Μεταγλωτιστές 2020

Στ' Εξάμηνο

Προγραμματιστική Εργασία 2

Ονοματεπώνυμο : Δημήτριος Σιδηρόπουλος

ΑΜ : Π2016147

Ερώτημα Νο1:

Εξαγωγή και εκτύπωση του τίτλου (οτιδήποτε βρίσκεται μεταξύ <title> και </title>).

('<title>(.+?)</title>')

Για να συνδιαστεί οτιδήποτε ανάμεσα στα <title> και </title> είναι απαραίτητο να χρησιμοποιήσουμε την τελεία (.) και τον μαθηματικό τελεστή της πρόσθεσης (+), όπως επίσης είναι απαραίτητη και η ύπαρξη τουλάχιστον ενός χαρακτήρα.

Ερώτημα Νο2:

Απαλοιφή των σχόλιων (οτιδήποτε βρίσκεται μεταξύ<!-- και -->).

('<!--.*?-->',re.DOTALL)

Για να μπορέσει να συνδιαστεί ό,τι βρίσκεται ανάμεσα στα <!-- και --> γίνεται χρήση της τελείας (.) και του αστερίσκου (*), θα μπορούσε να χρησιμοποιηθεί και το σμβολο της πρόσθεσης (+) αλλά θα απαιτούσε και την χρήση ενός επιπλέον χαρακτήρα.

Ερώτημα Νο3:

Απλοιφή των <script> και <style> tags με όλο τους το περιεχόμενο, μέχρι δηλαδή να

συναντήσετε το αντίστοιχο </script> ή </style> (και τα τελευταία).

(r'<(s(?:cript | tyle)).*?>.*?</\1>',re.DOTALL).

Επειδή το πρώτο γράμμα και των δυο tags είναι ίδιο (s), χρησιμοποιούμε το σύμβολο | για την επιλογή της αντίστοιχης ακολουθίας γαμμάτων μετά απο το συγκεκριμένο γράμμα. Χρησιμοποιούμε τους χαρακτήρες ?: με αυτόν τον συνδιασμό ωστέ να μην οριστεί αυτόματα νέο group με την ακολουθία cript ή tyle, χρησιμοποιώντας το \1 ταιριάζουμε ό,τι βρέθηκε και είναι αποθηκευμένο στο group(1).

Ερώτημα Νο4:

Εξαγωγή και εκτύπωση του συνδέσμου (ιδιότητα href) από <a> tags και του κειμένου

τους (ό,τι βρίσκεται δηλαδή μεταξύ των <a> και).

(r'<a.+?href="<.*?)"/*?>(.*?)',re.DOTALL)

Παραπάνω συνδιάζουμε τον σύνδεσμο της ιδιότητας href και τα περιεχόμενα των <a> tags. Η τελεία και οι τελεστές (?, *) ταιριάζουν το κείμενο στην αρχή των tags <a μέχρι και την ιδιότητα href και σε οτιδήποτε βρίσκεται εντός των " " μέχρι και το > και ανάμεσα και στα <a> . Τέλος, οι παρενθέσεις μέσα στα " " είναι για την αποθήκευση στο group(1) , όπως επίσης και εντός των <a> αποθηκεύοντας έτσι το αντίστοιχο περιεχόμενο και στο group(2).

Ερώτημα No5:

Απαλοιφή όλων των tags από το κείμενο.

a. (r'<.+?>|</.+?>',re.DOTALL)

b. (r'<.+?/>',re.DOTALL)

Στην πρώτη έκφραση περιέχονται tags της πρώτης κατηγορίας που ξεκινάει με <a> και τελειώνει με , ενώ στην δεύτερη κατηγορία ανήκουν τα self closing tags(πχ <meta /> και </meta>)

Ερώτημα No6:

Μετατροπή των ειδικών HTML entries που υπάρχουν στο κείμενο σύμφωνα με τον

πίνακα

('r&(amp | gt | lt | nbsp);')

Στόχος της συγκεκριμένης έκφρασης να συνδεθούν τα amp, gt, lt, nbsp και με την χρήση του συμβόλου της εναλλαγής (|) για να ταιριάζει κάθε φορα μια από τις 4 πιθανές επιλογές

Ερώτημα No7:

Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitspace σε ένα ακριβώς κενό,

βλ. και (link) (εδώ όμως διατηρούμε τα σημεία στίξης!).

(r'\s+')

Με την χρήση αυτής της έκφρασης ταιριάζουμε τις ακολουθίες συνεχόμενων χαρακτήρων whitespace