Μεταγλωττιστές 2020

Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Ελευθέριος Μπαΐλης

ΑΜ: Π2017080

Περιγραφή επεξεργασίας και κανονικών εκφράσεων

Πρώτο βήμα είναι η προσθήκη της βιβλιοθήκης re στον κώδικα και έπειτα το άνοιγμα του txt αρχείου που παραθέτει το περιεχόμενό της στην μεταβλητή text. Έπειτα:

i. Για την εκτύπωση του τίτλου χρησιμοποιείται η κανονική έκφραση **<title>(.+?)</title>** όπου αναγνωρίζει οτιδήποτε βρίσκεται μεταξύ των tags και με την εκτύπωση του group(1) εμφανίζεται ο τίτλος.
ii. Για την απαλοιφή των σχολίων χρησιμοποιείται η κανονική έκφραση **<!--.+?-->** όπου θα αναγνωρίσει τόσο και τα tags όσο και το σχόλιο και με την sub θα γίνει η απαλοιφή του.
iii. Παρόμοια με την απαλοιφή των σχολίων η κανονική έκφραση **<script(.+?)</script>|style=(.+?)</style>** θα αναγνωρίσει την κάθε περίπτωση και με την sub θα πετύχει απαλοιφή των tags και του περιεχομένου τους.
iv. Η κανονική έκφραση **<a.+?href="(.*?)".*?>(.*?)** θα αναγνωρίσει όλο το περιεχόμενο μεταξύ των Tags καθώς και τα ίδι τα tags αλλά με την εκτύπωση του group(1) θα εμφανίσει μόνο το link καθώς και τον τίτλο του link.
v. Η επόμενη κανονική έκφραση αναγνωρίζει όλες τις μορφές που μπορούν να έχουν τα tags στην HTML και τα διαγράφει με την sub
vi. Η χρήση της έκφρασης **(&|>|<|)** αναγνωρίζει τα HTML entities και μέσω της callback συνάρτησης η οποία καλείται μέσω της sub αντικαθιστά ανάλογα τα entities.
vii. Τέλος η κανονική έκφραση \s+ αναγνωρίζονται συνεχόμενα spaces καθώς και tabs και απαλείφονται μέσω της sub.

Στις κανονικές εκφράσεις 2,3 και 4 ήταν απαραίτητη η χρήση του flag re.DOTALL ώστε η αναγνώριση να γίνεται ακόμα στον χαρακτήρα του newline