Μεταγλωτιστές 2020

Προγραμματιστική Εργασία 2

Ονοματεπώνυμο : Ελευθέριος Ελευθεριάδης

ΑΜ : Π2017166

Ερώτημα 1ο:

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

(‘<title>(.+?)</title>’)

Χρήση της τελείας(.) και του plus sign (+) για να ταιριαστεί ό,τι βρίσκεται μεταξύ <title> και </title> (είναι απαραίτητη η ύπαρξη τουλάχιστον 1ος χαρακτήρα).

Ερώτημα 2ο:

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

(‘<!--.*?-->’,re.DOTALL)

Χρήση της τελείας(.) και του αστερίσκου (*) για να ταιριαστεί ό,τι βρίσκεται μεταξύ <!-- και -- > (δεν είναι απαραίτητη η ύπαρξη χαρακτήρα γιατί γίνεται η χρήση του αστερίσκου αντί του plus sign.

Ερώτημα 3ο:

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).

Χρήση του (s(?:cript | tyle) για ταίριασμα των tags <script></script> και <style></style>. Χρήση του \1 για να ταιριάξει ό,τι βρίσκεται στο group(1).

Ερώτημα 4ο:

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)

Πιο πάνω ταιριάζουμε τον σύνδεσμο της ιδιότητας href και τα περιεχόμενα μεταξύ <a> και . Η τελεία και οι τελεστές επανάληψης ? Και * ταιριάζουν το κείμενο από το <a μέχρι την ιδιότητα href σε ότι είναι εντός των “” μέχρι το > και μεταξύ των <a> και . Οι παρενθέσεις αποθηκεύουν των σύνδεσμο στο group(1) και των αντίστοιχο γίνεται στο group(2).

Ερώτημα 5ο:

Απαλοιφή όλων των tags από το κείμενο.

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

Η πρώτη ψάχνει tag της πρώτης κατηγορίας. Στην οποία περιέχονται tag που ξεκινούν με <a> και κλείνουν με . και το δεύτερο ψάχνει self closing tags.

Ερώτημα 6ο:

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

(‘r&(amp | gt | lt | nbsp);’)

Η κάθετος χρησιμοποιείται για αλλαγή της επιλογής μεταξύ των amp, gt, lt και nbsp.

Ερώτημα 7ο:

Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitspace σε ένα ακριβώς κενό, βλ. και (link) (εδώ όμως διατηρούμε τα σημεία στίξης!).

(r’\s+’)

Ταιριάζουμε χαρακτήρες whitespace.

Ερώτημα 8ο:

Στο τέλος τυπώστε το κείμενο όπως έχει διαμορφωθεί μετά τις προηγούμενες μετατροπές. Η έξοδος του προγράματος σας περιλαμβάνεται στα παραδοτέα της εργασίας.