Εργασια Ανακτηση Πληροφοριας 2022

Μάριος Ζίχναλης 3226

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

Επιλογη αρχειου

Το αρχειο βρεθηκε στο Kaggle. Ο συνδεσμος στον οποιο βρηκα το αρχειο ειναι ο εξης: https://www.kaggle.com/datasets/heyueyuan/rottentomatoesmoviesandcriticsdatasets

Το αρχειο περιεχει 2 εγγραφα απο τα οποια εγω θα χρησιμοποιησω το ενα με το ονομα rotten_tomatoes_movies.csv. Το εγγραφο εχει περιπου 16000 εγγραφες. Απο την προεπεξεργασια, κατα την αφαιρεση των null τιμων μενουν περιπου 7600 εγγραφες. Θα κρατηθουν οι πρωτες 20 εγγραφες για το πρωτο κομματι της αναφορας οπως ζητηθηκε στις οδηγειες, και θα παραδωθουν σε ενα .csv αρχειο στο github που εβαλα παραπανω. Η προεπεξεργασια αυτη εγινε με την χρηση python και συγκεκριμενα το αρχειο prep.py. Το αρχειο περιεχει 23 στηλες, απο τις οποιες θα κρατησω τα:

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

Προεπεξεργασια

Το αρχειο που θα χρησιμοποιηθει ειναι τυπου .csv. Η προεπεξεργασια των αρθρων για τη δημιουργία του εγγράφου γινεται με τη βοήθεια του **Standard Analyzer**. Ο αναλυτης χωριζει το κειμενο οπου βρισκει χαρακτηρες που δεν ειναι γράμματα, εφαρμοζει το lower-case φιλτρο και εξαλειφει λεξεις που ανήκουν στην stop-word λιστα. Η επιλογη του συγκεκριμένου αναλυτη εγινε λογω της δομης του υπαρχοντος εγγράφου. Η μοναδα εγγραφου είναι ουσιαστικα η γραμμη του κειμενου, η πληροφορια για ενα συγκεκριμενο αρθρο. Τα περιεχομενα ανηκουν ολα σε ενα απο τα προαναφερθεισα πεδια. Επιπλεον θα χρησιμοποιηθουν αντεστραμμενα ευρετηρια για ολα τα πεδια εκτος απο το πρωτο, το οποιο αφορα το link για την σελιδα στο rotten tomatoes. Αυτο θα γινει με σκοπο την μειωση του απαιτουμενου χρονου αναζητησης και εμφανιζονται τα αποτελεσματα απευθειας στον χρηστη.

Τροπος Αναζητησης

Περα απο το ζητουμενο το οποιο ειναι η αναζητηση με λεξει κλειδι, θα υποστηριζει και την αναζητηση με πεδιο, και συγκεκριμενα την εμφανιση συγκεκριμενων αποτελεσματων του πεδιου κανωντας την ερωτηση πανω στο αντεστραμμενο ευρετηριο του πεδιου. Τα κλειδια απο ολες αυτες τις αναζητησης θα αποθηκευονται σε μια λιστα και μεσω αυτων θα προτεινονται εναλλακτικα ερωτηματα. Καθε ερωτηση επισης θα διατηρηται και θα δημιουργηται ενα ιστορικο αναζητησεων στο οποιο ο χρηστης θα εχει προσβαση αν θελει να κανει την ιδια ερωτηση.

Παρουσιαση Αποτελεσματων

Οπως ζητηται και στην ασκηση, τα αποτελεσματα που θα παρουσιαζονται, θα περιεχουν την λεξη κλειδι την οποια ο χρηστης θα γραψει. Αν ο χρηστης διαλεξει να κανει ερωτηση πανω σε ενα συγκεκριμενο πεδιο, τοτε θα εμφανιστουν ολα τα αποτελεσματα του συγκεκριμενου πεδιου, τα οποια περιεχουν την απαιτουμενη λεξη. Να αναφερθει οτι η αναζητηση γινεται ελεγχοντας αν ολα τα γραμματα της λεξης περιεχονται στις εγγραφες. Αυτο μπορει να εμφανισει καποια μη επιθυμητα αποτελεσματα. Επιπροσθετα ενα αλλο προβλημα το οποιο μπορει να εμφανιστει ειναι η εμφανιση αποτελεσματων που απλα εχουν λεξιλογικη συνδεση. Αυτο θα ειναι πιο εμφανες σε περιπτωση αναζητησης χωρις πεδιο οπου ενας ηθοποιος και ενας director μπορει να εχουν παρομοιο η ιδιο ονομα και να εμφανιστουν ταινιες που δεν θα επρεπε να εμφανιστουν. Η λεξη κλειδι που θα χρησιμοποιηται θα εμφανιζεται με πιο εντονα γραμματα για να ξεχωριζει απο τις υπολοιπες λεξεις. Η παρουσιαση των αποτελεσματων θα γινεται σε ομαδες των 10, και ενδεχομενως να δινεται δυνατοτητα μειωσης η αυξησης αυτου του αριθμου και την επιλογη να πηγαινει ο χρηστης μπρος-πισω στις σελιδες αποτελεσματων.