ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

WIKIPEDIA MOVIE SEARCHER

ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

2019-2020

ΕΛΕΝΗ ΜΟΥΖΑΚΗ, 3280

ΠΑΝΑΓΙΩΤΗΣ ΠΑΠΑΪΩΑΝΝΟΥ, 3309

ΤΕΛΙΚΗ ΑΝΑΦΟΡΑ

Μάιος 2020

ΙΣΤΟΡΙΚΟ ΠΡΟΗΓΟΥΜΕΝΩΝ ΠΑΡΑΔΟΣΕΩΝ

Ημερομηνία	Περιγραφή	Συγγραφέας
25/04/2020	Σύντομη περιγραφή σχεδιασμού και συλλογής δεδομένων	Ελένη Μουζάκη Παναγιώτης Παπαϊωάννου
29/05/2020	Περιγραφή σχεδιασμού ολοκληρωμένης εργασίας	Ελένη Μουζάκη Παναγιώτης Παπαϊωάννου

1 ΠΡΟΛΟΓΟΣ

Σε αυτή την εργασία ασχοληθήκαμε με την αναζήτηση ταινιών και την ανάκτηση αποτελεσμάτων από την Wikipedia. Πιο συγκεκριμένα, ο χρήστης θα μπορεί να δίνει στο σύστημα ως είσοδο λέξεις κλειδιά. Το σύστημα με τη σειρά του, με βάση αυτές τις λέξεις αυτές θα του επιστρέφει μία λίστα αποτελεσμάτων. Με ένα κλικ από τον χρήστη, κάθε αποτέλεσμα μπορεί να τον μεταφέρει στο αντίστοιχο άρθρο της Wikipedia στο web.

Επιλέξαμε το πρόγραμμα μας να μην αποθηκεύει στο δίσκο τα άρθρα της Wikipedia, αλλά να μπορεί ο χρήστης να μεταβεί σε αυτά με τη χρήση του διαδικτύου, έτσι ώστε να παραμένει ενημερωμένο με την τελευταία ανανέωση του άρθρου. Αυτό θα το καθιστά χρήσιμο σε βάθος χρόνου.

2 ΣΥΛΛΟΓΗ ΕΓΓΡΑΦΩΝ

Για την υλοποίηση του Wikipedia Movie Searcher, συλλέξαμε άρθρα από τη Wikipedia σχετικά με ταινίες. Μέσω του Kaggle.com κατεβάσαμε το αρχείο 'wiki_movie_plots_deduped.csv' από το οποίο δημιουργήσαμε το database. Το αρχείο αυτό περιέχει πληροφορίες σχετικά με τις ταινίες (release date, title, origin/ethnicity, director, cast και genre) αλλά και το Wikipedia URL τους.

Μετά την ανάκτηση του παραπάνω αρχείου, γράψαμε ένα πρόγραμμα σε γλώσσα Pyhton, έτσι ώστε στο ίδιο αρχείο να υπάρχουν δύο επιπλέον στήλες πληροφοριών για την κάθε ταινία. Οι στήλες αυτές περιλαμβάνουν το rating value (αξιολόγηση στα 10) και rating count (αριθμός κριτικών) τους. Αυτές τις πληροφορίες τις συλλέξαμε από το IMDB.com, με web scrapping, με τη βοήθεια του Beautiful Soup και τις χρησιμοποιήσαμε για την εναλλακτική διάταξη των αποτελεσμάτων στην οποία θα αναφερθούμε στη συνέχεια.

Στο τέλος της διαδικασίας της συλλογής, η βάση δεδομένων μας αποτελούνταν από 34.873 άρθρα, δηλαδή 34.873 Wikipedia URLs.

3 ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΚΑΤΑΣΚΕΥΗ ΕΥΡΕΤΗΡΙΟΥ

Από το αρχείο csv που είχε πλέον δημιουργηθεί, με τη βοήθεια της Lucene κατασκευάσαμε ένα ανεστραμμένο ευρετήριο. Για την κατασκευή του, κρατήσαμε τα πεδία: title, release date, origin, director, cast, genre, URL και rating count. Σε κάθε πεδίο έγινε απαλοιφή stop words με τη βοήθεια analyzer, καθώς και tokenization.

3.1 ANALYZERS

Καθώς το database μας αποτελείται από αγγλικές λέξεις, ο analyzer που χρησιμοποιήθηκε, ο EnglishAnalyzer, διότι επεξεργάζεται ορθά τις αγγλικές λέξεις και πραγματοποιεί stemming, σε αντίθεση με τους υπόλοιπους analyzers. Επομένως η δημιουργία Porter Stemmer δεν θεωρήθηκε αναγκαία.

3.2 INDEXED AND STORED FEILDS

Τα πεδία που προστέθηκαν στο ανεστραμμένο ευρετήριο, αποθηκεύτηκαν (indexed and stored fields) για την μετέπειτα χρήση τους στην προβολή των αποτελεσμάτων της αναζήτησης. Ακόμη, έγιναν store τα Wikipedia URL και rating count, για την μετάβαση του χρήστη στο άρθρο της Wikipedia και την κατάταξη με βάση τη βαθμολόγηση, στα οποία θα αναφερθούμε στη συνέχεια. Φυσικά τα Wikipedia URL και rating count δεν συμπεριλήφθηκαν στο inverted index (stored but not indexed field), διότι .

4 ΑΝΑΖΗΤΗΣΗ

Στο κομμάτι της αναζήτησης οι ελάχιστες απαιτήσεις του προγράμματος ικανοποιούνται καθώς και η αναζήτησης πεδίου και το ιστορικό αναζητήσεων.

Οι αναζητήσεις υλοποιήθηκαν με τη βοήθεια Query, QueryParser και MultiFieldQueryParser.

4.1 ΑΝΑΖΗΤΗΣΗ ΜΕ ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Ο χρήστης μπορεί να θέσει ερωτήματα στο σύστημα, γράφοντας στο search bar οποιαδήποτε λέξη κλειδί επιθυμεί. Εκείνο θα το επεξεργαστεί με τη βοήθεια ενός analyzer,του EnglishAnalyzer και ενός tokenizer. Και ύστερα θα αναζητήσει το επεξεργασμένο ερώτημα που του τέθηκε σε όλα τα πεδία του inverted index. Αυτή η λειτουργία υλοποιείται με τη χρήση του απλού Search και όχι του Advanced Search.

4.2 ΑΝΑΖΗΤΗΣΗ ΠΕΔΙΟΥ

Όσον αφορά την αναζήτηση με βάση κάποιο πεδίο, το σύστημα διαθέτει την δυνατότητα Advanced Search. Πατώντας το αντίστοιχο κουμπί στον χρήστη εμφανίζεται ένα παράθυρο στο οποίο μπορεί να αναθέσει τιμές στα αντίστοιχα πεδία. Να σημειωθεί εδώ πως δεν είναι αναγκαίο να αναθέσει τιμές σε όλα, αλλά τουλάχιστον σε ένα. Έτσι ο χρήστης μπορεί να αναζητήσει με βάση ένα μόνο πεδίο (π.χ.Title) αλλά και με δύο (π.χ. Title, Director), τρία (π.χ. Year, Director, Cast), ακόμα και με όλα.

4.3 ΙΣΤΟΡΙΑ ΑΝΑΖΗΤΗΣΕΩΝ

Για την υλοποίηση του search history δημιουργείται ένα αρχείο txt έτσι ώστε να καταγράφει τις αναζητήσεις του χρήστη. Κάθε φορά που ο χρήστης πραγματοποιεί μία αναζήτηση είτε αυτή φέρει αποτελέσματα, είτε όχι το σύστημα την αποθηκεύει στο αρχείο αναζητήσεων. Έτσι στην επόμενη εκτέλεση του προγράμματος, στο πλαίσιο search history, θα μπορεί να δει παλαιότερες αναζητήσεις που ίσως τον ενδιαφέρουν.

5 ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η παρουσίαση των αποτελεσμάτων στον χρήστη έχει τις εξής ιδιότητες: παρουσίαση αποτελεσμάτων ανά δέκα, οι λέξεις κλειδιά παρουσιάζονται τονισμένες και διαφορετική διάταξη.

5.1 ΔΙΑΤΑΞΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Κάθε φορά που ο χρήστης πραγματοποιεί μία αναζήτηση στο σύστημα με λέξεις κλειδιά, είτε στο απλό Wikipedia Movie Searcher, είτε Advanced Search, ένα παράθυρο με τα αποτελέσματα της αναζήτησης του θα εμφανιστεί. Σε αυτό το παράθυρο θα περιέχονται τα πρώτα δέκα αποτελέσματα διατεταγμένα με βάση τη συνάφεια τους με το ερώτημα.

Η δυνατότητα για την εναλλακτική παρουσίαση αποτελεσμάτων υπάρχει μόνο στην αναζήτηση μέσω του Advanced Search. Στο Advanced Search παράθυρο, ο χρήστης μπορεί να διαλέξει ανάμεσα σε κατάταξη με βάση τη χρονιά(year), την αναλογία rating - rating count (rating) και φυσικά τη συνάφεια (default). Για την υλοποίηση των κατατάξεων έγινε χρήση του Sort της Lucene. Όσον αφορά την κατάταξη σύμφωνα με την αναλογία rating - rating count, αυτό είναι το σημείο όπου αξιοποιείται η αποθηκευμένη πληροφορία για το rating count.

Για την υλοποίηση των παραπάνω κατατάξεων αξιοποιήθηκαν τα TopDocs και ScoreDocs.

5.2 HIGHLIGHTING

Οι λέξεις κλειδιά με τις οποίες δημιούργησε το χρήστης το ερώτημα, εμφανίζονται τονισμένες, σε όποιο πεδία κι αν βρίσκονται, σε οποιοδήποτε απλό Search και Advanced.

5.3 ΜΕΤΑΒΑΣΗ ΣΤΟ WIKIPEDIA ΑΡΘΡΟ

Για να μεταβεί ο χρήστης στο άρθρο της Wikipedia για την ταινία που αναζητά, αυτό που χρειάζεται να κάνει είναι να πατήσει πάνω στο τίτλο. Πατώντας τον θα μεταβεί στο αντίστοιχο άρθρο, με τη βοήθεια του default bowser του. Αυτό υλοποιήθηκε δημιουργώντας ένα hyperlink, με όνομα το τίτλο της ταινίας, το οποίο περιέχει το Wikipedia URL του συγκεκριμένου άρθρου, το οποίο ανακτήθηκε από το database και αποθηκεύτηκε για αυτόν τον λόγο.

6 USER INTERFACE

Για την αναζήτηση των ταινιών έχει δημιουργηθεί ένα απλό user interface. Για να το χρησιμοποιήσει ο χρήστης δεν έχει παρά να τρέξει το πρόγραμμα. Το βασικό παράθυρο αποτελείται από ένα search bar ένα κουμπί αναζήτησης search ένα πλαίσιο στο οποίο εμφανίζεται η ιστορία αναζητήσεων και ένα κουμπί advanced search.

6.1 ΑΠΛΟ SEARCH

Για να πραγματοποιήσει μια απλή αναζήτηση ο χρήστης μπορεί να γράψει στη μπάρα κάποιες λέξεις κλειδιά και έπειτα να πατήσει το search.Τότε το πρόγραμμα θα του εμφανίσει ένα νέο παράθυρο στο οποίο θα βρίσκονται τα αποτελέσματα της αναζήτησης του. Κάθε αποτέλεσμα αποτελείται από τα χαρακτηριστικά του, ενώ οι λέξεις κλειδιά εμφανίζονται σε bold. Για να μεταβεί στο αντίστοιχο άρθρο ο χρήστης πρέπει να πατήσει πάνω στον τίτλο και αυτόματα θα μεταφερθεί στο Wikipedia άρθρο μέσω του default bowser του με χρήση του διαδικτύου.

6.2 ADVANCED SEARCH

Για να πραγματοποιήσει μια πιο συνθέτη αναζήτηση ο χρήστης πρέπει να πατήσει το κουμπί Advanced Search. Πατώντας το θα του ανοίξει ένα καινούριο παράθυρο με άδειες θέσεις για να γεμίσει με τις τιμές του επιθυμεί (δεν είναι απαραίτητο να γεμίσουν όλες). Μπορεί ακόμα να επιλέξει την κατάταξη με την οποία επιθυμεί να εμφανιστούν τα αποτελέσματα του. Επίσης υπάρχει ένα πλαίσιο για την ιστορία αναζητήσεων και εδώ. Εάν πατήσει το κουμπί Search, αφού έχει γεμίσει έστω ένα κενό, θα εμφανιστεί ένα νέο παράθυρο όμοιο με αυτό της απλής αναζήτησης.

7 ΣΗΜΕΙΩΣΕΙΣ

Για την μετάβαση στα Wikipedia άρθρα και επομένως για την ολοκληρωμένη λειτουργία του συστήματος απαιτείται η σύνδεση στο διαδίκτυο.

Οι λειτουργία embedding καθώς και η διόρθωση ορθογραφικών λαθών, δεν υλοποιήθηκαν.