Καλιόπη Οικονόμου 5099
Μηλίτσα Βουδούρη 5104

Μηχανή αναζήτησης επιστημονικών άρθρων

Github link

https://github.com/militsavdr/Information-Retrieval.git

Συλλογή εγγράφων - Corpus

Θα χρησιμοποιήσουμε επιστημονικά άρθρα από την παρακάτω συλλογή από το Kaggle:

https://www.kaggle.com/datasets/rowhitswami/nips-papers-1987-2019-updated/data?select=papers.csv

Η συλλογή αποτελείται από δύο αρχεία της μορφής csv τα οποία περιέχουν πληροφορίες για επιστημονικά άρθρα καθώς και για τους συγγραφείς τους . Πιο συγκεκριμένα, το πρώτο αρχείο, με την ονομασία papers.csv, περιέχει πληροφορίες σχετικά με τα επιστημονικά άρθρα:

- source_id : μοναδικό αναγνωριστικό πηγής
- year: χρονολογία δημοσίευσης
- title: τίτλος
- abstarct: περίληψη
- full_text: πλήρες κείμενο

ενώ το δεύτερο αρχείο authors.csv περιέχει πληροφορίες σχετικά με τους συγγραφείς :

- source_id: μοναδικό αναγνωριστικό
- first_name: όνομα συγγραφέα
- last_name: επίθετο συγγραφέα
- institution: ινστιτούτο στο οποίο ανήκει ο συγγραφέας

Για να συσχετίσουμε έναν συγγραφέα με τα δημοσιευμένα του άρθρα, μπορούμε να χρησιμοποιήσουμε το κοινό πεδίο source_id. Κάθε συγγραφέας έχει ένα μοναδικό source_id που τον συσχετίζει με τα άρθρα που έχει γράψει.

Με βάση τις απαιτήσεις του συστήματος, θα χρησιμοποιήσουμε όλα τα πεδία για την υλοποίηση των διαφορετικών μορφών αναζήτησης.

Εισαγωγή

Ο κύριος στόχος του συστήματος αναζήτησης είναι η διευκόλυνση της πρόσβασης σε επιστημονικά άρθρα και στους συγγραφείς τους, μέσω ενός αποτελεσματικού και

εύχρηστου περιβάλλοντος αναζήτησης. Εκτιμούμε ότι η λειτουργικότητα του συστήματός μας θα περιλαμβάνει αναζήτηση με βάση λέξεις-κλειδιά, κάποιο συγκεκριμένο πεδίο αλλά και συνώνυμα λέξεων. Σκοπός είναι να επικεντρώνεται ο σχεδιασμός στην ακρίβεια, την απλότητα και την ολοκληρωμένη κάλυψη των αποτελεσμάτων. Για την υλοποίηση, θα χρησιμοποιήσουμε την βιβλιοθήκη Lucene, η οποία μας παρέχει τα απαραίτητα εργαλεία και λειτουργίες.

Ανάλυση κειμένου και κατασκευή ευρετηρίου

Προεπεξεργασία

Απαραίτητο κρίνουμε ότι θα είναι τα άρθρα και οι πληροφορίες των συγγραφέων να υποστούν ενός είδους προεπεξεργασίας για την αφαίρεση της περιττής πληροφορίας ή και την εξασφάλιση ομοιομορφίας στην δομή των δεδομένων. Συγκεκριμένα σκοπεύουμε να χρησιμοποιήσουμε το API της Lucene **org.apache.lucene.analysis** για την επεξεργασία των κειμένων, μέσω του οποίου επιλέγουμε ως αναλυτή για την μετατροπή των λέξεων σε tokens, τον **Standard Analyzer**. Ο αναλυτής αυτός είναι υπεύθυνος για την αφαίρεση των συχνών λέξεων (stop words), την αφαίρεση σημείων στίξης, καθώς και την μετατροπή όλων των γραμμάτων σε lowercase.

Μονάδα εγγράφου και πεδία

Η μονάδα εγγράφου αποτελείται από όλη την συλλογή μας καθώς συγκεντρώνει όλες τις σχετικές πληροφορίες για ένα άρθρο και τον συγγραφέα του, οργανώνοντάς τες σε ένα ενιαίο σύνολο δεδομένων.

Ευρετήρια

Η δημιουργία των ευρετηρίων θα βασιστεί σε κείμενο για τους τίτλους, τις περιλήψεις και τα πλήρη κείμενα των επιστημονικών άρθρων, αλλά και στα ονόματα και επώνυμα συγγραφέων με την χρήση του API **org.apache.lucene.index** το οποίο παρέχει τις κλάσεις που αφορούν την δημιουργία και την διαχείριση των ευρετηρίων.

Για κάθε άρθρο και πληροφορία συγγραφέα, δημιουργούμε ένα αντικείμενο Document χρησιμοποιώντας το API **org.apache.lucene.document**. Αυτό το έγγραφο περιέχει πολλά πεδία (fields) που αντιπροσωπεύουν τα διάφορα μέρη του άρθρου ή τις πληροφορίες του συγγραφέα.

Αρχικά, η χρήση **ανεστραμμένου ευρετηρίου** είναι απαραίτητη για την αναζήτηση συγκεκριμένων λέξεων σε συγκεκριμένα πεδία, όπως ο τίτλος, η περίληψη κ.λπ. Αυτό επιτυγχάνεται με τον συσχετισμό κάθε λέξης με το αντίστοιχο έγγραφο που την περιλαμβάνει. Με αυτόν τον τρόπο, μπορούμε να

επιτύχουμε πιο γρήγορη και αποδοτική αναζήτηση σε συγκεκριμένα πεδία των εγγράφων.

Επίσης, ένα ακόμα είδος ευρετηρίου που θα χρειαστούμε είναι το **δυναμικό ευρετήριο (dynamic index)**, για τη διαχείριση του ιστορικού αναζητήσεων, καθώς μας παρέχει τη δυνατότητα προσθήκης νέων εγγράφων και ενημέρωσης των δεδομένων χωρίς να χρειάζεται να ανακτήσουμε το ευρετήριο από την αρχή. Αυτό επιτρέπει τη διατήρηση ενός ιστορικού αναζητήσεων και την εξοικονόμηση χρόνου στη διαδικασία αναζήτησης.

Αναζήτηση

Το σύστημα θα υποστηρίζει 3 βασικά είδη αναζήτησης εγγράφων:

(α) Αναζήτηση με λέξεις-κλειδιά (Keyword Search)

Αφορά την αναζήτηση αυτούσιων λέξεων ή φράσεων προκειμένου να γίνει μία πιο ειδική και ακριβής αναζήτηση με βάση συγκεκριμένη ορολογία του αντικειμένου ενδιαφέροντος. Η αναζήτηση αυτή μπορεί να γίνει σε όλα τα πεδία εξασφαλίζοντας ότι κανένα σημαντικό στοιχείο δεν παραβλέπεται και αυξάνοντας σημαντικά τη πιθανότητα ο χρήστης να βρει ακριβώς αυτό που ψάχνει.

(β) Αναζήτηση Πεδίου (Fielded Search)

Επιτρέπει την επικέντρωση σε συγκεκριμένα πεδία προκειμένου να προσφέρει τη δυνατότητα πιο στοχευμένης αναζήτησης από τον χρήστη (π.χ. αναζήτηση με βάση τη χρονολογία δημοσίευσης ενός άρθρου (πεδίο: year)). Αυτό επιτρέπει τη διευκόλυνση της εύρεσης συγκεκριμένων πληροφοριών και την προβολή των αποτελεσμάτων σε σχέση με τα συγκεκριμένα κριτήρια που επιλέγει ο χρήστης. Θα χρησιμοποιήσουμε το **org.apache.lucene.search** καθώς παρέχει τις κλάσεις που αφορούν την εκτέλεση των αναζητήσεων στο ευρετήριο και την επεξεργασία αποτελεσμάτων.

(γ) Αναζήτηση με Συνώνυμα (Synonym Search)

Προσφέρει μεγαλύτερο εύρος στοχευμένων αποτελεσμάτων και την κάλυψη περιπτώσεων διαφορετικής χρήσης ορολογίας ή τρόπου έκφρασης που χρησιμοποιούνται στα επιστημονικά άρθρα. Αυτή η αναζήτηση αφορά πεδία κειμένων (περίληψη και πλήρες κείμενο). Επιτρέπει την εύρεση εγγράφων που ίσως δεν είχαν εντοπιστεί με ένα πιο περιοριστικό ερώτημα αλλά και βελτιώνει σημαντικά την εμπειρία του χρήστη καθώς καλύπτει και αυτούς που δεν είναι εξοικειωμένοι με την ακριβή επιστημονική ορολογία. Έτσι, το σύστημα γίνεται προσιτό και εύχρηστο για όλους τους χρήστες, επιτρέποντας άτομα εκτός της επιστημονικής κοινότητας, αλλά και σε ερευνητές από άλλους επιστημονικούς τομείς, να χρησιμοποιούν τη

μηχανή αναζήτησης με μεγαλύτερη ευκολία και αποτελεσματικότητα. Για την επίτευξη αυτής της αναζήτησης θα χρησιμοποιήσουμε το ευρετήριο **SynonymMap** για τον ορισμό των συνωνύμων μέσω του API **org.apache.lucene.analysis.synonym.**

Παρουσίαση Αποτελεσμάτων

Το σύστημά μας θα παρουσιάζει τα αποτελέσματα της αναζήτησης διατεταγμένα με βάση τη συνάφειά τους με το ερώτημα του χρήστη. Χρησιμοποιώντας το **Java Swing** ως framework, διαθέτουμε ένα πλήρες σύνολο βιβλιοθηκών και εργαλείων για τη δημιουργία γραφικών διεπαφών χρήστη (GUI).

Τα αποτελέσματα θα παρουσιάζονται σε ομάδες των 10, προσφέροντας τη δυνατότητα στον χρήστη να προβεί στην περιήγηση μεταξύ των σελίδων με το πάτημα ενός κουμπιού.

Επιπλέον, οι λέξεις-κλειδιά καθώς και οι συνώνυμες θα εμφανίζονται στα αποτελέσματα τονισμένες για να επισημανθεί η συνάφειά τους με το ερώτημα του χρήστη.

Θα παρέχεται επίσης η δυνατότητα αναδιάταξης των αποτελεσμάτων με βάση τη χρονολογία που δημοσιεύτηκε το κάθε άρθρο.

Με αυτόν τον τρόπο, η εφαρμογή θα παρέχει μια ευέλικτη και χρήσιμη διεπαφή για τους χρήστες που αναζητούν πληροφορίες στο σύστημά μας.