Μεταγλωττιστές 2020
Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Αθανασάκη Παναγιώτα
Α.Μ.: Π2016003

Για την υλοποίηση της άσκησης χρησιμοποιήθηκε το πρόγραμμα Python 3.8 και ο κώδικας έτρεχε στο Python 3.8.2 Shell. Πραγματοποιήθηκε μελέτη στις σημειώσεις του μαθήματος. Ακολουθεί μια περιγραφή, σχετικά με το πώς έγινε η επεξεργασία των βημάτων στα ζητούμενα της άσκησης.

1. Εξαγωγή και εκτύπωση του τίτλου
Χρησιμοποιείται η κανονική έκφραση, rexp1 = re.compile(r'<title>(.+?)</title>') η οποία με την (.) συμβολίζει ότι ακολουθεί ένας οποιοσδήποτε χαρακτήρας, εκτός από newline, και το (+?) είναι η μη-άπληστη μορφή των τελεστών επανάληψης, όπου σταματάει το ταίριασμα.

2. Απαλοιφή των σχολίων
Για την απαλοιφή των σχολίων χρησιμοποιείται η κανονική έκφραση, rexp2 = re.compile(r'<!--.*?-->',re.DOTALL). Επιτυγχάνεται ταίριασμα με οποιοδήποτε σχόλιο εντός του κειμένου, το οποίο βρίσκεται μεταξύ <!-- και →. Η εντολή re.DOTALL, χρησιμοποιήθηκε επειδή μπορεί να υπάρχουν σχόλια πολλαπλών γραμμών.

3. Απαλοιφή script
pscript=(r'<script(.*?)</script>',re.DOTALL)

4. Απαλοιφή style
pstyle=(r'<style(.*?)</style>',re.DOTALL)

5. Απαλοιφή script και style με μία κανονική έκφραση

rexp3=re.compile(r'<(script|style).*?>.*?</(script|style)>',re.DOTALL)

Χρησιμοποιείται ο τελεστής | για να επιλέγονται και τα δύο και οι τελεστές (.*) για να γίνει ταίριασμα οποιουδήποτε χαρακτήρα.

6. Εξαγωγή και εκτύπωση του συνδέσμου (ιδιότητα href)

rexp4 = re.compile(r'<a.+?href="(.*?)".*?>(.*?)',re.DOTALL)

Ταίριασμα συνδέσμου href και περιεχομένου μεταξύ <a> και .

7. Απαλοιφή όλων των tags από το κείμενο

Έχουμε δύο περιπτώσεις, καθώς ένα tag μπορεί να είναι self-closing.

rexp5 = re.compile(r'<.+?>|</.+?>',re.DOTALL)

Απαλοιφή tags με διπλή μορφή, χρησιμοποιείται ο τελεστής εναλλαγής.

rexp5x = re.compile(r'<.+?/>',re.DOTALL)

Απαλοιφή tags με μονή μορφή-self-closing tags.

8. HTML entities

rexp6 = re.compile(r'&(amp|gt|lt|nbsp);')

Χρησιμοποιείται ο τελεστής της εναλλαγής, έτσι ώστε σε κάθε ταίριασμα να γίνεται χρήση μιας από τις 4 πιθανές επιλογές. Η μετατροπή των html entities όπως ζητείται γίνεται με τη χρήση μιας συνάρτησης και της δομή επανάληψης if-else.

9. Whitespace

rexp7 = re.compile(r'\s+')

Εξαγωγή whitespaces μία ή περισσότερες φορές. Ο συνδυασμός χαρακτήρων (\s), αντιπροσωπεύει τον χαρακτήρα whitespace και ο χαρακτήρας (+) δηλώνει το ταίριασμα ενός ή περισσότερων whitespace.

10. Άνοιγμα αρχείου και έξοδος προγράμματος

Αφού ολοκληρώθηκαν οι κανονικές εκφράσεις, έγινε η ανάγνωση του html αρχείου. Χρησιμοποιήθηκε μια μεταβλητή fp, για να διαβάζει το κείμενο από το αρχείο εισόδου. Τέλος, γίνεται η εκτύπωση του κειμένου, όπως έχει διαμορφωθεί από τις προηγούμενες αλλαγές.