Εργασία: Μηχανή αναζήτησης πληροφορίας για τραγούδια

Καταληκτικές Ημερομηνίες

Παρασκευή 7 Απριλίου 2022	Σύντομη περιγραφή του σχεδιασμού και της συλλογής δεδομένων
Παρασκευή 19 Μαΐου 2022	Παράδοση εργασίας
Εβδομάδα 22 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 στίχων τραγουδιών και άλλης πληροφορίας σχετικής με μουσικούς και τραγούδια. 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). Αρχικά, πρέπει να συλλέξετε τα έγγραφα που θα αποτελούν τη συλλογή σας. Το έγγραφα σας θα είναι έγγραφα σχετικά με τραγούδια, συλλογές τραγουδιών ή μουσικούς.

Μπορείτε να κατασκευάσετε τη συλλογή από τα άρθρα με όποιο τρόπο θέλετε, όπως να χρησιμοποιείστε έτοιμες συλλογές εγγράφων, ή να κατεβάσετε ιστοσελίδες (π.χ., με scrapping), ή να συλλέξετε δημοσιεύσεις από κοινωνικά δίκτυα. Τα έγγραφα θα πρέπει απαραίτητα να περιέχουν κείμενο.

Η συλλογή πρέπει να περιλαμβάνει *τουλάχιστον* 500 έγγραφα, για παράδειγμα στίχους από τουλάχιστον 500 τραγούδια.

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 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, όνομα δημιουργού).
(2) Να διατηρεί πληροφορία για την ιστορία των αναζητήσεων. Χρησιμοποιείστε αυτήν την πληροφορία για να προτείνετε εναλλακτικά ερωτήματα.

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) Να παρέχει δυνατότητα ομαδοποίησης των αποτελεσμάτων με κάποιο κριτήριο που θα ορίσετε εσείς.

Προαιρετικό Ερώτημα. Το σύστημα θα πρέπει να παρέχει τη δυνατότητα σημασιολογικής ανάκτησης (λεπτομερής εκφώνηση θα δοθεί τις επόμενες εβδομάδες).