Πανεπιστήμιο Ιωαννίνων
Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Ανάκτηση Πληροφορίας

1η Φάση

Μέλη ομάδας:

- Γεώργιος Δεληγιώργης, 4662
- Ελευθέριος-Μάριος Μανίκας, 4723

ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2024
ΠΑΡΑΔΟΣΗ: Τετάρτη 10 Απριλίου

Περιεχόμενα

1.Github link

https://github.com/manikas07/Anaktish-Plhroforias.git

2. Συλλογή Εγγράφων

Όπως μας υποδεικνύεται από την εκφώνηση κατεβάσαμε τα έγγραφα με τα επιστημονικά κείμενα από το παρακάτω link https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated/data?select=papers.csv. Επιλέξαμε από ένα σύνολο 9680 άρθρων 310 άρθρα για να τα χρησιμοποιήσουμε ως την βάση δεδομένων μας. Το format του αρχείου είναι .csv(comma separated values).

Στην πρώτη στήλη του αρχείου έχουμε το source_id του κάθε άρθρου το οποίο δεν είναι μοναδικό.

Στην δεύτερη στήλη έχουμε την χρονιά που δημοσιοποιήθηκε το κάθε άρθρο.

Στην τρίτη στήλη τον τίτλο του.

Στην τέταρτη στήλη το abstract που είναι μια μικρή περίληψη του άρθρου και μπορεί να παραλείπεται.

Στην πέμπτη στήλη είναι όλο το υπόλοιπο κείμενο του άρθρου.

3. Περιγραφή Σχεδιασμού Συστήματος

3.1 Στόχος & Λειτουργικότητα

Ο στόχος για την συγκεκριμένη εργασία είναι να δημιουργήσουμε ένα interface στο οποίο ο χρήστης θα μπορεί να θέτει ερωτήματα τα οποία θα σχετίζονται με επιστημονικά άρθρα που είναι αποθηκευμένα στην Βάση Δεδομένων. Αυτό επιτυγχάνεται μέσω μιας γραφικής διεπαφής (GUI) και το σύστημα μας θα τις διαχειρίζεται με την βοήθεια της βιβλιοθήκης ανοιχτού κώδικα Lucene.

3.2 Ανάλυση κειμένου και κατασκευή ευρετηρίου

Για την ανάλυση του .csv αρχείου επιλέξαμε τον StandardAnalyzer της Lucene διότι μετατρέπει όλους τους χαρακτήρες σε lower case και απομακρύνει τις περιττές λέξεις(π.χ. the, and, or...). Θα δημιουργήσουμε ένα QueryParser όπου θα εισάγουμε όλες τις πιθανές ερωτήσεις που μπορεί να κάνει ο χρήστης. Μέσω αυτής του API πακέτου org.apache.lucene.index θα δημιουργήσουμε τον IndexWriter και IndexReader του ευρετηρίου που διαπερνούν τα δεδομένα μας. Επίσης θα χρησιμοποιήσουμε το API org.apache.lucene.search για να φτιάξουμε τους TermQuery, PhraseQuery που θα χρησιμοποιηθούν για τον κάθε τύπο ερώτησης.

3.3 Αναζήτηση

Αρχικά, η αναζήτηση θα γίνεται μέσω ενός γραφικού περιβάλλοντος στο οποίο ο χρήστης θα εισάγει την εκάστοτε ερώτηση που επιθυμεί.

Οι επιτρεπτές ερωτήσεις αφορούν συγκεκριμένα πεδία :

- Τον τίτλο του άρθρου.
- Όροι που υπάρχουν στο abstract.
- Όροι που υπάρχουν στο full_text.
- Αναζήτηση άρθρων με βάση το πεδίο year.

3.4 Παρουσίαση Αποτελεσμάτων

Το σύστημα θα επιστρέψει στον χρήστη τα άρθρα που ταιριάζουν περισσότερο με το ερώτημα που υπέβαλλε ο χρήστης ταξινομημένα κατά φθίνουσα σειρά συνάφειας. Επίσης θα εμφανίζονται ανά 10, δίνοντας την δυνατότητα στον χρήστη να προχωρήσει στα επόμενα 10 αν υπάρχουν. Επιπρόσθετα οι λέξεις κλειδιά θα είναι bold στο αποτέλεσμα και θα υπάρχει η δυνατότητα να ταξινομηθούν τα αποτελέσματα με βάση την χρονιά δημοσίευσης.