ΕΡΓΑΣΙΑ ΑΝΑΚΤΗΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ - 1η ΦΑΣΗ

Κωνσταντίνος Παπαντωνίου-Χατζηγιώσης 4769, Βασίλειος Σωμάκος 4806

Α) Η συλλογή εγγράφων που θα χρησιμοποιήσουμε στην εργασία αποτελείται από 200 επιστημονικά άρθρα που περιλαμβάνονται στο σύνολο δεδομένων papers.csv και authors.csv από το Kaggle. Θα χρησιμοποιήσουμε τα παρακάτω πεδία:

Source_id: Μοναδικό αναγνωριστικό του άρθρου.

year: Έτος δημοσίευσης του άρθρου.

title: Τίτλος του άρθρου.

abstract: Περίληψη του άρθρου.

full_text: Κείμενο του άρθρου.

First_name/Last name: Μικρό όνομα και επώνυμο του συγγραφέα του άρθρου(foreign key -> source_id του άρθρου)

Αυτά τα πεδία παρέχουν τις βασικές πληροφορίες που απαιτούνται για την ανάλυση και την αναζήτηση επιστημονικών άρθρων. Το πεδίο "title" παρέχει τον τίτλο του άρθρου, ενώ το "abstract" περιέχει μια σύντομη περίληψη του περιεχομένου του. Το "full_text" περιλαμβάνει το πλήρες κείμενο του άρθρου, ενώ το first_name και last_name αναφέρει το όνομα και το επώνυμο του συγγραφέα καθώς και το source_id του, το οποίο είναι ο δείκτης που υποδηλώνει το άρθρο του κάιε συγγραφέα.

Με τη χρήση αυτών των πεδίων, μπορούμε να δημιουργήσουμε ένα σύστημα αναζήτησης πληροφοριών που θα επιτρέπει την αναζήτηση με βάση τον τίτλο, την περίληψη, το περιεχόμενο του άρθρου και τους συγγραφείς.

Β) **Εισαγωγή:** Η ανάπτυξη ενός ευρετηρίου όπου θα περιέχει επιστημονικά έγγραφα και θα δίνει στο χρήστη δυνατότητα αναζήτησης επιστημονικών άρθρων με διάφορους τρόπους, όπως αναζήτηση με λέξεις κλειδιά, αναζήτηση πεδίου (όπως τίτλος, περίληψη, κείμενο άρθρου) και αναζήτηση με βάση λέξεις κλειδιά οι οποίες εμπεριέχονται στα άρθρα τα οποία έχουν συγκεκριμένο συγγραφέα.

Ανάλυση κειμένου και κατασκευή ευρετηρίου: Για την ανάλυση κειμένου και την κατασκευή ευρετηρίου, προβαίνουμε σε διάφορα βήματα προεπεξεργασίας των άρθρων προκειμένου να δημιουργήσουμε ένα συστηματικό και αποτελεσματικό ευρετήριο. Τα βασικά βήματα περιλαμβάνουν:

1. *Καθαρισμός του κειμένου*: Αφαιρούμε όλους τους μη αλφαριθμητικούς χαρακτήρες και σημεία στίξης ή άλλα μη επιθυμητά στοιχεία.

2. *Διαχωρισμός σε λέξεις*: Το κείμενο διαχωρίζεται σε λέξεις ή τερματικά (tokens), ώστε να μπορέσουμε να εφαρμόσουμε περαιτέρω επεξεργασία σε κάθε λέξη ξεχωριστά.

3. *Μετατροπή σε πεζά*: Όλες οι λέξεις μετατρέπονται σε πεζά ώστε να μην υπάρχει διάκριση μεταξύ κεφαλαίων και πεζών γραμμάτων.

4. Αφαίρεση λέξεων: Αφαιρούμε τις συχνά εμφανιζόμενες λέξεις που δεν προσφέρουν πολλή πληροφορία για το περιεχόμενο του κειμένου, όπως "or", "the", "and" κλπ.

5. Κανονικοποίηση κειμένου: Μετατρέπουμε τις λέξεις σε κανονικοποιημένη μορφή για να μειώσουμε τον αριθμό των μορφών μιας λέξης.

   Τα 1,2 και 3 μπορούν να πραγματοποιηθούν με το StandardAnalyzer της Lucene.

Αναζήτηση: Η αναζήτηση θα γίνεται με βάση τις διάφορες επιλογές που θα παρέχονται στον χρήστη μέσω ενός γραφικού περιβάλλοντος χρήστη (GUI). Οι επιλογές αυτές μπορεί να περιλαμβάνουν:

1. Αναζήτηση με βάση τον συγγραφέα (Author): Ο χρήστης μπορεί να εισάγει το όνομα του συγγραφέα που ενδιαφέρεται και να πραγματοποιήσει αναζήτηση για άρθρα που έχουν συγγραφεί από αυτόν.

2. Αναζήτηση με βάση τον τίτλο (Title): Ο χρήστης μπορεί να εισάγει έναν τίτλο άρθρου και να πραγματοποιήσει αναζήτηση για άρθρα που περιέχουν αυτόν τον τίτλο.

3. Αναζήτηση με βάση το έτος δημοσίευσης (Year of Publication): Ο χρήστης μπορεί να επιλέξει ένα συγκεκριμένο έτος δημοσίευσης και να πραγματοποιήσει αναζήτηση για άρθρα που δημοσιεύθηκαν εκείνο το έτος.

4. Αναζήτηση με βάση το αναγνωριστικό του άρθρου (source_id)

Ύστερα με ένα drop down menu θα υπάρξει ταξινόμηση που θα έχει τα εξής:

1. Ταξινόμηση ανάλογα με το χρόνο δημοσίευσης (Date of Publication): Ο χρήστης μπορεί να επιλέξει την επιλογή "Date of Publication" από ένα drop-down μενού και να πραγματοποιήσει αναζήτηση με βάση τον χρόνο δημοσίευσης, με δυνατότητα ταξινόμησης των αποτελεσμάτων ανά χρονολογική σειρά.

2. Ταξινόμηση αλφαβητικά (Alphabetical Order): Ο χρήστης μπορεί να επιλέξει την επιλογή "Alphabetical Order" από το drop-down μενού και να πραγματοποιήσει αναζήτηση τον τίτλο ή των ονομάτων των συγγραφέων.

Παρουσίαση Αποτελεσμάτων: Τα αποτελέσματα της αναζήτησης θα παρουσιάζονται σε μια ευανάγνωστη διάταξη μέσω του γραφικού περιβάλλοντος χρήστη (GUI). Κάποια στοιχεία που θα πρέπει να περιλαμβάνονται στην παρουσίαση των αποτελεσμάτων είναι:

1. Τίτλος άρθρου: Κάθε αποτέλεσμα θα πρέπει να περιλαμβάνει τον τίτλο του αντίστοιχου άρθρου για ευκολότερη αναγνώριση.

2. Συγγραφείς: Μπορεί να εμφανίζονται οι συγγραφείς του άρθρου για περαιτέρω πληροφόρηση και επιβεβαίωση της ακρίβειας της αναζήτησης.

3. Χρόνος Δημοσίευσης (Year of Publication): Η ημερομηνία δημοσίευσης του άρθρου μπορεί να εμφανιστεί για περαιτέρω πληροφόρηση και ταξινόμηση των αποτελεσμάτων.

Επιπρόσθετα, μπορεί να παρέχονται διάφορες επιλογές για την προβολή και τη διαχείριση των αποτελεσμάτων, όπως:

1. Προβολή αποτελεσμάτων ανά σελίδα: Τα αποτελέσματα μπορούν να παρουσιάζονται ανά δεκάδες, με δυνατότητα πλοήγησης μεταξύ των σελίδων και επιλογή του επιθυμητού άρθρου.

2. Υπογράμμιση λέξεων κλειδιών: Οι λέξεις κλειδιά που αντιστοιχούν στην αναζήτηση μπορεί να υπογραμμίζονται στα αποτελέσματα για ευκολότερη εντοπισμό.

3. Ταξινόμηση αποτελεσμάτων: Οι χρήστες μπορούν να επιλέξουν τη μέθοδο ταξινόμησης των αποτελεσμάτων, όπως ανά χρόνο δημοσίευσης ή αλφαβητικά.