

## Εργασία: Μηχανή αναζήτησης άρθρων σχετικών με τον COVID-19

### Αρχική περιγραφή

#### Καταληκτικές Ημερομηνίες

**Δευτέρα 19 Απριλίου 2021**, Σύντομη περιγραφή σχεδιασμού και συλλογή δεδομένων

**Δευτέρα 24 Μαΐου 2021**, Παράδοση εργασίας

**Προφορική εξέταση στο διάστημα 25 με 31 Μαΐου (οι ακριβείς ημέρες 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 άρθρων σχετικών με το Covid-19.

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

**Συλλογή εγγράφων (corpus).** Αρχικά, πρέπει να συλλέξετε τα έγγραφα που θα αποτελούν τη συλλογή σας. Το έγγραφο σας θα είναι έγγραφο σχετικά με το covid-19.

Μπορείτε να κατασκευάσετε τη συλλογή από τα άρθρα με όποιο τρόπο θέλετε, όπως να χρησιμοποιείτε έτοιμες συλλογές εγγράφων, ή να κατεβάσετε ιστοσελίδες (π.χ., με scrapping), ή να συλλέξετε δημοσιεύσεις από κοινωνικά δίκτυα.

Η συλλογή πρέπει να περιλαμβάνει τουλάχιστον 500 έγγραφα στην περίπτωση άρθρων και 2.000 έγγραφα στην περίπτωση σύντομων κειμένων από κοινωνικών δίκτυα.

**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 εξηγήστε την επιλογή σας.

**Αναζήτηση.** Το σύστημα σας θα πρέπει να υποστηρίζει αναζήτηση εγγράφων με λέξεις κλειδιά.

Επιπρόσθετα, θα πρέπει

- (1) Να υποστηρίζει και άλλα είδη ερωτήσεων, για παράδειγμα αναζήτηση πεδίου, δηλαδή, την εμφάνιση όρων σε συγκεκριμένα πεδία (πχ. στον τίτλο).
- (2) Να διατηρεί πληροφορία για την ιστορία των αναζητήσεων. Χρησιμοποιείστε αυτήν την πληροφορία για να προτείνετε εναλλακτικά ερωτήματα.
- (3) Να χρησιμοποιεί embedding για να βελτιώσει την αναζήτηση.

**Παρουσίαση Αποτελεσμάτων.** Το σύστημα σας θα πρέπει να παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους με το ερώτημα.

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) Δυνατότητα ομαδοποίησης των αποτελεσμάτων με βάση κάποιο κριτήριο (πχ, συναίσθημα (sentiment), την τιμή κάποιου πεδίου (π.χ., χρόνο δημοσίευσης) κλπ.