Μεταγλωτιστές 2020
Προγραμματιστική Εργασία #2
Ονοματεπώνυμο: Δημήτρης Μισαηλίδης
ΑΜ: Π2017066

1. **Συνοπτική περιγραφή της σειράς βημάτων επεξεργασίας στον κώδικά σας**

Πρώτα απ' όλα, άνοιξα το αρχείο testpage.txt με την python και έλεγξα εκτυπώνοντάς το στην κονσόλα ότι ανοίγει κανονικά. Έπειτα, ξεκίνησα διαδοχικά τα ερωτήματα κάνοντας πρώτα μια κανονική έκφραση και έπειτα μέσα στην with (που ανοίγει το αρχείο), δοκίμαζα την σωστή λειτουργία της κανονικής έκφρασης. Αφού έγιναν όλα τα ερωτήματα έλεγξα προσεκτικά το output αρχείο.

2. **Περιγραφή της κανονικής έκφρασης που χρησιμοποιήσατε σε κάθε βήμα.**

Ερώτημα 1: `<title>(.+?)</title>`

Επιλογή ενός ή παραπάνω χαρακτήρων που βρίσκεται μέσα στο title tag. Τελεστής . οποιοσδήποτε χαρακτήρας, + μια ή περισσότερες φορές και είναι μέσα στην παρένθεση για την εξαγωγή του.

Ερώτημα 2: `<!--.*?-->',re.DOTALL`
Επιλογή σχολίων. Αφού απαιτείται απαλοιφή των σχολίων, δεν απομονώνουμε τα περιεχόμενα των σχολίων. Επίσης υπάρχει πιθανότητα κενού σχολίου οπότε ο τελεστής + δεν προτιμήθηκε. Τέλος το re.DOTALL χρησιμοποιήθηκε επειδή μπορεί να υπάρχουν σχόλια πολλαπλών γραμμών.

Ερώτημα 3: `<(script|style).*?>.*?</(script|style)>',re.DOTALL`
Επιλογή όλων των script και style tags. Αυτό γίνεται με τον τελεστή | για να επιλέγονται και τα 2 καθώς και με τους τελεστές .* για να μπαίνουν όλα τα περιεχόμενα. Αφού απαιτείται απαλοιφή, οι παρενθέσεις είναι περιττές.

Ερώτημα 4: `<a.+?href="(.*?)".*?>(.*?)</a>',re.DOTALL`
Εξαγωγή όλων των περιεχομένων του href και του περιεχόμενου του a, βρίσκονται απομονωμένα στις παρενθέσεις.

Ερώτημα 5: `<.+?>|</.+?>',re.DOTALL , <.+?/>',re.DOTALL`
Δύο κανονικές εκφράσεις χρησιμοποιούνται γιατί ένα tag μπορεί να είναι self-closing. Οπότε υπάρχουν 2 περιπτώσεις.

Ερώτημα 6: `&(amp|gt|lt|nbsp);`
Εξαγωγή των html entities που υπάρχουν στον πίνακα.

Ερώτημα 7: `\s+`
Εξαγωγή whitespaces (\s) μια ή περισσότερες φορές (+).

3. **Αναφορά σε πηγές που πιθανόν χρησιμοποιήσατε**

Σημειώσεις στο site του μαθήματος και εργαστηριακές ασκήσεις.