NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Unix System Programming

-3rd Project-

By Kalopisis Ioannis
AM: 1115201500059
June 5, 2018

1 Γενική Περιγραφή

Στην άσκηση αυτή όλες οι εφαρμογές που δημιουργήθηκαν συνεργάζονται μεταξύτους για να παράξουν ένα όλοκληρωμένο αποτέλεσμα, παρόμοιο με αυτό ορισμένων εφρμογών αναζήτησης, όπως η Google. Η εφαρμογή δημιουργεί ιστοσελίδες από όποιο κείμενο της δοθεί, εξυπηρετεί αιτήματα από τον browser μέσα από τον server και τέλος κατεβάζει και αναλύει όλες τις ιστοσελίδες.

2 Bach Script

Η εφαρμογή κελύφους $webcrator.sh$ δημιουργεί W sites με P ιστοσελίδες το καθένα μέσα, όπου W και P δίνονται ως παράμετροι κατά την εκτέλεση. Δημιουργεί επίσης για κάθε ιστοσελίδα ένα σύνολο από f = (p/2) + 1 εσωτερικά links προς σελίδες του ίδιου site εκτός από τον εαυτό της και ένα σύνολο από q = (w/2) + 1 εξωτερικά links προς άλλα sites. Αυτά τα τοποθετεί καθώς δημιουργεί την ιστοσελίδα κάθε m γραμμές, όπου m τυχαίος αριθμός με 1000<m<2000. Τέλος εμφανίζει και αν όλες οι σελίδες έχουν εισερχόμενα links. Μια ενδεικτική εκτέλεση είναι: ./webcreator.sh root_dir text_file 5 4, όπου root_dir είναι ο φάκελος που θα αποθηκευτούν τα sites, το text_file είναι το κείμενο που θα διαβάσει, 5 είναι ο αριθμός των site και 4 ο αριθμός των σελίδων σε κάθε site.

3 Server

Σε αυτό το κομμάτι της εφαρμογής έχει υλοποιηθεί ένας απλός server που εξυπηρετεί αιτήματα προς τις σελίδες από κάποιον browser. Για παράδειγμα όταν ανοίξουμε κάποια από τις ιστοσελίδες που φτιάξαμε στο προηγούμενο κομμάτι με κάποιον browser μπορούμε να πατήσουμε πάνω σε κάποιο από τα links που εμφανίζονται στη σελίδα. Τότε ο browser θα στείλει αίτημα στο port που ακούει ο server μας και αυτός θα αναλύση και εξυπηρετήσει το αίτημα αν υπάρχει η σελίδα και έχει δικαιώματα πάνω σε αυτή στέλνοντάς την πίσω. Σε αυτή την περίπτωση ο browser θα μας εμφανίσει την καινούργια σελίδα. Αν δεν έχει δικαιώματα ο server πάνω σε αυτή τη σελίδα ή αν δεν υπάρχει η σελίδα που ζητήθηκε τότε επιστρέφει στον browser το κατάλληλο μήνυμα. Ένα ενδεικτικό αίτημα από τον browser στον server είναι: http://127.0.0.1:5050/site0/page0_8552.html. Επίσης ο server ακούει και σε ένα ακόμα port στο οποίο δέχεται εντολές μέσω browser πάλι. Οι εντολές αυτές είναι STATS για να εμφανίσει στατιστικά για το πόση ώρα λειτουργεί, το πόσα αιτήματα εξυπηρέτησε και το πόσα bytes έχει στείλει και SHUTDOWN για τον τερματισμό του. Ενδεικτικές εκτελέσεις είναι: http://127.0.0.1:5050/STATS και http://127.0.0.1:5050/SHUTDOWN.

4 Crawler

Στο τελευταίο κομμάτι της εφαρμογής έχει υλοποιηθεί ένας απλός crawler που παίρνει σαν όρισμα κατά της εκτέλεση το port που ακούει ο server που φτιάξαμε και ένα αρχικό link. Στέλνει άιτημα στον server για αυτό το link και το κατεβάζει από τον server. Έπειτα το αναλύει και παίρνει όλα τα links που έχει μέσα και τα τοποθετεί σε μία ουρά. Στη συνέχεια κατεβάζει από το server κάθε link που έχει στην ουρά και επεναλαμβάνει την ίδια διαδικασία μέχρι να αδειάσει η ουρά. Στο τέλος θα έχουμε σαν αποτέλεσμα ένα αντίγραφο του φακέλου root_dir που περιέχει όλα τα sites και τις ιστοσελίδες. Αφού ολοκληρώσει αυτή τη διαδικασία τότε σε ένα άλλο port που το παίρνει και αυτό σαν όρισμα κατά την εκτέλεση δέχεται εντολές από τον browser. Οι εντολές είναι: η STATS που δείχνει όπως και πριν το χρόνο εκτέλεσης, τις πόσες σελίδες κατέβασε και τα πόσα bytes κατέβασε, η SEARCH/word1/word2/word3.../word10 όπου εδώ έχει ενσωματοθεί ο κώδικας της προηγούμενης άσκησης μέσω socket και απαντάει σε search ερωτήματα για τις λέξεις που δώθηκαν μέσω browser και η SHUTDOWN που τερματίζει τον crawler. Ενδεικτικές εκτελέσεις των εντολών είναι: http://127.0.0.1:9000/STATS, http://127.0.0.1:9000/SEARCH /word1/word2/.../word10 και http://127.0.0.1:9000/SHUTDOWN.

5 Threads

Ο server και ο crawler τρέχουν με την βοήθεια νημάτων. Στον server τα αιτήματα μπαίνουν σε έναν πίνακα και εξυπηρετούνται από νηματα, ενώ στον crawler το κατέβασμα των ιστοσελίδων γίνεται από τα νήματα που δημιουργούνται στην αρχή.