

Εργασία: Μηχανή αναζήτησης πληροφορίας από ειδησεογραφικά άρθρα

Καταληκτικές Ημερομηνίες

Δευτέρα 14 Απριλίου 2025	Σύντομη περιγραφή του σχεδιασμού και της συλλογής εγγράφων
Παρασκευή 23 Μαΐου 2025	Παράδοση εργασίας
Εβδομάδα 26 Μαΐου 2025	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.

1. Εισαγωγή

Η εργασία αφορά στο σχεδιασμό και υλοποίηση ενός συστήματος αναζήτησης πληροφορίας από ειδησεογραφικά άρθρα.

Μπορείτε να κατεβάσετε τα άρθρα από το kaggle:

<https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning/data>

Η συλλογή σας θα πρέπει να περιλαμβάνει τουλάχιστον 1.000 από αυτά τα άρθρα.

Θα χρησιμοποιήσετε τη βιβλιοθήκη **Lucene** <https://lucene.apache.org/>, μια βιβλιοθήκη ανοικτού κώδικα για την κατασκευή μηχανών αναζήτησης από αρχεία κειμένου.

Το κύριο μέρος της εργασίας αφορά αναζήτηση με λέξεις κλειδιά. Η αναζήτηση αυτή θα επεκταθεί για την υποστήριξη διανυσματικής αναζήτησης.

Θα βαθμολογηθεί και η ευχρηστία του συστήματός σας.

2. Αναζήτηση με λέξεις κλειδιά [80% του βαθμού]

Η μηχανή αναζήτησης θα πρέπει να υποστηρίζει τα παρακάτω:

(α) **Ανάλυση κειμένου και κατασκευή ευρετηρίου.** Η Lucene παρέχει τη δυνατότητα για

stemming, απαλοιφή stop words, επέκταση συνωνύμων, κλπ. 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 εξηγήστε την επιλογή σας.

(β) Αναζήτηση. Το σύστημά σας θα πρέπει να υποστηρίζει: (α) αναζήτηση με λέξεις κλειδιά και (β) αναζήτηση πεδίου, δηλαδή, εμφάνιση όρων σε συγκεκριμένα πεδία (π.χ., επικεφαλίδα, συγγραφέας, ημερομηνία).

Επιπρόσθετα, το σύστημά σας θα πρέπει να διατηρεί πληροφορία από την ιστορία των αναζητήσεων. Συγκεκριμένα, το σύστημα θα πρέπει να αποθηκεύει τα 20 προηγούμενα ερωτήματα μαζί με τα 5 πρώτα άρθρα της απάντησης. Χρησιμοποιήστε αυτή την πληροφορία με έναν κατάλληλο τρόπο ώστε να βελτιώσετε την αναζήτηση.

(γ) Παρουσίαση Αποτελεσμάτων. Το σύστημά 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.

(2) Να εμφανίζει τις λέξεις κλειδιά τονισμένες στο αποτέλεσμα.

(3) Να παρέχει τη δυνατότητα αναδιάταξης των αποτελεσμάτων αλφαβητικά με βάση τον τίτλο τους.

3. Διανυσματική Αναζήτηση [20% του βαθμού]

Στόχος είναι η υποστήριξη διανυσματικής αναζήτησης.

Ο βασικός τρόπος διανυσματικής αναζήτησης που υποστηρίζει η Lucene είναι η δημιουργία ενός διανύσματος ανά έγγραφο. Αν θέλετε, μπορείτε να χωρίσετε το έγγραφο σε μικρότερα τμήματα (chunks) και να δημιουργήσετε διανύσματα για αυτά, αλλά θα χρειαστεί ειδική διαχείριση και δεν ζητείται.

Θα πρέπει πρώτα να δημιουργήσετε διανύσματα (embeddings) με χρήση κάποιου υπάρχοντος μοντέλου (π.χ.μ all-MiniLM-L6-v2) και μετά να τα προσθέσετε σε αντίστοιχο πεδίο στο ευρετήριο (KnnVectorField).

Στη συνέχεια, θα δημιουργείται embeddings για κάθε ερώτημα και θα χρησιμοποιείτε το KnnVectorQuery για την επεξεργασία τους.

Χρήσιμη βιβλιοθήκη για τον υπολογισμό των embeddings: Deep Java Library (DJL)