Ηλίας Μαυριάνος 2753

Κωνσταντίνα Βαταβάλη 2649

Κώδικας: https://drive.google.com/open?id=1_IbSZ_JE507RS0pSHHgraaEFoE0X-DjS

Μηχανή αναζήτησης Wikipedia άρθρων

Σύντομη περιγραφή σχεδιασμού και συλλογής δεδομένων

Στόχος της συγκεκριμένης εργασίας, ήταν να δημιουργηθεί μία μηχανή αναζήτησης που να επιστρέφει άρθρα της Wikipedia, με την χρήση της βιβλιοθήκης Lucene. Επιλέξαμε τα άρθρα μας να αφορούν κινηματογραφικές ταινίες.

1. Συλλογή εγγράφων (corpus):

Τα έγγραφα που αποτελούν την συλλογή μας είναι από το archive του Kaggle και συγκεκριμένα ένα csv αρχείο, που περιέχει περιγραφές 34.886 ταινιών από όλον τον κόσμο. Διαλέξαμε από αυτές 140 ταινίες που θα αποτελούν το corpus μας, χωρίς κάποιο συγκεκριμένο κριτήριο, παρά μόνο να έχουν κοντινές ημερομηνίες κυκλοφορίας.

Οι στήλες του αρχείου είναι οι εξής: Release Year, Title, Origin/Ethnicity, Director, Genre, Wiki Page, Plot. Στην συνέχεια, αυτές οι στήλες θα αποτελέσουν τα Fields των Documents μας, εκτός του Origin/Ethnicity. Οι ταινίες είναι αποθηκευμένες στο αρχείο movies_wiki.csv. Κάθε γραμμή του, αντιπροσωπεύει μία ταινία.

Διατρέξαμε κάθε γραμμή του csv αρχείου και με κατάλληλη προ-επεξεργασία αυτής, κρατήσαμε τις πληροφορίες που μας ενδιαφέρουν. Η κάθε στήλη ήταν χωρισμένη με κόμμα, εκτός της στήλης plot που ξεκινούσε και τελείωνε με double quotes και εμπεριείχε κενά.

2. Ανάλυση κειμένου και κατασκευή ευρετηρίου.

Δημιουργήσαμε ένα Document, για κάθε ταινία. Τα Fields που θα περιέχει το καθένα από αυτό θα είναι τα εξής:

-**Release Year:** Η χρονιά κυκλοφορίας της ταινίας. Δεν θα γίνει ANALYZED, γιατί θέλουμε την χρονολογία ως ενιαίο TOKEN, όμως θα γίνει STORED επειδή θα την προβάλλουμε στον χρήστη, με τα αποτελέσματα της αναζήτησης.

-**Title:** Ο τίτλος της ταινίας. Θα γίνει ANALYZED, εφόσον ο τίτλος μπορεί να αποτελείται από μία ή παραπάνω λέξεις και STORED γιατί θα γίνεται προβολή του στα αποτελέσματα της αναζήτησης

-**Director - Director(s):** Ο σκηνοθέτης(ή οι σκηνοθέτες) της ταινίας. Θα γίνει ANALYZED, αλλά όχι STORED. Δεν θα εμφανίζεται στα αποτελέσματα, όμως ο χρήστης μπορεί να κάνει αναζήτηση κάποιον σκηνοθέτη ή σκηνοθέτες.

-Cast: Οι ηθοποιοί της ταινίας. Θα γίνει ANALYZED, αλλά όχι STORED. Παρέχουμε στον χρήστη την δυνατότητα να κάνει αναζήτηση με το όνομα ενός ηθοποιού, όμως δεν μας ενδιαφέρει να εμφανίζεται στην λίστα των αποτελεσμάτων.

-Genre: Το είδος (ή τα είδη) της ταινίας. Θα γίνει ANALYZED, αλλά όχι STORED. Όπως και προηγουμένως μπορεί να γίνει αναζήτηση του είδους της ταινίας, όμως δεν μας ενδιαφέρει η εμφάνιση στον χρήστη.

-Wiki Page: Το URL της WIKIPEDIA σελίδας από την οποία συλλέχθηκε η πληροφορία. Δεν θα γίνει ANALYZED, αλλά θα γίνει STORED. Αυτό θα μας δώσει την δυνατότητα, να εμφανίζεται η σελίδα Wikipedia της ταινίας με μορφή hyperlink.

-Plot: Περιγραφή της πλοκής της ταινίας. Θα γίνει ANALYZED, και STORED ώστε να προβληθεί ένα μέρος της στον χρήστη, που θα περιέχει την λέξη κλειδί.

Μέσω των κλάσεων και των μεθόδων τους, που προσφέρει η βιβλιοθήκη Lucene, δημιουργήσαμε τα απαραίτητα πεδία. Έπειτα, τα προσθέσαμε στα Documents που είχαμε δημιουργήσει προηγουμένως. Όταν ολοκληρώθηκε η σύνταξη των Documents, τότε κατασκευάστηκε το ανεστραμμένο ευρετήριο, χωρισμένο σε τμήματα βάση των πεδίων.

Ο analyzer που χρησιμοποιήσαμε για την λεκτική ανάλυση των πεδίων, δημιουργήθηκε από εμάς. Αυτό που κάναμε, ήταν να προσθέσουμε ένα επιπλέον φίλτρο, το Porter Stemmer Filter στον ήδη υπάρχων Standard Analyzer. Αυτό, δίνει την δυνατότητα stemming στα πεδία.

Τα φίλτρα που περιέχει ήδη ο Standard Analyzer, είναι το LowerCaseFilter για μετατροπή κεφαλαίων γραμμάτων σε μικρά και το Stop Filter για αποκοπή λέξεων όπως άρθρων ή συνδέσμων.

Παραδείγματος χάριν, η πρόταση The quick brown foxes jumped over the lazy dog, με τον analyzer που κατασκευάστηκε θα μετατραπεί σε: [quick] [brown] [fox] [jumped] [over] [lazy] [dog].

Το ευρετήριο θα δημιουργηθεί μία φορά, και θα αποθηκευτεί στον δίσκο, και όχι στην μνήμη.

3. Αναζήτηση άρθρων:

Ο χρήστης θα μπορεί να αναζητάει άρθρα, με λέξεις – κλειδιά, φράσεις, BOOLEAN ερωτήσεις, ερωτήσεις wild-card.

Παραδείγματος χάριν, εάν πληκτρολογήσει com*dy , θα εμφανιστούν όλες οι ταινίες που ταιριάζουν με τον όρο comedy, ή αν γράψει co*, θα εμφανιστούν όλες οι ταινίες που περιέχουν λέξεις που ξεκινούν με το δίγραμμα co.

Αν ρωτήσει (walk OR night) AND day, η μηχανή θα παρουσιάσει ταινίες που περιέχουν σίγουρα τον όρο day και έναν από τους όρους walk και night ή και τους δύο.

Αν η ερώτηση του είναι περικλειόμενη από double quotes, τότε τα αποτελέσματα θα πρέπει να περιέχουν την φράση που αναζήτησε.

Ο χρήστης θα έχει την δυνατότητα να πραγματοποιεί αναζήτηση λέξεων – κλειδιά σε συγκεκριμένο πεδίο. Αυτό το καθιστά δυνατό ο χωρισμός του ευρετηρίου σε ξεχωριστά τμήματα μέσω των Fields που είχαμε δημιουργήσει στην αρχή.

Ο χρήστης μόλις πληκτρολογήσει την ερώτηση του στο ειδικό πλαίσιο, έχει δυο επιλογές. Αν πατήσει το πλήκτρο SEARCH, τότε θα εμφανιστούν αποτελέσματα που περιέχουν τις λέξεις-κλειδιά, σε οποιοδήποτε field του Document.

Επιπρόσθετα, έχει την δυνατότητα να πραγματοποιήσει αναζήτηση σε συγκεκριμένο field. Αν πατήσει στο menu SEARCH BY, τότε θα εμφανιστεί μία λίστα με όλα τα δυνατά πεδία αναζήτησης. Αυτά είναι τα εξής: release year, title, cast, genre, director.

Χρησιμοποιήσαμε δύο μεθόδους για την αναζήτηση των όρων, μία για κάθε τρόπο αναζήτησης. Στην δεύτερη περίπτωση προσθέσαμε το αλφαριθμητικό (παραδείγματος χάριν «Title:») που αντιστοιχεί στο επιθυμητό field, στο αλφαριθμητικό που περιέχει την ερώτηση του χρήστη. Με το συγκεκριμένο συντακτικό, η μηχανή αναζήτησης καταλαβαίνει ότι πρέπει να ψάξει σε συγκεκριμένο field.

4. Παρουσίαση Αποτελεσμάτων.

Ο χρήστης βλέπει τα αποτελέσματα ανά 5, με δυνατότητα να προχωρήσει στα επόμενα 5, μέχρι να μην υπάρχουν πλέον έγγραφα που να ταιριάζουν με την αναζήτηση. Επιλέξαμε τον συγκεκριμένο αριθμό από έγγραφα να εμφανίζεται σε κάθε σελίδα, ως αναλογία των συνολικών ταινιών που υπάρχουν στο corpus μας.

Αν πατήσει στο κουμπί GO TO NEXT PAGE, τότε θα γίνει ξανά η αναζήτηση των ταινιών, που θα επιστρέφει πλέον επιπλέον 5 συναφείς ταινίες, και στην οθόνη θα εμφανίζεται η τελευταία πεντάδα. Αυτό μπορεί να το κάνει επαναλαμβανόμενα, μέχρι να τελειώσουν οι συναφείς ταινίες. Όταν δεν υπάρχουν πλέον άλλα αποτελέσματα, εμφανίζεται το μήνυμα NO MORE RESULTS στην οθόνη του χρήστη.

Τα αποτελέσματα της αναζήτησης έχουν της εξής διάταξη. Εμφανίζεται ένας αύξων αριθμός, έπειτα στην ίδια γραμμή ο τίτλος της ταινίας και η ημερομηνία κυκλοφορίας της. Στις επόμενες γραμμές υπάρχει ο σύνδεσμος στο άρθρο της Wikipedia. Αν οι λέξεις κλειδιά εμφανίζονται στο plot της ταινίας, τότε από κάτω υπάρχουν προτάσεις που περιέχουν τις λέξεις αυτές τονισμένες.

Οι ταινίες που επιστρέφονται στον χρήστη είναι διατεταγμένες με βάση την συνάφεια τους. Όμως, ο χρήστης έχει την δυνατότητα για αναδιάταξη των αποτελεσμάτων. Μπορεί να επιλέξει το κουμπί SHOW OLDER MOVIES FIRST, πριν την αναζήτηση. Τότε οι ταινίας θα διαταχθούν με βάση την ημερομηνία κυκλοφορίας τους. Στην αρχή θα εμφανίζονται οι παλαιότερες ταινίες και στην συνέχεια οι πιο πρόσφατες.