Μεταγλωττιστές 2020
Προγραμματιστική Εργασία #2
Ονοματεπώνυμο : Παππά Δήμητρα
ΑΜ : Π2015015

Το πρώτο βήμα του κώδικα είναι η εισαγωγή της βιβλιοθήκης re κανονικών εκφράσεων.
Στη συνέχεια γίνεται άνοιγμα του ζητούμενου αρχείου για ανάγνωση και εκχωρείται στην μεταβλητή text.

1. Για το πρώτο βήμα επεξεργασίας χρησιμοποιήθηκε η έκφραση <title>(.+?)</title> κατά την οποία αναγωρίζονται όλοι οι τίτλοι μέσα στο tag. Έγινε εύρεση και εκτύπωση του τίτλου.
2. Στο δεύτερο βήμα, χρησιμοποιήθηκε η έκφραση <--!(*?)--> κατά την οποία αναγνωρίζονται τα σχόλια ανάμεσα στο tag και η αφαίρεσή τους έγινε με την χρήση της sub καθώς αντικαταστήθηκαν με κενό (whitespace). Τα νέα αποτελέσματα εισάγονται στην μεταβλητή text του κειμένου.
3. Στο τρίτο βήμα, χρησιμοποιήθηκε η έκφραση <script(.*?)</script>|style=(.*?) για την αναγνώριση των scripts & styles tags στο κείμενο και η αφαίρεσή τους έγινε με την χρήση της sub καθώς αντικαταστήθηκαν με κενό (whitespace). Τα νέα αποτελέσματα εισάγονται στην μεταβλητή text του κειμένου.
4. Στο τέταρτο βήμα, χρησιμοποιήθηκε η έκφραση r'<a.+?href="(.*?)".*?>(.*?)' κατά την οποία αναγνωρίζονται οι σύνδεσμοι του κειμένου και ο τίτλος των συνδέσμων μεταξύ των tags <a> και γίνεται η εκτύπωσή τους. Τα νέα αποτελέσματα εισάγονται στην μεταβλητή text του κειμένου.
5. Στο πέμπτο βήμα, έγινε χρήση της έκφρασης r'<.+?>(.*?)</.?>' κατά την οποία αναγνωρίζονται τα html tags του κειμένου και η απαλοιφή τους έγινε με την βοήθεια της sub καθώς αντικαταστήθηκαν με κενό (whitespace). Τα νέα αποτελέσματα εισάγονται στην μεταβλητή text του κειμένου.
6. Στο έκτο βήμα, έγινε δημιουργία μιας function κατά την οποία γίνεται αντικατάσταση των &, &gt, &lt,   με &, >, <, (whitespace) αντίστοιχα. Έπειτα χρησιμοποιείται η έκφραση r'&(amp|gt|lt|nbsp); για την αναγνώριση των html entities. Γίνεται κλήση της function μέσω της sub γίνεται η εισαγωγή των νέων αποτελεσμάτων στην μεταβλητή text του κειμένου.
7. Στο έκτο βήμα, χρησιμοποιήθηκε η έκφραση r'\s+' κατά την οποία αναγνωρίζονται όλα τα κενά σύμβολα (ένα ή και περισσότερα από ένα). Με την βοήθεια της sub έγινε η αντικατάστασή τους με έναν χαρακτήρα κενού (whitespace). Τα νέα αποτελέσματα εισάγονται στην μεταβλητή text του κειμένου.
8. Στο τέλος γίνεται η εκτύπωση του κειμένου με όλες τις αλλαγές που έχουν εισαχθεί.

Σχόλια: Η χρήση της re.DOTALL στα βήματα 2,3,4,5 έγινε ώστε ο χαρακτήρας της τελείας (.) να ταιριάζει με οποιονδήποτε χαρακτήρα, συμπεριλαμβανομένης και της νέας γραμμής (newline). Αυτό γίνεται γιατί επεξεργαζόμαστε ένα κείμενο το οποίο έχουμε μετατρέψει σε strings πολλών γραμμών.
Για την ολοκλήρωση της εργασίας ως πηγές χρησιμοποιήθηκαν οι σημειώσεις του μαθήματος
https://gist.github.com/mixstef/66cf94177914eee5a68facdd2749a955
καθώς και την ιστοσελίδα της python για την βιβλιοθήκη re κανονικών εκφράσεων
https://docs.python.org/2/library/re.html#