Πανεπιστήμιο Κρήτης, Τμήμα Επιστήμης Υπολογιστών
HY463 - Συστήματα Ανάκτησης Πληροφοριών
2007-2008 Εαρινό Εξάμηνο

Προγραμματιστική Άσκηση
MyIR-A

Ανάθεση: 4 Μαρτίου
Παράδοση: 18 Μαρτίου

Σκοπός αυτής της άσκησης είναι να κατανοήσετε κάποιες βασικές έννοιες και τεχνικές, φτιάχνοντας εξ' αρχής ένα δικό σας IR system. Για να δοκιμάσετε το σύστημα σας, μπορείτε να χρησιμοποιήσετε μια συλλογή που θα σας δοθεί.

Μπορείτε να ακολουθήσετε τα εξής βήματα:

Π1) Γράψτε (κατά προτίμηση σε Java) ένα πρόγραμμα το οποίο να μπορεί να διαβάζει αρχεία κειμένου και να τυπώνει το πλήθος των διαφορετικών λέξεων και την κάθε διαφορετική λέξη συνοδευόμενη από το πλήθος εμφανίσεών της[1]. Επίσης να μπορεί να τυπώσει αρχείο εξόδου το οποίο να είναι αναγνώσιμο από το gnuplot με σκοπό την εμφάνιση γραφικής παράστασης στην οποία ο Χ άξονας θα έχει ένα σημείο για κάθε διαφορετική λέξη, αρχίζοντας από την πιο συχνά εμφανιζόμενη, και το f(x) μιας λέξης x θα είναι το πλήθος των εμφανίσεων της λέξης. Επίσης συντάξτε ένα .plt αρχείο το οποίο να παρουσιάζει το παραπάνω γράφημα σε κανονική και σε λογαριθμική κλίμακα[2].

Π2) Επεκτείνετε το σύστημα ώστε να μπορεί να διαβάσει όχι μόνο ένα αλλά πολλά αρχεία (π.χ. όσα βρίσκονται σε ένα συγκεκριμένο φάκελο του λειτουργικού συστήματος). Να μπορεί να κάνει ότι περιγράφεται στο (Π1) τόσο για κάθε αρχείο ξεχωριστά, όσο και συγκεντρωτικά για όλα τα αρχεία του φακέλου (ήτοι να μπορεί να μεταχειριστεί το σύνολο των αρχείων σαν να ήταν ένα αρχείο).

Π3) Συντάξτε μια αναφορά (κατά προτίμηση σε latex) η οποία να περιέχει τα γραφήματα για τουλάχιστον 10 αρχεία καθώς και το συγκεντρωτικό γράφημα[3].

Π4) Επεκτείνετε το σύστημα ώστε να δημιουργεί ένα φάκελο CollectionIndex στον οποίο να δημιουργεί ένα αρχείο (Lexicon.txt) που θα καταγράφει όλες τις διαφορετικές λέξεις σε αύξουσα (λεξικογραφική) σειρά. Δίπλα σε κάθε λέξη να καταγράφεται το πλήθος των εγγράφων στα οποία εμφανίζεται.

Π5) Επεκτείνετε το σύστημα ώστε να δημιουργεί ένα αρχείο (DocumentIDs.txt) που θα καταγράφει για κάθε αρχείο της συλλογής ένα ζευγάρι αποτελούμενο από ένα μοναδικό αριθμητικό αναγνωριστικό και το πλήρες μονοπάτι του αρχείου. Οι εγγραφές αυτές να είναι καταγεγραμμένες σε αύξουσα σειρά ως προς το αναγνωριστικό.

---

[1] Μπορείτε να δείτε (ή να θυμηθείτε) τα παραδείγματα που υπάρχουν στο
http://www.csd.uoc.gr/~hy252/Lectures07/pdf/CS252CollectionClassesInterfaces07.pdf

[2] Το GnuPlot μπορείτε να το κατεβάσετε από:
http://sourceforge.net/project/showfiles.php?group_id=2055.
Online documentation βρίσκεται στο: http://www.gnuplot.info/docs/gnuplot.html.
Σε pdf μορφή: http://www.gnuplot.info/docs/gnuplot.pdf.
Tutorials θα βρείτε στην σελίδα: http://www.gnuplot.info/help.html.

[3] Ένα πρότυπο κείμενο σε latex μπορείτε να βρείτε στην διεύθυνση:
http://groogle.csd.uoc.gr/apache2-default/index.php/Document_Templates
Tutorial για Latex βρίσκετε στην σελίδα: http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/

Π6) Επεκτείνετε το σύστημα ώστε να δημιουργεί ένα αρχείο για κάθε λέξη (που εμφανίζεται στο Lexicon.txt) όπου μέσα του θα καταγράφονται οι εμφανίσεις της κάθε λέξης σε αύξουσα σειρά (ως προς το αναγνωριστικό του εγγράφου και τη θέση εμφάνισης της λέξης στο έγγραφο).
Κάντε και μια παραλλαγή όπου στο αρχείο της κάθε λέξης θα εμφανίζεται το αναγνωριστικό του αρχείου στο οποίο εμφανίζεται και το TF της λέξης στο αντίστοιχο κείμενο.
Τα αρχεία αυτά να δημιουργούνται στο φάκελο CollectionIndex/InvertedLists.

Π7) Συγκρίνετε το μέγεθος της συλλογής κειμένων που χρησιμοποιήσατε με αυτού του ευρετηρίου σας και παραδώστε σχετική αναφορά με μετρήσεις. Το μέγεθος του ευρετηρίου σας είναι το άθροισμα του μεγέθους του Lexicon.txt, του DocumentsIDs καθώς και των αρχείων εμφανίσεων της κάθε λέξης. Να γίνει ξεχωριστή μέτρηση για την παραλλαγή των αρχείων εμφανίσεων λέξεων (που έχουν μόνο αναγνωριστικό αρχείου και TF).

Προαιρετικά
Στελέχωση κειμένου, υποστήριξη λίστας αποκλεισμού, διευθυνσιοδότηση «τμημάτων» (block addressing), GUI.