ΜΗΧΑΝΗ ΑΝΑΖΗΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΑΠΟ ΕΠΙΣΤΗΜΟΝΙΚΑ ΑΡΘΡΑ

ΜΑΘΗΜΑ: ΑΝΑΚΤΗΣΗ ΠΛΗΡΟΦΟΡΙΑΣ

ΙΑΤΡΑΚΗΣ ΙΩΑΝΝΗΣ 5116
ΓΚΟΥΔΙΝΑΚΗΣ ΒΑΣΙΛΕΙΟΣ 5130

Στόχος και λειτουργικότητα του συστήματος:

Το πεδίο της επιστημονικής έρευνας αυξάνεται με εκθετικούς ρυθμούς, με νέα άρθρα και εργασίες να δημοσιεύονται καθημερινά. Το σύστημα αναζήτησης πληροφοριών από επιστημονικά άρθρα που θα δημιουργήσουμε έχει ως στόχο να παρέχει στους χρήστες τη δυνατότητα να εντοπίζουν, να προβάλλουν και να ανακτούν σχετική επιστημονική πληροφορία σύμφωνα με τα keywords που αναζητούν. Η βασικότερη λειτουργικότητα είναι να επιστρέφει στους χρήστες όσο το δυνατόν πιο ακριβής και αξιόπιστα αποτελέσματα από τα επιστημονικά άρθρα που έχει το dataset που επιλέξαμε. Επίσης, με αυτό τον τρόπο μπορεί να επιταχύνει την διαδικασία αναζήτησης, ψάχνοντας ανάμεσα σε συγκεκριμένα αποτελέσματα σύμφωνα με τα keywords, αντί να πρέπει να ψάχνει σε όλο το dataset. Τελευταίο αλλά εξίσου σημαντικό είναι, ακόμα και απλή χρήστες χωρίς μεγάλη εμπειρία , να μπορούν να χρησιμοποιήσουν τη μηχανή αναζήτησης μας και να είναι σαφείς οι λειτουργικότητες που παρέχονται. Συνοψίζοντας, η εφαρμογή μας στοχεύει στην ανάπτυξη ενός συστήματος ανάκτησης πληροφορίας προσαρμοσμένο για επιστημονικά άρθρα, εστιάζοντας ειδικά στην παροχή ακριβών, αξιόπιστων και σχετικών αποτελεσμάτων ενώ, έχει σχεδιαστεί για να καλύπτει τις ανάγκες διαφόρων ομάδων χρηστών, συμπεριλαμβανομένων ερευνητών, φοιτητών και επαγγελματιών του κλάδου.

Συλλογή Δεδομένων:

Το dataset είναι εμπνευσμένο από το dataset "NIPS Papers" του Ben Hamner και περιλαμβάνει 9680 επιστημονικά άρθρα τα οποία καλύπτουν την χρονική περίοδο 1987 2019.

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

Η συλλογή των εγγράφων που επιλέξαμε να χρησιμοποιήσουμε είναι τα πρώτα 1000 επιστημονικά άρθρα (με τον ίδιο τρόπο γίνεται και σε ολόκληρο το dataset) τα οποία βρίσκονται στο παραπάνω dataset. Τα πεδία που αποτελούν το dataset είναι το source_id , το οποίο κρατάει ένα μοναδικό id για κάθε άρθρο ώστε να μπορούμε να τα ξεχωρίσουμε. Ακόμη, περιέχει το year το οποίο δημοσιεύτηκε κάθε άρθρο με εύρος τιμών από το 1987 έως το 2019. Επίσης, υπάρχει και το title του, καθώς και το abstract που περιλαμβάνει τα κυριότερα σημεία, τις βασικές ιδέες ή την ουσία του άρθρου. Τέλος, έχει και το full_text ,το οποίο περιέχει το περιεχόμενο του άρθρου, μαζί με κάποιες πληροφορίες για τους συγγράφεις όπως το όνομα ,το επώνυμο, institution κ.α.

Ανάλυση κειμένου και κατασκευή ευρετηρίου:

Ανάλυση κειμένου:

Μετά από μελέτη που κάναμε στο dataset διαπιστώσαμε ότι κάθε άρθρο περιλαμβάνει στην αρχή τα πεδία που αναφέρθηκαν παραπάνω ενώ ο συμβολισμός για να διαπιστώσουμε ότι το άρθρο έφτασε στο τέλος του είναι \f\" . Η μονάδα εγγράφου που θα χρησιμοποιήσουμε είναι ένα άρθρο το οποίο θα έχει ως πεδία τα εξής: source_id ,year,title, abstract, full_text.

Κατασκευή ευρετηρίου:

Στην αναζήτηση με λέξεις κλειδιά θα χρησιμοποιήσουμε ένα ενιαίο ευρετήριο για κάθε άρθρο. Αυτό σημαίνει ότι όλες οι λέξεις κλειδιά θα αναζητούνται σε ένα συνδυαστικό ευρετήριο που περιλαμβάνει όλα τα πεδία του άρθρου.

Ενιαίο Ευρετήριο: Το ενιαίο ευρετήριο αποτελείται από ένα συνδυασμό των πεδίων title, abstract, και full_text, παρέχοντας τη δυνατότητα στον χρήστη να αναζητήσει πληροφορίες σε ολόκληρο το περιεχόμενο του άρθρου. Αυτό είναι ιδιαίτερα χρήσιμο για χρήστες που δεν γνωρίζουν συγκεκριμένα πεδία αλλά αναζητούν γενικές πληροφορίες.

Στην αναζήτηση πεδίου καθώς και στη Boolean αναζήτηση που θα υλοποιήσουμε, θα έχουμε ένα παραμετρικό ευρετήριο για τον τίτλο , το abstract , και το full_text. Αυτό σημαίνει ότι όλες οι λέξεις κλειδιά θα αναζητούνται σε ένα ευρετήριο που περιλαμβάνει ένα τα πεδία του άρθρου.

Παραμετρικό Ευρετήριο: Στην αναζήτηση πεδίου θα έχουμε ένα παραμετρικό ευρετήριο για τον **title**, το **abstract**, και το **full_text**. Το παραμετρικό ευρετήριο επιτρέπει πιο εξειδικευμένες αναζητήσεις, καθώς οι χρήστες μπορούν να περιορίσουν τα αποτελέσματα σε συγκεκριμένα πεδία του άρθρου. Για παράδειγμα, αν κάποιος ενδιαφέρεται μόνο για τους τίτλους των άρθρων, μπορεί να πραγματοποιήσει την αναζήτηση μόνο στο πεδίο title.

Αναζήτηση:

Για την αναζήτηση , ο χρήστης θα πρέπει να εισάγει ένα text μέσα στο πλαίσιο αναζήτησης για το οποίο θέλει να βρει πληροφορίες ώστε το πρόγραμμα να του επιστρέψει σχετικά άρθρα. Για πιο συγκεκριμένη αναζήτηση θα υπάρχει δυνατότητα να ψάξει με βάση κάποιο από τα πεδία επιλέγοντας το στο dropdown menu αριστερά από το πλαίσιο αναζήτησης . Επιπλέον, η Boolean αναζήτηση είναι ένας τρόπος αναζήτησης που χρησιμοποιεί τους λογικούς τελεστές (AND , OR , NOT) για να συνδυάσει ή να εξαιρέσει λέξεις - κλειδιά στα αποτελέσματα αναζήτησης. Τα ερωτήματα σε αυτή την περίπτωση θα είναι της μορφής “text” OPERATION “text”. Μια ακόμη λειτουργία που διαθέτει η μηχανή αναζήτησης είναι ότι κρατάει ιστορικό αναζητήσεων. Ο χρήστης πατώντας το κουμπί «History» πάνω αριστερά , του εμφανίζονται τα 5 τελευταία keywords που έψαξε. Εάν επιλέξει κάποιο από αυτά , το συγκεκριμένο keyword εμφανίζεται στο πλαίσιο αναζήτησης και έτσι ο χρήστης δεν χρειάζεται να πληκτρολογεί ξανά παλαιότερες αναζήτησης που έκανε.

Παρουσίαση Αποτελεσμάτων:

Στο χρήστη θα παρουσιάζονται τα αποτελέσματα ανά 10 με δυνατότητα να προχωρήσει στα επόμενα και να γυρίσει στα προηγούμενα. Στα αποτελέσματα θα εμφανίζονται τονισμένοι με κίτρινο χρώμα οι οροί που συμπεριλαμβάνονται στην αναζήτηση, κάνοντας με αυτό τον τρόπο πιο εύκολη την εντόπιση του σημείου που ενδιαφέρει το χρήστη μέσα σε ένα άρθρο. Τα αποτελέσματα των αναζητήσεων του χρήστη θα εμφανίζονται σε μια στήλη (like Google) με τον τίτλο του άρθρου και με την περίληψη του (abstract πεδίο του άρθρου). Τα αποτελέσματα θα παρουσιάζονται σε διάταξη με βάση την συνάφεια τους με το ερώτημα που έθεσε εκείνη την στιγμή ο χρήστης. Συνεπώς , στις πρώτες σελίδες των αποτελεσμάτων θα εμφανίζονται τα άρθρα , τα οποία το περιεχόμενο τους είναι πιο κοντά στην αναζήτηση.

Ακόμη, ο χρήστης θα μπορεί να έχει την επιλογή ταξινόμησης των αποτελεσμάτων ανάλογα με το έτος δημοσίευσης του άρθρου. Η λειτουργία αυτή θα ενεργοποιείται επιλέγοντας τον αντίστοιχο κουμπί. Τα αποτελέσματα θα εμφανίζονται κατά φθίνουσα σειρά, δηλαδή πρώτα θα είναι τα άρθρα που έχουν δημοσιευτεί πιο πρόσφατα και όσο προχωράει ο χρήστης θα βλέπει πιο παλιά. Έτσι, ο χρήστης θα μπορεί πιο γρήγορα να εντοπίσει τα νέα άρθρα ,αφού θα εμφανίζονται μαζεμένα στις πρώτες σελίδες των αποτελεσμάτων.

Υλοποίηση user interface:

Το UI υλοποιείτε με τη χρήση της βιβλιοθήκης JavaSwing για τη δημιουργία του γραφικού περιβάλλοντος. Η JavaSwing είναι μια βιβλιοθήκη γραφικών χρήστη (GUI) για τη γλώσσα προγραμματισμού Java. Η JavaSwing παρέχει ένα σύνολο στοιχείων GUI, όπως πλαίσια, κουμπιά, πεδία κειμένου, μενού τα οποία μπορούν να χρησιμοποιηθούν για τη δημιουργία διαδραστικών εφαρμογών. Ο τρόπος αυτός επιλέχθηκε καθώς συνδυάζει ευκολία και καλή τεκμηρίωση του API , διευκολύνοντας τη διαδικασία ανάπτυξης γραφικών εφαρμογών.