Εργασία (προκαταρτική περιγραφή)

Καταληκτική Ημερομηνία Παράδοσης:

Τετάρτη 24 Απριλίου 2013, μια 2-σέλιδη περιγραφή του συστήματος όπου θα αναφέρεται η συλλογή και ο αρχικός σχεδιασμός

Τελική παράδοση την ημέρα της εξέτασης του μαθήματος

Σχεδιάστε και υλοποιήστε ένα Σύστημα Ανάκτησης Πληροφορίας για μια ειδική συλλογή εγγράφων.

Μπορείτε να χρησιμοποιήστε οποιαδήποτε γλώσσα προγραμματισμού.

Η συλλογή θα πρέπει να περιέχει τουλάχιστον 40 έγγραφα που θα περιέχουν συνολικά τουλάχιστον 8,000 tokens. Τα έγγραφα μπορεί να είναι emails, tweets, απλά έγγραφα κειμένου. Κάθε έγγραφο θα είναι αποθηκευμένο σε διαφορετικό αρχείο στο δίσκο.

Στη συνέχεια περιγράφονται οι ελάχιστες απαιτήσεις από το σύστημα.

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. [Βαθμοί 2/10]. Επιπρόσθετη επεξεργασία όπως κανονικοποίηση των όρων σε κλάσεις ισοδυναμίας, stemming, κλπ θα μετρήσουν θετικά.

2. Να κατασκευάζει ένα αντεστραμμένο ευρετήριο για τους όρους. Κατά ελάχιστον οι καταχωρήσεις θα πρέπει να διατηρούν το DocID του εγγράφου και τη συχνότητα εμφάνισης του όρου στο έγγραφο. Το ευρετήριο (λεξικό και λίστες καταχωρήσεων) θα αποθηκεύεται στο δίσκο και θα φορτώνεται κάθε φορά που θα αρχίζει το πρόγραμμα σας [Βαθμοί 4/10]. Ευρετήρια εγγύτητας (positional indexes) και άλλες επεκτάσεις θα μετρήσουν θετικά.

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 3/10]. Υποστήριξη ερωτημάτων εγγύτητας, φράσεων και διόρθωση ορθογραφικών λαθών θα μετρήσουν θετικά.

4. Να απαντά σε ερωτήσεις Boolean με χρήση AND και OR τελεστών. Κάθε ερώτημα μπορεί να έχει μόνο AND ή μόνο OR. [Βαθμός 1/10]. Υποστήριξη ερωτημάτων με συνδυασμό AND, OR και NOT θα μετρήσει θετικά.