Πανεπιστήμιο Ιωαννίνων
Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Εργασία: Μηχανή αναζήτησης πληροφορίας από επιστημονικά άρθρα

Καταληκτικές Ημερομηνίες

Τετάρτη 17 Απριλίου 2024	Σύντομη περιγραφή του σχεδιασμού και της συλλογής εγγράφων
Τετάρτη 22 Μαΐου 2024	Παράδοση εργασίας
Εβδομάδα 27 Μαΐου	Προφορική εξέταση (οι ακριβείς ημέρες και ώρες θα ανακοινωθούν αργότερα)

Η εργασία μπορεί να γίνει σε ομάδες έως 2 ατόμων.
Η εργασία μετράει σε ποσοστό 50% στο βαθμό σας στο μάθημα.

Η εργασία αφορά στο σχεδιασμό και υλοποίηση ενός συστήματος αναζήτησης πληροφορίας από επιστημονικά άρθρα. Για την υλοποίηση, θα χρησιμοποιήστε τη βιβλιοθήκη **Lucene** https://lucene.apache.org/, μια βιβλιοθήκη ανοικτού κώδικα για την κατασκευή μηχανών αναζήτησης κειμένου.

Συλλογή εγγράφων (corpus). Θα χρησιμοποιείστε επιστημονικά άρθρα την παρακάτω συλλογή από το kaggle**:**

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

Η συλλογή πρέπει να περιλαμβάνει *τουλάχιστον* 200 επιστημονικά άρθρα.

Ανάλυση κειμένου και κατασκευή ευρετηρίου. Η Lucene παρέχει τη δυνατότητα για stemming, απαλοιφή stop words, επέκταση συνωνύμων, κλπ.

Επίσης, κάποιες λειτουργίες, όπως η διόρθωση τυπογραφικών λαθών, ή η επέκταση ακρωνύμων, μπορούν να γίνουν εναλλακτικά κατά τη διάρκεια της αναζήτησης (τροποποιώντας το ερώτημα).

Επιλέξτε το είδος της ανάλυσης που θεωρείτε κατάλληλο και εξηγείστε την επιλογή σας.

Αναζήτηση. Το σύστημά σας θα πρέπει να υποστηρίζει (α) αναζήτηση με λέξεις κλειδιά και (β) αναζήτηση πεδίου, δηλαδή, την εμφάνιση όρων σε συγκεκριμένα πεδία (στον τίτλο, abstract, full text) (γ) έναν ακόμα τρόπο αναζήτησης της επιλογής σας

Επίσης, να διατηρεί πληροφορία από την ιστορία των αναζητήσεων. Χρησιμοποιείστε αυτήν την πληροφορία για να προτείνετε εναλλακτικά ερωτήματα.

Προαιρετικό ερώτημα: συμπεριλάβατε στο έγγραφο και το όνομα του συγγραφέα και υποστηρίξτε αναζήτηση πεδίου και με το όνομα.

Παρουσίαση Αποτελεσμάτων. Το σύστημα σας θα πρέπει να παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους με το ερώτημα όπως αυτή υπολογίζεται από την Lucene.

Επιπρόσθετα, θα πρέπει
(1) Να παρουσιάζει τα αποτελέσματα ανά 10, με δυνατότητα στο χρήστη να προχωρήσει στα επόμενα.
(2) Οι λέξεις κλειδιά να παρουσιάζονται τονισμένες στο αποτέλεσμα.
(3) Να παρέχει δυνατότητα αναδιάταξης των αποτελεσμάτων με βάση τη χρονιά που εμφανίστηκε το άρθρο.

Θα βαθμολογηθεί και η ευχρηστία στην παρουσίαση των αποτελεσμάτων.