Μεταγλωτιστές 2020
Προγραμματιστική Εργασία #2
Ονοματεπώνυμο: Κιορπελίδης Κωνσταντίνος
ΑΜ: Π2016096

Βήμα 1:
Εξαγωγή και εκτύπωση τίτλου. Χρήση (‘<title>(.+?)</title>’), όπου τα σύμβολα “.” και “+” χρησιμοποιούνται για το ταίριασμα κάθε χαρακτήρα και τουλάχιστον ενός.
Βήμα 2:
Απαλοιγή των σχολίων. Χρήση ('<!--.*?-→',re.DOTALL), όπου τα σύμβολα “.” και “*” ταιριάζουν οτιδήποτε υπάρχει ανάμεσα ακόμα και αν δεν υπάρχει χαρακτήρας.
Βήμα 3:
Απαλοιφή των <script> και <style> tags με όλο τους το περιεχόμενο, μέχρι δηλαδή να συναντήσετε το αντίστοιχο </script> ή </style>. Χρήση (r'<(s(?:cript|tyle)).*?>.*?</\1>',re.DOTALL), όπου το <(s(?:cript|tyle)).*?> εξυπηρετεί ταυτόχρονα την αρχή και των δύο tags και του \1 για αντιστοιχία όποιου από τα δύο έχει ήδη συναντήσει.
Βήμα 4:
Εξαγωγή και εκτύπωση του συνδέσμου (ιδιότητα href) από <a> tags και του κειμένου τους. Χρήση (r'<a.+?href="(.*?)".*?>(.*?)',re.DOTALL) .
Βήμα 5:
Απαλοιφή όλων των tags. Χρήση (r'<("[^"]*"|\'[^\']*\'|[^\'">])*>',re.DOTALL), όπου φροντίζει να αφαιρέσει κάθε υπολοιπόμενο συνδυασμό.
Βήμα 6:
Μετατροπή των ειδικών HTML entities που υπάρχουν στο κείμενο. Χρήση (r'&(nbsp|gt|lt|amp);'), όπου ελέγχει κάθε μία από τις 4 εκδοχές.
Βήμα 7:
Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό. Χρήση (r'\s+').
Βήμα 8:
Εκτύπωση όλου του κειμένου μετά την εφαρμογή των αλλαγών.