ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2017-2018
ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ
ΜΑΘΗΜΑ: ΜΗΧΑΝΕΣ ΑΝΑΖΗΤΗΣΗΣ

Project
Προθεσμία Υποβολής: 21/01/2018 23:59

Η εργασία είναι ομαδική και κάθε ομάδα θα αποτελείται από 2 άτομα. Τα παραδοτέα σας θα τα τοποθετήσετε σε ένα zip ή rar αρχείο και θα τα ανεβάσετε στη σελίδα του μαθήματος στο eclass.

1ο Τμήμα

Θα πειραματιστείτε στην κατασκευή των inverted indexes για μια συλλογή εγγράφων. Η είσοδος στο σύστημά σας θα είναι η συλλογή των εγγράφων ενώ η έξοδος θα είναι το ευρετήριο των όρων όπως έχουμε ήδη συζητήσει στις διαλέξεις του μαθήματος.

Η υλοποίηση της εργασίας είτε θα γίνει με χρήση κάποιου έτοιμου πακέτου που έχει δημιουργηθεί για το σκοπό αυτό ή δημιουργώντας κώδικα από την αρχή σε μια γλώσσα προγραμματισμού της αρεσκείας σας. Παράδειγμα εργαλείου που μπορεί να χρησιμοποιήσετε είναι: Terrier IR Platform (University of Glasgow - Java). Επίσης, εύκολα μπορείτε να υιοθετήσετε το API του εργαλείου στον κώδικά σας και να ανακτήσετε πληροφορίες σχετικά με το indexing που θα κάνετε (π.χ. δείτε εδώ http://terrier.org/docs/v4.2/quickstart-integratedsearchdisk.html). Κατά την κατασκευή των ευρετηρίων θα πρέπει να απορρίψετε τα stop words όπως αυτά φαίνονται εδώ http://www.lextek.com/manuals/onix/stopwords1.html, https://www.ranks.nl/stopwords. Επίσης, θα πρέπει να ενεργοποιήσετε / υλοποιήσετε λογισμικό για stemming.

Να κατασκευάσετε ευρετήρια για τις ακόλουθες συλλογές εγγράφων: ClueWeb09-B, ClueWeb12, Cranfield.

Το παραδοτέο σας θα είναι οι εντολές του εργαλείου / κώδικας που χρησιμοποιήσατε καθώς και documentation με μέγεθος 1-2 σελίδων το μέγιστο. Στο παραδοτέο να παρουσιάσετε και κάποια στατιστικά για τα ευρετήρια που θα δημιουργήσετε.

2ο Τμήμα

Στο δεύτερο τμήμα της εργασίας καλείστε να υλοποιήσετε ένα απλό κώδικα ο οποίος θα αναθέτει ερωτήματα προς τη μηχανή αναζήτησης και τα ευρετήρια που κατασκευάσατε στο 1ο Τμήμα. Ακολουθώντας τις οδηγίες που παρέχονται εδώ http://terrier.org/docs/v4.2/quickstart-integratedsearchdisk.html θα καταγράψετε σε ένα αρχείο κειμένου ένα σύνολο τυχαίων ερωτημάτων (π.χ. 100 ερωτήματα – συνδυασμούς όρων [1,2,3,.. όροι]). Στη συνέχεια θα πειραματιστείτε με διάφορα μοντέλα retrieval. Θα πρέπει να πάρετε αποτελέσματα και στατιστικά για τα εξής μοντέλα: bm25, idf, tf, tf_idf, WeightingModel. Για το indexing υιοθετείστε τις ίδιες συλλογές δεδομένων όπως και στο 1ο τμήμα.

Το παραδοτέο σας θα είναι οι εντολές του εργαλείου / κώδικας που χρησιμοποιήσατε καθώς και documentation με μέγεθος 1-2 σελίδων το μέγιστο. Στο παραδοτέο να παρουσιάσετε και κάποια στατιστικά για τα ερωτήματα που εκτελέσατε.

3ο Τμήμα

Θα προχωρήσετε σε ερωτήματα πάνω σε XML έγγραφα. Θα κατεβάσετε τα XML έγγραφα που σχετίζονται με auctions από εδώ http://aiweb.cs.washington.edu/research/projects/xmltk/xmldata/www/repository.html#courses και

στη συνέχεια θα τα χρησιμοποιήσετε στο Terrier framework. Για να το επιτύχετε θα ακολουθήσετε τις οδηγίες που υπάρχουν εδώ http://terrier.org/docs/v4.2/configure_retrieval.html. Προφανώς θα πρέπει να έχει προηγηθεί το indexing των εγγράφων. Για το weighting υιοθετήστε τις μετρικές bm25 & tf_idf καθώς και query expansion. Δημιουργήστε κάποια ερωτήματα της αρεσκείας σας και αποτυπώστε τα αποτελέσματα.

Το παραδοτέο σας θα είναι οι εντολές του εργαλείου / κώδικας που χρησιμοποιήσατε καθώς και documentation με μέγεθος 1-2 σελίδων το μέγιστο.