Εργασία: Μηχανή αναζήτησης επιχειρήσεων
(αρχική περιγραφή)

Καταληκτικές Ημερομηνίες
Παρασκευή 19 Απριλίου 2019, Σύντομη περιγραφή σχεδιασμού
Παρασκευή 24 Μαΐου 2019, Παράδοση εργασίας
Δευτέρα 27 ή/και Τρίτη 28 Μαΐου 2019, 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 επιχειρήσεων.

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 πληροφορίες για επιχειρήσεις από το σύστημα Yelp (https://www.yelp.com/).

Ως πρώτο βήμα δημιουργείστε τη συλλογή σας κατεβάζοντας δεδομένα από το Yelp Open Dataset (https://www.yelp.com/dataset).

Διαλέξτε ένα υποσύνολο των διαθέσιμων δεδομένων τα οποία να αφορούν επιχειρήσεις σε **μία** από τις 10 πόλεις που καλύπτει το dataset.

Ελάχιστες απαιτήσεις:

- 10000 επιχειρήσεις
- 1000000 κριτικές και υποδείξεις για αυτές τις επιχειρήσεις.

Ανάλυση 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 επιχειρήσεων *τουλάχιστον* με βάση:

(1) Το όνομα της επιχείρησης,
(2) Την κατηγορία της επιχείρησης,
(3) Λέξεις κλειδιά και φράσεις (phrase queries) που εμφανίζονται:
   a. στο *πλήρες κείμενο* των κριτικών για την επιχείρηση (για παράδειγμα επιχειρήσεις των οποίων οι κριτικές περιλαμβάνουν τη λέξη «sesame»),
   b. στο *πλήρες κείμενο* των υποδείξεων για την επιχείρηση (για παράδειγμα επιχειρήσεις των οποίων οι υποδείξεις περιλαμβάνουν τη λέξη «sesame»),
(4) Συνδυασμό των παραπάνω με χρήση Boolean queries.

Παρουσίαση Αποτελεσμάτων

Διάταξη αποτελεσμάτων
Εξηγείστε τον τρόπο με τον οποίο γίνεται η διάταξη των αποτελεσμάτων.

Επίσης, να παρέχετε η δυνατότητα διάταξης με βάση τον αριθμό των αστεριών. Σε περίπτωση ισοβαθμίας στον αριθμό των αστεριών, να προηγείται η επιχείρηση με το μεγαλύτερο αριθμό κριτικών.

Άλλες Απαιτήσεις
Στο αποτέλεσμα, να γίνεται επισήμανση (highlight) των όρων αναζήτησης.

Επιπρόσθετη λειτουργικότητα

Το σύστημα σας θα πρέπει να διατηρεί πληροφορία για την ιστορία των αναζητήσεων (π.χ., clickthrough-rate, δημοφιλείς ερωτήσεις, κλπ).

Χρησιμοποιείστε αυτήν την πληροφορία για:
(1) να αναδιατάξετε τα αποτελέσματα της αναζήτησης, και
(2) να προτείνετε εναλλακτικά ερωτήματα.