Μεταγλωττιστές 2020
Προγραμματιστική Εργασία#2
Ημερομηνία Υποβολής:29/05/2020

Ονοματεπώνυμο:Πασχάλη Μαρία
Α.Μ:Π2016033

Σύντομη αναφορά για τον κώδικα

Έκανα ένα ένα τα ερωτήματα με τη σειρά.Πρώτα ορίζω την κανονική έκφραση και μετά ανάλογα με την εκφώνηση είτε επιστρέφω στο κείμενο ό,τι ταιριάζει(ερωτήματα για τίτλο και συνδέσμους) είτε αντικαθιστώ με κενό όπου ζητείται.

Εξήγηση κανονικών εκφράσεων :

1. <title>(.*)</title> : <title>(ένας ή περισσότεροι οποιοιδήποτε χαρακτήρες εκτός από new line)</title>.Το κομμάτι στην παρένθεση είναι αυτό που κρατάω στο text
2. <!.+?> : <! ένας ή περισσότεροι χαρακτήρες > (δεδομένου ότι τα σχόλια είναι της μορφής <!--οτιδήποτε-->)χρήση μη άπληστων τελεστών για να πάρουμε το εσωτερικό κάθε σχολίου ξεχωριστά
3. <script.*?</script>|<style.*?</style> : <script 0 ή περισσότεροι χαρακτήρες</script> **ή** <style 0 ή περισσότεροι χαρακτήρες </style>
4. <a.*? href="(.*)">.* : <a 0 ή περισσότεροι χαρακτήρες href=”(οποιοσδήποτε χαρακτήρας πολλές φορές)”>οποιοσδήποτε χαρακτήρας πολλές φορές .Η παρένθεση έχει μπει στο σύνδεσμο,το κομμάτι που θέλουμε να κρατήσουμε
5. <.+?> : <ένας ή περισσότεροι χαρακτήρες>,αναγνώριση των tags και του περιεχομένου τους
6. (&)|(>)|(<)|( ) : (&)ή(>)ή(<)ή( )
7. \s+ : ένας ή περισσότεροι whitespace χαρακτήρες

Υπάρχουν και σχόλια επεξήγης πάνω στον κώδικα για τη χρήση του flag re.DOTALL,για τη συνάρτηση sub και τι επιστρέφει κάθε φορά,για την finditer και τι επιστρέφει και αυτή κάθε φορά και για την cb για την αντικατάσταση των οντοτήτων html.

Πηγές

1. http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html
2. Τα gists των εργαστηρίων
3. https://www.tutorialspoint.com/python/python_reg_expressions.htm
4. http://regexlib.com/Default.aspx