Εργασία Ανάκτησης Πληροφορίας

Δημιουργία μηχανής αναζήτησης ακαδημαϊκών εργασιών

Η εργασία αφορά στην πρακτική εξάσκηση και εμπειρία στη δημιουργία μιας απλής μηχανής αναζήτησης με σκοπό την κατανόηση των θεμελιωδών εννοιών της ανάκτησης πληροφορίας, της ευρετηρίασης, της κατάταξης και της αναζήτησης πληροφορίας, καθώς και πρακτικές δεξιότητες στην επεξεργασία φυσικής γλώσσας και την εφαρμογή αλγορίθμων αναζήτησης.

Στόχοι της εργασίας:

- Ο σχεδιασμός και υλοποίηση ενός συστήματος ανάκτησης εγγράφων που μπορεί να ευρετηριάσει και να αναζητήσει αποτελεσματικά σε μια συλλογή εγγράφων κειμένου.
- Η ανάπτυξη αλγορίθμων ανάκτησης και η αξιολόγηση της απόδοσή τους χρησιμοποιώντας κοινές μετρήσεις αξιολόγησης, όπως ακρίβεια, ανάκληση και βαθμολογία F1.
- Η παροχή μιας φιλικής προς το χρήστη διεπαφής για τους χρήστες να εισάγουν ερωτήματα και να ανακτούν σχετικά έγγραφα.
- Η απόκτηση πρακτικής εμπειρίας σε διάφορες τεχνικές ανάκτησης πληροφοριών, όπως Boolean retrieval, Vector Space Model και Probabilistic retrieval models.

Περιγραφή εργασίας
Υλοποιείστε μια μηχανή αναζήτησης που ανακτά ακαδημαϊκές εργασίες (research papers) με βάση τα ερωτήματα των χρηστών. Η μηχανή αναζήτησης θα έχει δύο κύριες συνιστώσες:

α. έναν ανιχνευτή ιστού για τη συλλογή ακαδημαϊκών εργασιών και
β. μια διεπαφή αναζήτησης για τους χρήστες όπου θα αναζητούν και θα ανακτούν έγγραφα.

Βήμα 1. Σταχυολογητής (Web Crawler):

α. Επιλέξτε έναν ιστότοπο-στόχο ή ένα αποθετήριο ακαδημαϊκών εργασιών (π.χ. arXiv, PubMed ή αποθετήριο πανεπιστημίου).
β. Υλοποιήστε έναν web crawler σε Python (π.χ. με BeautifulSoup) για τη συλλογή μεταδεδομένων ακαδημαϊκών εργασιών (τίτλος, συγγραφείς, περίληψη, ημερομηνία δημοσίευσης κ.λπ.) από την επιλεγμένη πηγή.
γ. Αποθηκεύστε τα δεδομένα που συλλέγονται σε δομημένη μορφή, όπως JSON ή CSV.

Βήμα 2. Προεπεξεργασία κειμένου (Text Processing):
Κάντε προεπεξεργασία του κειμενικού περιεχομένου των ακαδημαϊκών εργασιών για την προετοιμασία τους για ευρετηρίαση και αναζήτηση. Αυτό μπορεί να περιλαμβάνει εργασίες όπως tokenization, stemming/lemmatization και stop-word removal και αφαίρεση ειδικών χαρακτήρων (removing special characters).

Βήμα 3. Ευρετήριο (Indexing):

α. Δημιουργήστε μια ανεστραμμένη δομή δεδομένων ευρετηρίου (inverted index) για την αποτελεσματική αντιστοίχιση όρων στα έγγραφα στα οποία εμφανίζονται.
β. Εφαρμόστε μια δομή δεδομένων για την αποθήκευση του ευρετηρίου.

Βήμα 4. Μηχανή αναζήτησης (Search Engine):

α. Αναπτύξτε μια διεπαφή χρήστη για την αναζήτηση ακαδημαϊκών εργασιών χρησιμοποιώντας την Python (π.χ. μια διεπαφή γραμμής εντολών ή μια απλή διεπαφή ιστού).
β. Υλοποιήστε πολλαπλούς (τουλάχιστον 3) αλγόριθμους ανάκτησης, όπως Boolean retrieval, Vector Space Model (VSM) και Probabilistic retrieval models (π.χ. Okapi BM25) για να ανακτήσετε σχετικές εργασίες με βάση τα ερωτήματα των χρηστών. Ο χρήστης θα μπορεί να επιλέγει τον αλγόριθμο ανάκτησης.
γ. Επιτρέψτε στους χρήστες να φιλτράρουν τα αποτελέσματα αναζήτησης με διάφορα κριτήρια, όπως η ημερομηνία δημοσίευσης ή ο συγγραφέας.

Επεξεργασία ερωτήματος (Query Processing): Αναπτύξτε ένα module επεξεργασίας ερωτημάτων που θα προεπεξεργάζεται τα ερωτήματα που λαμβάνει από τον χρήστη, τα αναλύει και ανακτά σχετικά έγγραφα χρησιμοποιώντας το ανεστραμμένο ευρετήριο. Μπορείτε να χρησιμοποιήσετε απλά ερωτήματα βάσει λέξεων (όρων). Οι χρήστες θα πρέπει να μπορούν να αναζητούν έγγραφα χρησιμοποιώντας μία ή περισσότερες λέξεις. Το module θα λαμβάνει ερωτήματα χρηστών τα οποία τα γίνονται tokenized και θα εκτελεί λειτουργίες Boolean (AND, OR και NOT).

Κατάταξη αποτελεσμάτων (Ranking): Εφαρμόστε έναν βασικό αλγόριθμο κατάταξης. Μπορείτε να ξεκινήσετε με έναν απλό αλγόριθμο κατάταξης TF-IDF (Term Frequency-Inverse Document Frequency) και αργότερα μπορείτε να συμπεριλάβετε πιο προηγμένες τεχνικές κατάταξης. Ταξινομήστε και παρουσιάστε τα αποτελέσματα αναζήτησης σε φιλική προς το χρήστη μορφή.,

Βήμα 5. Αξιολόγηση συστήματος:

Παρουσιάστε στο επόμενο βήμα (τεκμηρίωση) αναλυτικά τη μεθοδολογία (σύνολα δεδομένων, μετρικές, βιβλιοθήκες python κλπ) που θα ακολουθούσατε για την αξιολόγηση της αποτελεσματικότητα της μηχανής αναζήτησης που υλοποιήσατε.

Βήμα 6. Αναφορά και τεκμηρίωση:

Γράψτε μια ολοκληρωμένη **αναφορά και τεκμηρίωση** που εξηγεί το σχεδιασμό, την υλοποίηση και την αξιολόγηση της μηχανής αναζήτησης (δεν χρειάζεται επεξήγηση κώδικα). Αναφέρατε τις δυσκολίες που αντιμετωπίσατε και τις προτεινόμενες βελτιώσεις. Συμπεριλάβετε μελέτες περιπτώσεων από ερωτήματα χρηστών με εικόνες screenshots, για επίδειξη της λειτουργικότητας της μηχανής αναζήτησης.

Κριτήρια αξιολόγησης εργασίας:
Η αξιολόγηση της εργασίας θα γίνει με βάση την ποιότητα και πληρότητα τόσο της αναφοράς όσο και της μηχανής αναζήτησης, την αποτελεσματικότητα του αλγορίθμου κατάταξης, τις μετρήσεις αξιολόγησης αλλά και την ικανότητα τους να εξηγήσετε τις έννοιες και τις αποφάσεις που λάβατε κατά τη διάρκεια του εργασίας υπερασπίζοντας τις σχεδιαστικές τους επιλογές και την απόδοση στην αξιολόγηση τους συστήματος ανάκτησης.

Εξέταση:
Η εξέταση για το εργαστηριακό μέρος θα γίνει μαζί με την εξέταση του θεωρητικού, στην εκάστοτε εξεταστική. Ο βαθμός από την εργασία θα προσμετρήσει μόνο εφόσον απαντηθούν οι ερωτήσεις που θα αντιστοιχούν στο εργαστηριακό μέρος.

Την εργασία μπορεί να την αναλαμβάνει ομάδα μέχρι 2 άτομα

Παραδοτέα: *Αναφορά σε μορφή* ***.pdf*** *με όνομα αρχείου* **`Surname1 AM1-Surname2 AM2.pdf`** *και κώδικας* με τη μηχανή αναζήτησης ολοκληρωμένη σε αρχείο **zip**

Ημ/νία κατάθεσης (από το **ένα μόνο μέλος** της ομάδας) στο eclass: **22/01/2024**

χωρίς δυνατότητα εκπρόθεσμης υποβολής

Απορίες θα λύνονται μόνο κατά τη διάρκεια του εργαστηριακού μαθήματος και όχι με e-mail