Διαχείριση Περιεχομένου Παγκόσμιου Ιστού και Γλωσσικά Εργαλεία

Ακαδημαϊκό Έτος 2011-2012

Γενικά

Στόχος της άσκησης είναι η υλοποίηση ενός συστήματος διαχείρισης των κειμενικών δεδομένων που συγκεντρώνει τυπικά ένας χρήστης στον υπολογιστή του. Στα πλαίσια του project θα υλοποιήσετε ένα σύστημα το οποίο θα διαχειρίζεται πόρους που συναντάμε στον υπολογιστή ενός χρήστη και θα δίνει σαν έξοδο τις μοντελοποιημένες πληροφορίες που συγκεντρώσατε από τα διαθέσιμα δεδομένα.

Πιο συγκεκριμένα, το σύστημα που θα υλοποιήσετε θα πρέπει να ικανοποιεί τις παρακάτω απαιτήσεις:

1. Θα είναι σε θέση να ανατρέχει στα αρχεία ενός χρήστη και να εξάγει κείμενο από συγκεκριμένους τύπους αρχείων.
2. Για όλα τα κείμενα που έχει εξάγει θα παράγει ανεστραμμένο ευρετήριο στο οποίο θα υπάρχει η δυνατότητα αποθήκευσής του, φόρτωσης και αναζήτησης σε αυτό.
3. Με βάση τα κείμενα που εξήγαγε θα παράγει ένα προφίλ χρήστη, δηλαδή θα επιχειρεί μια προσέγγιση των ενδιαφερόντων του χρήστη.

Ακολουθεί παρακάτω πιο λεπτομερής περιγραφή των απαιτήσεων.

Σημειώνεται ότι για το project θα βαθμολογηθούν εκτός από την ορθότητα του συστήματος τα ακόλουθα:

- Σωστός σχεδιασμός με βάση τις απαιτήσεις σε χώρο και χρόνο.
- Σχεδιασμός λαμβάνοντας υπόψη τις ανάγκες ενός πραγματικού συστήματος.
- Όπου ζητάται από εσάς να επιλέξετε τον αλγόριθμο, θα στηριχτείτε πάνω στη θεωρία ή σε δοκιμές που θα πραγματοποιήσετε για να κάνετε την επιλογή. Θα βαθμολογηθεί το σκεπτικό σας και η διαδικασία που ακολουθήσατε για την επιλογή.
- Μη φοβηθείτε να πάρετε πρωτοβουλίες αρκεί να τις στηρίξετε και να δώσουν αποτελέσματα ακόμα κι αν δεν είναι 100% σωστά!
- **Σημαντικό**: Θα πάρετε βαθμό για ότι μπορέσετε να τεκμηριώσετε στην προφορική εξέταση και με βάση τις απαντήσεις που θα δώσετε. Θα πρέπει να μπορείτε να απαντήσετε σε ερωτήσεις πάνω στον κώδικα και να αναλύσετε εκτενώς το σκεπτικό της υλοποίησης και του σχεδιασμού σας με βάση τη θεωρία που έχετε διδαχθεί. Δεν βαθμολογείται ο κώδικας, αλλά η προφορική σας εξέταση πάνω στον κώδικα.

Απαιτήσεις

Α. Πηγές Δεδομένων

Αρχικά το σύστημά σας θα πρέπει να «μαζεύει» από τον υπολογιστή στον οποίο τρέχει ένα σύνολο δεδομένων που ανήκουν στον χρήστη και θεωρητικά είναι ενδεικτικά των ενδιαφερόντων του. Ένας χρήστης στον υπολογιστή του πραγματοποιεί ενέργειες οι οποίες ομαδοποιούνται σε δύο κατηγορίες:

- Αποθηκεύει και διαχειρίζεται αρχεία τοπικά
- Αλληλεπιδρά με τον Παγκόσμιο Ιστό.

Από τις δύο αυτές κατηγορίες ενεργειών επιδιώκουμε να συλλέξουμε πληροφορίες για τα ενδιαφέροντα του χρήστη. Ακολουθούν προτάσεις για πηγές τις οποίες θα επεξεργαστείτε για το project. Οποιαδήποτε επιπλέον πηγή δεδομένων σκεφτείτε και υλοποιήσετε την επεξεργασία της θα είναι bonus στη βαθμολογία σας. Σημειώστε τις επιλογές σας και τι υλοποιήσατε τελικά σε σύντομη αναφορά.

Τοπικά Αρχεία

Το σύστημά σας θα δέχεται ως είσοδο έναν κεντρικό φάκελο ο οποίος θα περιέχει τα αρχεία του χρήστη. Θα διατρέχει τα περιεχόμενα του φακέλου και θα εντοπίζει αρχεία των ακόλουθων τύπων:

- Αρχεία κειμένου (txt)
- Αποθηκευμένα αρχεία ιστοσελίδων (html ή htm)
- PDF
    - Για την εξαγωγή κειμένου δείτε σε python το PDFMiner και το pdf2txt εργαλείο που το συνοδεύει:
        - http://www.unixuser.org/~euske/python/pdfminer/index.html

Θα εξάγει το κείμενο για κάθε ένα από αυτά τα αρχεία για να συνεχίσει με την επεξεργασία.

Δεδομένα του Παγκόσμιου Ιστού

Τα δεδομένα που θα εκμεταλλευτείτε από την αλληλεπίδραση του χρήστη με τον παγκόσμιο ιστό είναι τα bookmarks (ή favourites) του χρήστη. Οι browser δίνουν τη δυνατότητα εξαγωγής των bookmarks σε ένα html αρχείο, το οποίο περιέχει μια λίστα από τις διευθύνσεις που έχει κάνει bookmark ο χρήστης. Θα εξάγετε αυτό το αρχείο και θα το δώσετε ως είσοδο στο σύστημά σας. Με αυτόματο τρόπο θα το επεξεργαστείτε, θα εξάγετε τις διευθύνσεις που περιέχει και θα κατεβάσετε τις ιστοσελίδες που έχει σημειώσει ο χρήστης. Από αυτές θα εξάγετε το καθαρό κείμενο που θα χρησιμοποιήσετε.

Στον Firefox η διαδικασία εξαγωγής περιγράφεται στο:
http://support.mozilla.org/en-US/kb/Exporting%20bookmarks%20to%20an%20HTML%20file
Ανάλογη διαδικασία υπάρχει για το Chrome. Αν έχετε browser που δεν υποσηρίζει κάτι τέτοιο, πάρτε τα bookmarks του συνεργάτη σας ή εγκαταστήστε firefox και κάντε import.

Β. Ευρετηριοποίηση των δεδομένων

Εφόσον έχετε συλλέξει τα κείμενα όπως περιγράφεται προηγουμένως θα δημιουργήσετε ένα ανεστραμμένο ευρετήριο των περιεχομένων τους εφαρμόζοντας τις τεχνικές που έχουν συζητηθεί στο

μάθημα. Επιλέξατε τις πιο αποδοτικές μεθόδους που έχουμε συζητήσει για την προεπεξεργασία του κειμένου στο αντίστοιχο φροντιστήριο:

http://www.dblab.upatras.gr/download/courses/DIAXEIRISH_PERIEXOMENOU/2011_12/frontistirio/3_preprocessing.pdf

Θα υλοποιήσετε εφαρμογή που θα πραγματοποιεί όλη τη διαδικασία προεπεξεργασίας των κειμένων και δημιουργίας του ανεστραμμένου ευρετηρίου. Πρέπει να λάβετε υπόψη σας τα ζητήματα απόδοσης που έχουν συζητηθεί. Επίσης θα δίνεται δυνατότητα αποθήκευσης του ανεστραμμένου ευρετηρίου σε ό,τι μορφή θέλετε εσείς καθώς και φόρτωσής του έτσι ώστε να πραγματοποιούνται αναζητήσεις σε αυτό. Θα πρέπει να δίνεται η δυνατότητα αναζήτησης στο ευρετήριο κατά την οποία θα επιστρέφεται full path όταν το αρχείο βρίσκεται τοπικά αποθηκευμένο ή url όταν είναι στον παγκόσμιο ιστό.

Σημειώστε στην αναφορά τι επιλογές εργαλείων κάνατε για την προεπεξεργασία, καθώς και ποια μέθοδο υπολογισμού βαρών επιλέξατε. Επίσης σύντομα περιγράψτε τη μορφή αποθήκευσης του ανεστραμμένου ευρετηρίου που επιλέξατε.

Γ. Εξαγωγή προφίλ του χρήστη

Από τα κείμενα που έχετε συλλέξει θα εξάγετε το αντίστοιχο προφίλ του χρήστη. Το προφίλ του χρήστη που σας ζητάται είναι 20 ουσιαστικά τα οποία θα αντιπροσωπεύουν τα ενδιαφέροντά του. Τα ουσιαστικά αυτά θα τα εξάγετε (με αυτόματο τρόπο) από τα κείμενα που συλλέξατε με οποιαδήποτε μέθοδο θεωρείτε εσείς ως αποδοτικότερη. Ο αλγόριθμος εξαγωγής του προφίλ που θα υλοποιήσετε θα πρέπει να είναι τεκμηριωμένος με βάση τη θεωρία ή τις προσωπικές σας παρατηρήσεις. Περιγράψτε τον αλγόριθμο και την διαδικασία τεκμηρίωσης στην αναφορά σας.

Τα 20 ουσιαστικά που θα προκύψουν θα τα δώσετε ως είσοδο σε αλγόριθμο αποσαφήνισης που βασίζεται στο WordNet, με στόχο να έχετε ως έξοδο μια λίστα εννοιών που αντιπροσωπεύουν τις προτιμήσεις του χρήστη. Η μέθοδος που θα ακολουθήσετε θα είναι και πάλι της δικιάς σας επιλογής. Περιγράψτε στην αναφορά τη μέθοδο που επιλέξατε και τα προβλήματα που προέκυψαν.

Θα πρέπει επίσης να πραγματοποιήσετε δοκιμές και να σχολιάσετε αν η μέθοδός σας κατάφερε να προσεγγίσει αποτελεσματικά το προφίλ σας. Σχολιάστε τη σχέση της ακρίβειας του προφίλ με τον αριθμό και το είδος των κειμένων που συγκεντρώσατε.

Ζητήματα Υλοποίησης

Ελληνικά

Πολλά από τα αρχεία που έχετε στον υπολογιστή σας είναι ελληνικά. Τα εργαλεία όπως ο tagger δεν μπορούν να τα επεξεργαστούν. Τα ελληνικά δεδομένα και τις ελληνικές λέξεις απλά θα τις αγνοείτε και θα επεξεργαστείτε τα αγγλικά δεδομένα που έχετε. Για παράδειγμα, αν σε ένα έγγραφο υπάρχουν ανακατεμένες αγγλικές και ελληνικές λέξεις θα αγνοήσετε τις ελληνικές εντελώς και θα δουλέψετε μόνο με τις αγγλικές. Αν υπάρχουν πάλι μόνο ελληνικές λέξεις θα αγνοήσετε όλο το έγγραφο!

Απόδοση των εργαλείων

Τα εργαλεία που θα χρησιμοποιήσετε λόγω της άγνωστης φύσης των δεδομένων που υπάρχουν στον υπολογιστή σας μπορεί να παρουσιάσουν κάποια προβλήματα και να μην έχουν απόδοση 100%. Αγνοήστε τα προβλήματα και μην αναλώσετε χρόνο στο να τα επιλύσετε εφόσον δεν σας περιορίζουν τα δεδομένα δραματικά. Για παράδειγμα μπορεί να μην είναι δυνατόν να εξάγετε κείμενο από κάποιο PDF που είναι πολύ παλιό. Το σύστημά σας θα πρέπει να αγνοεί τα σφάλματα στην επεξεργασία ενός αρχείου και να συνεχίζει με την επεξεργασία των υπολοίπων.

Υλοποίηση

Όλα όσα περιγράφονται θα πρέπει να γίνονται αυτόματα από το σύστημα που θα παραδώσετε. Το μόνο που θα κάνετε «με το χέρι» θα είναι η εξαγωγή των bookmarks από τον browser. Η εξαγωγή κειμένου από τα αρχεία και το «κατέβασμα» των bookmarks αντίθετα θα πραγματοποιούνται αυτόματα από το σύστημα. Γενικά έχετε υπόψη σας ότι το σύστημα θα πρέπει να ανταποκρίνεται σε πραγματικές ανάγκες, δηλαδή να δουλεύει όσο το δυνατόν πιο ολοκληρωμένα και χωρίς ενδιάμεσες παρεμβάσεις.

Αναφορά

Η αναφορά που θα παραδώσετε δεν χρειάζεται να είναι ιδιαίτερα εκτενής ούτε "επίσημη". Ουσιαστικά το μόνο που χρειάζεται να περιέχει είναι ο σχολιασμός που ζητάται και τα σημεία που βαθμολογούνται και θα με βοηθήσουν στο να διορθώσω την άσκησή σας. Και φυσικά ότι νομίζετε εσείς ότι πρέπει να διευκρινήσετε πάνω στον κώδικα.

Η παράδοση θα γίνει τη μέρα της εξέτασης η οποία θα πραγματοποιηθεί προφορικά μετά το τέλος της εξέτασης του Ιουνίου. Παραδίδετε τα πάντα σε ηλεκτρονική μορφή. Δεν χρειάζεται να εκτυπώσετε την αναφορά και δεν είναι απαραίτητο να γράψετε CD. Αρκεί να φέρετε μόνο τα αρχεία κώδικα και την αναφορά σε φλασάκι στην εξέταση. Προσοχή: μην φέρετε μαζί και μη μου αφήσετε προσωπικά δεδομένα, όπως είναι πχ το προφίλ σας και τα δεδομένα σας! Αρκεί η περιγραφή των δεδομένων στην αναφορά. Θα τρέξουμε τη μέθοδό σας στον δικό μου υπολογιστή για να εξάγουμε προφίλ (αφού έχω σβήσει τα bookmarks που δεν θέλω να δείτε :P).