Ανάκτηση Πληροφορίας – Χειμερινό Εξάμηνο 2013-2014
Εργασία

1. Βασικά Στοιχεία

Στόχος της εργασίας είναι η ενασχόληση με thread programming με τη γλώσσα C++. Πιο συγκεκριμένα, καλείστε να υλοποιήσετε μία παράλληλη μηχανή ανάκτησης κειμένου που θα έχει περιορισμένες δυνατότητες και θα χρησιμοποιεί τους πυρήνες ενός πολυπήρυνου συστήματος. Προτείνεται η εκπόνηση της εργασίας να βασιστεί σε Linux ώστε να αποφύγουμε προβλήματα συμβατότητας (π.χ. με POSIX threads κλπ). Ωστόσο, και σε MacOS δε θα υπάρχει πρόβλημα.

Θα πρέπει να χρησιμοποιηθεί ο compiler gcc/g++ έκδοση τουλάχιστον 4.8. Εάν δεν τον έχετε μπορείτε να δείτε πως θα τον εγκαταστήσετε διαβάζοντας κάποιο σχετικό post, π.χ. αυτό:
http://ubuntuhandbook.org/index.php/2013/08/install-gcc-4-8-via-ppa-in-ubuntu-12-04-13-04/

Όσοι δεν εργάζονται σε Linux το πιο εύκολο είναι η χρήση virtualbox και η εγκατάσταση Linux ως virtual machine. Έτσι δε θα χρειαστεί να πειράξετε την εγκτατάστασή σας. Φυσικά μπορείτε να χρησιμοποιήσετε όποιο IDE θέλετε (Eclipse, NetBeans, CodeBlocks, κλπ). Μπορείτε να χρησιμοποιήσετε είτε POSIX threads (pthreads) είτε να χρησιμοποιήσετε τις ευκολίες που παρέχει το c++11 standard για τη διαχείριση threads.

2. Περιγραφή

Το σύστημα που θα υλοποιήσετε θα υποστηρίζει δύο βασικές λειτουργίες που είναι 1) η δημιουργία αντεστραμμένου καταλόγου μίας συλλογής εγγράφων και 2) η επεξεργασίας ενός συνόλου ερωτημάτων. Πιο συγκεκριμένα:

Δημιουργία Καταλόγου: δίνεται ως είσοδος ένα αρχείο txt που σε κάθε γραμμή περιέχει τον κωδικό του εγγράφου και στην συνέχεια τις λέξεις του. Η πρώτη γραμμή του αρχείου περιέχει το πλήθος των εγγράφων που υπάρχουν συνολικά στο αρχείο. Για παράδειγμα:

```
3
1 this is the first document
2 this is the second document
3 this is another document
```

Θα υποθέσουμε ότι ο κατάλογος που θα δημιουργηθεί θα χωρά στην κύρια μνήμη. Προφανώς, η διαδικασία κατασκευής θα πρέπει να λάβει υπόψη τον παραλληλισμό οπότε θα πρέπει να χρησιμοποιεί threads.

Επεξεργασία Ερωτημάτων: δίνεται ως είσοδος ένα αρχείο txt που σε κάθε γραμμή περιέχει τον κωδικό ενός ερωτήματος, το πλήθος των απαντήσεων και τις λέξεις του ερωτήματος. Η πρώτη γραμμή περιέχει το πλήθος των ερωτημάτων. Για παράδειγμα:

```
3
1       5       information retrieval
2       4       data structures
3       3       data mining
```

Η επεξεργασία των ερωτημάτων θα γίνεται με βάση το κλασικό διανυσματικό μοντέλο που θα χρησιμοποιεί το cosine ως μέτρο ομοιότητας και τα βάρη των διανυσμάτων θα δημιουργούνται με βάση TF x IDF. Μέσα στον αντεστραμμένο κατάλογο μπορείτε να αποθηκεύσετε ότι πληροφορίες θεωρείτε απαραίτητες για την υποστήριξη του μοντέλου.

3. Παράδοση και Βαθμολόγηση

Η εργασία θα πρέπει να εκπονηθεί σε ομάδες των δύο ατόμων και να παραδοθεί στο τέλος της εξεταστικής Ιανουαρίου και θα εξεταστεί προφορικά. Ο βαθμός της εργασίας αντιστοιχεί στο 50% του τελικού βαθμού ενώ θα υπάρχει και bonus +1 βαθμός για παραπάνω δυνατότητες που θα υποστηριχθούν. Βασικό κριτήριο βαθμολογίας θα είναι ο χρόνος εκτέλεσης της κατασκευής του καταλόγου και ο χρόνος επεξεργασίας των ερωτημάτων. Φροντίστε ώστε οι δύο φάσεις να είναι διακριτές και μετά την εκτέλεση της κάθε φάσης να τυπώνεται ο χρόνος που χρειάστηκε το σύστημα για την επεξεργασία. Ο καλύτερος τρόπος είναι να χρησιμοποιήσετε την **getrusage()** που δίνει ακριβείς πληροφορίες για το χρόνο εκτέλεσης.

Για απορίες και παρατηρήσεις papadopo@csd.auth.gr