

Εργασία Ανακτηση Πληροφοριας 2022



Μάριος Ζίχναλης 3226

<https://github.com/marioszih/Data-Retrieval-2022>

Επιλογή αρχείου

Το αρχείο βρέθηκε στο Kaggle. Ο συνδεσμος στον οποίο βρήκα το αρχείο είναι ο εξής:

<https://www.kaggle.com/datasets/heyueyuan/rottentomatoesmoviesandcriticsdatasets>

Το αρχείο περιέχει 2 εγγραφές από τα οποία εγώ θα χρησιμοποιήσω το ένα με το όνομα rotten_tomatoes_movies.csv. Το έγγραφο έχει περίπου 16000 εγγραφές. Από την προεπεξεργασία, κατά την αφαίρεση των null τιμών μένουν περίπου 7600 εγγραφές. Πέρα από εκεί το αρχείο περιέχει 23 στήλες, από τις οποίες θα κρατήσω τα:

- Rotten Tomatoes Link
- Movie Title
- Movie Info
- Critics Consensus
- Rating
- Genre
- Directors
- Writers
- Cast
- Runtime
- Tomatometer Rating
- Audience Rating

Θα κρατηθούν οι πρώτες 20 εγγραφές για το πρώτο κομμάτι της αναφοράς όπως ζητήθηκε στις οδηγίες, και θα παραδωθούν σε ένα .csv αρχείο στο github που έβαλα παραπάνω.

Προεπεξεργασία

Το αρχείο που θα χρησιμοποιηθεί είναι τύπου .csv. Η προεπεξεργασία των άρθρων για τη δημιουργία του εγγράφου γίνεται με τη βοήθεια του **Standard Analyzer**. Ο αναλυτής χωρίζει το κείμενο όπου βρίσκει χαρακτήρες που δεν είναι γράμματα, εφαρμόζει το lower-case φίλτρο και εξαλείφει λέξεις που ανήκουν στην stop-word λίστα. Η επιλογή του συγκεκριμένου αναλυτή έγινε λόγω της δομής του υπάρχοντος εγγράφου. Η μονάδα εγγράφου είναι ουσιαστικά η γραμμή του κειμένου, η πληροφορία για ένα συγκεκριμένο άρθρο. Τα περιεχόμενα ανήκουν όλα σε ένα από τα προαναφερθείσα πεδία. Επιπλέον θα χρησιμοποιηθούν αντεστραμμένα ευρετήρια για όλα τα πεδία εκτός από το πρώτο, το οποίο αφορά το link για την σελίδα στο rotten tomatoes. Αυτό θα γίνει με σκοπό την μείωση του απαιτούμενου χρόνου αναζήτησης και εμφανίζονται τα αποτελέσματα απευθείας στον χρήστη.

Τρόπος Αναζήτησης

Πέρα από το ζητούμενο το οποίο είναι η αναζήτηση με λέξει κλειδί, θα υποστηρίζει και την αναζήτηση με πεδίο, και συγκεκριμένα την εμφάνιση συγκεκριμένων αποτελεσμάτων του πεδίου κάνοντας την ερώτηση πάνω στο αντεστραμμένο ευρετήριο του πεδίου. Τα κλειδιά από όλες αυτές τις αναζητήσεις θα αποθηκευονται σε μια λίστα και μέσω αυτών θα προτείνονται εναλλακτικά ερωτήματα. Κάθε ερώτηση επίσης θα διατηρηθεί και θα δημιουργηθεί ένα ιστορικό αναζητήσεων στο οποίο ο χρήστης θα έχει πρόσβαση αν θέλει να κάνει την ίδια ερώτηση.

Παρουσίαση Αποτελεσμάτων

Όπως ζητείται και στην άσκηση, τα αποτελέσματα που θα παρουσιάζονται, θα περιέχουν την λέξη κλειδί την οποία ο χρήστης θα γράψει. Αν ο χρήστης διαλέξει να κάνει ερώτηση πάνω σε ένα συγκεκριμένο πεδίο, τότε θα εμφανιστούν όλα τα αποτελέσματα του συγκεκριμένου πεδίου, τα οποία περιέχουν την απαιτούμενη λέξη. Να αναφερθεί ότι η αναζήτηση γίνεται ελεγχοντας αν όλα τα γράμματα της λέξης περιέχονται στις εγγραφές. Αυτό μπορεί να εμφανίσει κάποια μη επιθυμητά αποτελέσματα. Επιπροσθέτα ένα άλλο πρόβλημα το οποίο μπορεί να εμφανιστεί είναι η εμφάνιση αποτελεσμάτων που απλά έχουν λεξιλογική συνδεση. Αυτό θα είναι πιο εμφανές σε περίπτωση αναζήτησης χωρίς πεδίο όπου ένας ηθοποιός και ένας director μπορεί να έχουν παρόμοιο ή ίδιο όνομα και να εμφανιστούν ταινίες που δεν θα έπρεπε να εμφανιστούν. Η λέξη κλειδί που θα χρησιμοποιηθεί θα εμφανίζεται με πιο έντονα γράμματα για να ξεχωρίζει από τις υπολοίπες λέξεις. Η παρουσίαση των αποτελεσμάτων θα γίνεται σε ομάδες των 10, και ενδεχομένως να δίνεται δυνατότητα μείωσης ή αύξησης αυτού του αριθμού και την επιλογή να πηγαίνει ο χρήστης μπροσ-πίσω στις σελίδες αποτελεσμάτων.