Εργασία: Μηχανή αναζήτησης Wikipedia άρθρων

(αρχική περιγραφή, θα ακολουθήσει πιο αναλυτική)

Καταληκτικές Ημερομηνίες
Δευτέρα 26 Απριλίου 2020, Σύντομη περιγραφή σχεδιασμού και συλλογή δεδομένων
Παρασκευή 29 Μαΐου 2020, Παράδοση εργασίας
Όταν ανοίξει το Πανεπιστήμιο, Εξέταση εργασίας

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 της wikipedia.

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)

Το δεδομένα σας θα είναι άρθρα της Wikipedia.

Μπορείτε να τα συλλέξετε τα άρθρα με όποιο τρόπο θέλετε (web scrapping, από κάποιο archive, κλπ).

Ελάχιστες απαιτήσεις: 5000 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 σύστημα σας θα πρέπει να επιτρέπει αναζήτηση άρθρων με λέξεις κλειδιά.

Θα πρέπει να υποστηρίζονται διάφορα είδη ερωτήσεων.

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

Τέλος, χρησιμοποιήστε embedding για να βελτιώσετε τα αποτελέσματα της αναζήτησης.

Παρουσίαση Αποτελεσμάτων

Παρουσιάστε τα αποτελέσματα ανά 10, με δυνατότητα στο χρήστη να προχωρήσει στα επόμενα.