



ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών



ΤΜΗΜΑ
ΠΛΗΡΟΦΟΡΙΚΗΣ &
ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Μάθημα
Προγραμματισμός Συστήματος

Project 3

Όνομα
Κυλάφη Χριστίνα-Θεανώ

ΑΜ
1115201200077

Αθήνα, 2018

Η άσκηση έχει υλοποιηθεί σε γλώσσα C, στο πρόγραμμα “sublime text” ενώ παράλληλα δοκιμαζόταν σε περιβάλλον linux μέσω απομακρυσμένης σύνδεσης μέσω τερματικού macOS.

Στο φάκελο, περιλαμβάνονται τα εξής αρχεία εκτός από το readme :

makefile, server.c, structs.h, mycrawler.c, webcreator.sh (script δημιουργίας ιστοσελίδων), book.txt (το αρχείο με του οποίου το περιεχόμενο γεμίζουν οι ιστοσελίδες), header.txt, footer.txt.

Με την εντολή “**make**” παράγονται τα εκτελέσιμα “**myhttpd**” και “**mycrawler**”.

Έχω ήδη δημιουργήσει ενδεικτικά, 2 websites , από 3 σελίδες στο καθένα για τη δοκιμή μου ως εξής:

./webcreator.sh sites book.txt 2 3

Η σύνταξη για την εκτέλεσή των παραπάνω, είναι:

./myhttpd -p serving_port -c command_port -t num_of_threads -d root_dir

./mycrawler -h host_or_IP -p port -c command_port -t num_of_threads -d save_dir starting_URL

(το starting url είναι της μορφής: **http://linux10.di.uoa.gr:8080/site0/page0_3626.html**)

Έχω προσθέσει ελέγχους λανθασμένης σύνταξης.

Ο webcreator, δημιουργεί τις ιστοσελίδες ακριβώς με τον τρόπο που εξηγείται στην εκφώνηση.

Οι τυχαίοι σύνδεσμοι που τοποθετούνται στις σελίδες, είναι της μορφής: **/site0/page0_3626.html** .

Εαν εκτελεστεί ο crawler, ξεκινώντας από το αρχικό path που δίνεται σαν όρισμα, ζητάει κάθε σελίδα που βρίσκεται στη λίστα από τον server που τρέχει στο δωσμένο host, την αντιγράφει στο νέο φάκελο και έπειτα τη διατρέπει για την ανάκτηση όλων των links που περιέχει, τα οποία τοποθετεί στη λίστα. Αυτή η διαδικασία γίνεται μέσα στο critical section που έχω δημιουργήσει με τα monitors , όπως αναφέρω παρακάτω.

Παράλληλα και αφού τελειώσει το κατέβασμα όλων των σελίδων, ο crawler ακούει στις εντολές “STATS” και “SHUTDOWN”, όπως ακριβώς και ο server, στο αντίστοιχό του command port.

Τυπώνονται τα κατάλληλα μηνύματα, σταματούν τα threads εαν δεν έχουν σταματήσει ήδη μετά την ολοκλήρωση του download των σελίδων, απελευθερώνεται η μνήμη και τερματίζει το πρόγραμμα.

Παρατηρήσεις:

- Για το συγχρονισμό των threads και στις 2 περιπτώσεις, χρησιμοποίησα monitors (condition variables προστατευμένα απο locks). Συγκεκριμένα, έχω την αντίστοιχη condition variable σε κάθε περίπτωση, στην οποία θα περιμένουν τα threads εαν υπάρχει ήδη ένα thread που επεμβαίνει(παίρνει ή τοποθετεί στοιχείο) επάνω στην αντίστοιχη λίστα κάθε φορά (fd list / path list). Στην πρώτη περίπτωση, υπάρχει επιπλέον η συνθήκη που δηλώνει πως η λίστα είναι κενή, κάτι που μπλοκάρει τα threads μέχρι το main thread να τοποθετήσει ένα στοιχείο στην λίστα.
- Η παραπάνω διαδικασία , συνεχίζεται μέχρι να δωθεί κάποιο command στο κατάλληλο port από το χρήστη.
- Στην πρώτη περίπτωση, επειδή έχουμε 2 ports, χρησιμοποιείται η select().
- Εαν υπάρχει ήδη φάκελος με τα αντίγραφα (mycrawler), διαγράφεται και δημιουργείται ένας κενός με το ίδιο όνομα που έχει δωθεί σαν όρισμα.
- Στην πρώτη περίπτωση, αν δεν υπάρχει ο φάκελος που δίνεται σαν όρισμα, το πρόγραμμα σταματάει.
- Η ζητούμενη λειτουργία, επιτυγχάνεται με τη χρήση threads και η επικοινωνία server - client, με τη χρήση blocking sockets.

Ελλείψεις:

- Δεν έχει υλοποιηθεί η εντολή “**Search**” στον Crawler.

Ενδεικτικές εκτελέσεις:

```
./myhttpd -p 8080 -c 9090 -t 2 -d sites
```

```
linux10:/home/users/sdi1200077/syspro/pro3>./myhttpd -p 8080 -c 9090 -t 2 -d sites

Serving Port: 8080
Command port: 9090
# of threads: 2
Directory: sites

Main thread: 140625049900800
Created thread 140625041598208
Created thread 140625033205504

Server waiting for connections...
I am the newly created thread 140625041598208
I am the newly created thread 140625033205504
Serving Port(8080): Accepted connection from localhost (socket fd: 5)

Server waiting for connections...

Fetching page: ./sites/site0/page0_3626.html ...
```

```
./mycrawler -h linux10.di.uoa.gr -p 8080 -c 9090 -t 3 -d save_dir http://  
linux10.di.uoa.gr:8080/site0/page0_3626
```

```
[linux08:/home/users/sdi1200077/syspro/pro3>./mycrawler -h linux10.di.uoa.gr -p 8080 -c 9090 -t 3 -d save_dir http://linux10.di.uoa.gr:8080/site0/page0_3626.html

-----I am the CRAWLLLLLEEEEEER-----

Host: linux10.di.uoa.gr
Port: 8080
Command port: 9090
# of threads: 3
Directory: save_dir
Initial path: /site0/page0_3626.html


Main thread: 140188704208640
Created thread 140188695906048
Created thread 140188687513344


I am the newly created thread 140188695906048
Created thread 140188679120640


I am the newly created thread 140188679120640


I am the newly created thread 140188687513344
GET /site0/page0_3626.html
Page found!
Content-Length: 67175
Copying file /site0/page0_3626.html


GET /site0/page0_7523.html
Page found!
Content-Length: 83977
Copying file /site0/page0_7523.html


GET /site0/page0_10189.html
Page found!
Content-Length: 60833
Copying file /site0/page0_10189.html


GET /site1/page1_1420.html
Page found!
Content-Length: 81352
Copying file /site1/page1_1420.html


GET /site1/page1_2448.html
Page found!
Content-Length: 78258
Copying file /site1/page1_2448.html


GET /site1/page1_20908.html
Page found!
Content-Length: 66293
Copying file /site1/page1_20908.html


Thread 140188695906048 exiting...
Thread 140188679120640 exiting...
Thread 140188687513344 exiting...


Command Port(9090): Accepted connection from localhost (socket fd: 4)

- - - CRAWLLLLLEEEEEER OUT.... - - -
```