

Μεταγλωτιστές 2020

Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: ΛΑΖΑΡΟΣ ΠΑΠΑΝΙΚΟΛΑΟΥ

ΑΜ: Π2017085

Α.Συνοπτική περιγραφή της σειράς βημάτων επεξεργασίας στον κώδικά σας.

Αρχικά ανοίξα το testpage.txt με την ρύθση και το εκτυπώσα με τις καταλλήλες εντολές για να δω αν τρέχει σωστά. Στην συνέχεια ανοίξα και έτρεξα το testpage.txt με την μέθοδο with όπως μας δίνεται από τις διαφανείες του μαθήματος και ελέγξα αν τρέχει και με αυτήν την μέθοδο σωστά. Τέλος διαδοχικά έλυσα όλα τα ερωτήματα που μας δίνει η εκφώνηση του μαθήματος.

Β.Περιγραφή της κανονικής έκφρασης που χρησιμοποιήσατε σε κάθε βήμα.

Για το ερώτημα 1. Χρησιμοποίησα την εντολή `<title>(.+?)</title>` καθώς γίνεται επιλογή ενός ή παραπάνω χαρακτήρων που βρίσκεται μέσα στο title tag. Χρησιμοποιώ τον τελεστή `.` για να επιλέξει οποιοσδήποτε χαρακτήρας και τον τελεστή `+` για να επιλέξει μια ή περισσότερες φορές

Για το ερώτημα 2. Χρησιμοποίησα την εντολή `<!--.*?-->`, `re.DOTALL` για να επιλέξω τα σχόλια. Χρησιμοποιώ αυτήν την εντολή για να επιλέξω τα περιεχόμενα των σχολίων. Το `re.DOTALL` χρησιμοποιείται καθώς είναι εφικτό να υπάρχουν σχόλια πολλαπλών γραμμών.

Για το ερώτημα 3. Χρησιμοποίησα την συγκεκριμένη εντολή για να επιλέξω τα `scripts` και τα `style` tags. `<(script|style).*?>.*?</(script|style)>`, `re.DOTALL` με αυτήν την εντολή απαιτείται απαλοιφή και οι παρενθέσεις είναι περιττές. Κυριο ρόλο σε όλο αυτό είναι ο τελεστής `.` και ο τελεστής `|`.

Για το ερώτημα 4. Χρησιμοποίησα την εντολή `<a.+?href="(.*?)".*?>(.*?)`, `re.DOTALL` με σκοπό να εξαγωγή όλων των `href` και του περιεχομένου το `<a>` που βρίσκονται απομονωμένα στις παρενθέσεις.

Για το ερώτημα 5. Χρησιμοποίησα την εντολή `<.+?>|</.+?>`, `re.DOTALL`, `<.+?/>`, `re.DOTALL` με σκοπό τα tags με διπλό κλείσιμο και τα tags με μόνο κλείσιμο.

Για το ερώτημα 6. Χρησιμοποίησα την εντολή `&(amp|gt|lt|nbsp);` με σκοπό την εξαγωγή των html entities που υπάρχουν στον πίνακα.

Για το ερώτημα 7. Χρησιμοποίησα την εντολή `\s+` για την Εξαγωγή whitespaces (`\s`) μια ή περισσότερες φορές (+).

Γ.Αναφορά σε πηγές που πιθανόν χρησιμοποιήσατε

Αντλήσα πληροφορίες από διαφάνειες του μαθήματος που μας δίνονται, από scripts που δίνονται από το github gist καθώς και από tutorials της Python.