Εργασία: Μηχανή αναζήτησης Wikipedia άρθρων

Καταληκτικές Ημερομηνίες
Τετάρτη 29 Απριλίου 2020, 5μμ Σύντομη περιγραφή σχεδιασμού και συλλογή δεδομένων
Παρασκευή 29 Μαΐου 2020, 5μμ 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.
Οι καταληκτικές ημερομηνίες είναι αυστηρές, *δε θα γίνουν δεκτές αργοπορημένες παραδόσεις ασκήσεων*

Σύντομη περιγραφή

Η εργασία αφορά στο σχεδιασμό και υλοποίηση ενός συστήματος αναζήτησης άρθρων της wikipedia.

Αρχικά, θα συλλέξετε έναν αριθμό από άρθρα της. Wikipedia. Αυτά τα άρθρα θα αποτελούν τη συλλογή σας (corpus).

Στη συνέχεια, θα υλοποιήσετε μια μηχανή αναζήτησης αυτών των άρθρων.

Συγκεκριμένα, ο χρήστης θα θέτει ερωτήματα. Το αποτέλεσμα θα είναι τα συναφή με το ερώτημα άρθρα της συλλογής σας σε διάταξη με βάση τη συνάφεια τους με το ερώτημα.

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

Αναλυτική περιγραφή

Συλλογή εγγράφων (corpus). Αρχικά, πρέπει να συλλέξετε τα έγγραφα που θα αποτελούν τη συλλογή σας. Το έγγραφα σας θα είναι άρθρα της Wikipedia.

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

Απαιτούμενος αριθμός: ~~5000~~ 100 άρθρα. Bonus αν η συλλογή είναι μεγαλύτερη.

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. Η *ελάχιστη απαίτηση* από το σύστημα σας είναι να υποστηρίζει αναζήτηση άρθρων με λέξεις κλειδιά.

Επιπρόσθετα, μια άριστη εργασία, 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. Η *ελάχιστη απαίτηση* από το σύστημα σας είναι να παρουσιάζει τα αποτελέσματα σε διάταξη με βάση τη συνάφεια τους με το ερώτημα.

Επιπρόσθετα, μια άριστη εργασία, 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 διαφορετικής διάταξης των αποτελεσμάτων