Εργασία (προκαταρτική περιγραφή)

Καταληκτικές Ημερομηνίες Παράδοσης
Παρασκευή 11 Απριλίου 2014, Περιγραφή Αρχικού Σχεδιασμού
Τετάρτη 14 Μαΐου 2014 Παράδοση 1ης Φάσης (ερωτήματα 1-4)
Τετάρτη 28 Μαΐου 2014, Παράδοση και Εξέταση Τελικής Εργασίας (ερωτήματα 1-6)

Η εργασία μπορεί να γίνει σε ομάδες έως 2 ατόμων.
Η εργασία μετράει σε ποσοστό 65% στο βαθμό σας στο μάθημα.

Η εργασία αφορά στο σχεδιασμό και υλοποίηση ενός σύστημα ανάκτησης πληροφορίας για μια συλλογή εγγράφων. Τα έγγραφα θα είναι αποθηκευμένα στο δίσκο.

Για την υλοποίηση, θα χρησιμοποιήστε το σύστημα Lucene[1].

Μπορείτε να διαλέξετε κάποια από τις παρακάτω συλλογές εγγράφων

- άρθρα εγκυκλοπαίδειας, π.χ., κάποιο υποσύνολο από τα άρθρα της wikipedia[2]
- κριτικές, πχ κάποια υποσύνολο από το yelp[3]
- περιεχόμενο σελίδων του Facebook, πχ αφού κατεβάσετε (όλο ή τμήμα) του περιεχομένου της σελίδας σας
- συλλογές από νέα, όπως the 20 newsgroup dataset[4]
- κάποια συλλογή από το TREC[5]

Το σύστημα σας θα πρέπει:

1. Να υποστηρίζει ανάλυση (parsing) και κάποια γλωσσική επεξεργασία. Κατά ελάχιστον θα πρέπει να αναγνωρίζει τα tokens των εγγράφων και να παράγει τους όρους. Επιπρόσθετη επεξεργασία όπως κανονικοποίηση των όρων σε κλάσεις ισοδυναμίας, stemming, κλπ θα μετρήσουν θετικά. [Βαθμοί 1.5/10].

2. Να κατασκευάζει ένα κατάλληλο ευρετήριο [Βαθμοί 1.5/10].

3. Να απαντά σε ερωτήσεις ελεύθερου κειμένου χρησιμοποιώντας τη διανυσματική αναπαράσταση. Ο χρήστης θα δίνει ως είσοδο έως 6 όρους και θα επιστρέφονται τα καλύτερα έγγραφα σε διάταξη [Βαθμοί 2/10].

4. Πέρα από τη βασική λειτουργικότητα, θα πρέπει να υποστηρίζονται ερωτήματα εγγύτητα (δηλαδή, στη διάταξη των αποτελεσμάτων να προηγούνται έγγραφα στα οποία οι όροι της ερώτησης να βρίσκονται είναι πιο κοντά καθώς και διόρθωση ορθογραφικών λαθών [Βαθμοί 2/10].

5. Υποστήριξη γραφικού περιβάλλοντος για τη διατύπωση των ερωτημάτων και την προβολή των αποτελεσμάτων. Ο χρήστης θα μπορεί να επιλέξει κάποιες από τις απαντήσεις και σε αυτήν την

[1] https://lucene.apache.org/
[2] http://en.wikipedia.org/wiki/Wikipedia:Database_download
[3] http://www.yelp.com/dataset_challenge/
[4] http://qwone.com/~jason/20Newsgroups/
[5] http://trec.nist.gov/

περίπτωση θα βλέπει το αντίστοιχο κείμενο με τονισμένους τους όρους της ερώτησης. Προαιρετικά μαζί με την αρχική απάντηση θα παρέχεται και μια περίληψη του κειμένου [Βαθμοί 1.5/10].

6. Διάταξη των εγγράφων που θα χρησιμοποιεί επιπρόσθετη πληροφορία, όπως πληροφορία συνδέσεων ή δημοτικότητας [1.5/10].