αξιοποιώντας τα προηγούμενα Indexes στον φάκελο που θα αποθηκευτούν , καθώς και ένα Query Parser, θα ελέγχει βάση του Standard Analyzer τις ταινίες στα Document μας, επιστρέφοντας τις δέκα πρώτες ταινίες(TopDocs) βάσει ενός score. Έχοντας τα κλειδιά των κορυφαίων Documents, θα δίνεται η δυνατότητα να εντοπιστούν οι ταινίες και να τυπωθούν στον χρήστη πληροφορίες για τα υπόλοιπα πεδία τους. Το προεπιλεγμένο πεδίο αναζήτησης θα είναι το “Title” ωστόσο θα υποστηρίζεται και η δυνατότητα αναζήτησης λέξεων/φράσεων στα υπόλοιπα πεδία ταυτόχρονα , επιλέγοντας το κατάλληλο πεδίο αναζήτησης συνοδευόμενο από τον όρο (λέξη κλειδί) που επιθυμεί ο χρήστης. Τέλος, για κάθε Document που θα αποθηκεύεται στο index θα φορτώνεται το ID του σε ένα ξεχωριστό database το οποίο θα χρησιμοποιηθεί για ιστορικό αναζήτησης του χρήστη και την πρόταση διαφορετικών ερωτημάτων .

Παρουσίαση Αποτελεσμάτων

Η μηχανή αναζήτησης θα παρέχει στον χρήστη ένα στοιχειώδες παράθυρο GUI με μερικά κουμπιά για τις διάφορες λειτουργίες του συστήματος.Το σύστημα θα παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους με το ερώτημα. Η προεπιλεγμένη διάταξη όσον αφορά τα πεδία αναζήτησης θα εμφανίζει τους δέκα πιο σχετικούς τίτλους ταινιών σύμφωνα με την αντίστοιχη λέξη/φράση κλειδί και θα υπάρχει δυνατότητα προβολής πληροφοριών των υπολοίπων πεδίων της κάθε ταινίας.
Παρατίθενται μερικές επιπλέον ιδέες υλοποίησης σχετικά με την παρουσίαση αποτελεσμάτων :

- Αναζήτηση της λέξης/φράσης που έχει γραφτεί
- Δυνατότητα επιλογής προτεινόμενων αναζητήσεων .
- Αναζήτησή του ιστορικού προηγούμενων αναζητήσεων .
- Καθάρισμα των αποτελεσμάτων η/και των Index/Documents.
- Δυνατότητα αναδιάταξης των αποτελεσμάτων με βάση το “Vote Average” των χρηστών .

Παράδειγμα του Dataset

	Unnamed: 0	title	overview	original_language	vote_count	vote_average
0	0	Ad Astra	The near future, a time when both hope and har...	en	2853	5.9
1	1	Bloodshot	After he and his wife are murdered, marine Ray...	en	1349	7.2
2	2	Bad Boys for Life	Marcus and Mike are forced to confront new thr...	en	2530	7.1
3	3	Ant-Man	Armed with the astonishing ability to shrink i...	en	13611	7.1
4	4	Percy Jackson: Sea of Monsters	In their quest to confront the ultimate evil, ...	en	3542	5.9
5	5	Birds of Prey (and the Fantabulous Emancipatio...	Harley Quinn joins forces with a singer, an as...	en	2639	7.1
6	6	Live Free or Die Hard	John McClane is back and badder than ever, and...	en	3714	6.5
7	7	Cold Blood	A legendary but retired hit man lives in peace...	fr	119	5.1
8	8	Underwater	After an earthquake destroys their underwater ...	en	584	6.5
9	9	The Platform	A mysterious place, an indescribable prison, a...	es	1924	7.2
10	10	Jumanji: The Next Level	As the gang return to Jumanji to rescue one of...	en	2974	6.8
11	11	The Twilight Saga: Eclipse	Bella once again finds herself surrounded by d...	en	5687	6.1
12	12	Sonic the Hedgehog	Based on the global blockbuster videogame fran...	en	2066	7.4
13	13	Star Wars: The Rise of Skywalker	The surviving Resistance faces the First Order...	en	3800	6.5
14	14	Onward	In a suburban fantasy world, two teenage elf b...	en	956	8.0
15	15	Emma.	In 1800s England, a well-meaning but selfish y...	en	148	7.1
16	16	Pocahontas II: Journey to a New World	When news of John Smith's death reaches Americ...	en	845	5.3
17	17	Lara Croft: Tomb Raider - The Cradle of Life	Lara Croft ventures to an underwater temple in...	en	2896	5.7
18	18	The Invisible Man	When Cecilia's abusive ex takes his own life a...	en	1249	7.2
19	19	Blood Father	An ex-con reunites with his estranged wayward ...	en	946	6.1
20	20	A Rainy Day in New York	Two young people arrive in New York to spend a...	en	783	6.6