

Μεταγλωτιστές 2020

Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Χρήστος Δήμας

A.M.: Π2017204

Αρχικά έκανα `import` το module `re` της `python`. Το επόμενο βήμα ήταν να δημιουργήσω μία συνάρτηση `function`, η οποία θα καλείται μέσω της μεθόδου `sub` για την αντικατάσταση των χαρακτήρων `<>` του εξωτερικού αρχείου `testpage.txt`. Έπειτα, δημιούργησα τις μηχανές ταιριάσματος και τις κανονικές εκφράσεις. Τέλος, διάβασα το αρχείο `testpage.txt`, όπου και έγινε η χρήση των μηχανών ταιριάσματος, ώστε να γίνουν οι επιθυμητές απαλοιφές και εξαγωγές του κειμένου. Αφού έγιναν τα παραπάνω βήματα πραγματοποιήθηκε η τελική εκτύπωση του νέου κειμένου.

Περιγραφή των κανονικών εκφράσεων στα πλαίσια των ερωτημάτων της εργασίας:

Ερώτημα 1:

`('<title>(.*?)</title>')`

Έγινε η χρήση της παραπάνω έκφρασης για την επιλογή των χαρακτήρων που βρίσκονται στο tag `title`. Πιο συγκεκριμένα χρησιμοποιήθηκε ο τελεστής `.` και `+` για να γίνει το ταιρίασμα στους χαρακτήρες που βρίσκονται ανάμεσα στα `title`.

Ερώτημα 2:

`('<!--.*?-->',re.DOTALL)`

Έγινε η χρήση της παραπάνω έκφρασης για το ταιρίασμα όλων των σχολίων, τα οποία βρίσκονται ανάμεσα `<!--` και `-->` και χρησιμοποιώντας παράλληλα τους τελεστές `.` και `*`. Για την απαλοιφή ολόκληρων των σχολίων δεν χρησιμοποιήθηκαν παρενθέσεις αλλά η χρήση του `group(0)`.

Ερώτημα 3:

`(r'<(s(?:cript|tyle)).*?>.*?</\1>',re.DOTALL)`

Έγινε η χρήση της παραπάνω έκφρασης για το ταιρίασμα του περιεχομένου των tags `<script></script>` και `<style></style>` χρησιμοποιώντας παράλληλα τους τελεστές `.` και `*` για το ταιρίασμα των χαρακτήρων μετά τα `<script` και `<style` μέχρι τον

χαρακτήρα > καθώς επίσης και η χρήση του \1 για να ταιριάζει ότι έχει βρεθεί νωρίτερα από / και βρίσκεται στο group(1).

Ερώτημα 4:

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

Έγινε η χρήση της παραπάνω έκφρασης για το ταίριασμα των περιεχομένων του href και των tags <a>. Πιο συγκεκριμένα ότι βρίσκεται μεταξύ του <a και του href, οι σύνδεσμοι που είναι μέσα σε " " και τέλος μεταξύ σε " " και >.

Ερώτημα 5:

```
(r'<.+?>|</.+?>',re.DOTALL)
```

```
(r'<.+?/>',re.DOTALL)
```

Έγινε η χρήση των δύο παραπάνω εκφράσεων για το ταίριασμα των περιεχομένων των tags < > </> καθώς επίσης και για τα self closing tags </>.

Ερώτημα 6:

```
(r'&(amp|gt|lt|nbsp);')
```

Έγινε η χρήση της παραπάνω έκφρασης για το ταίριασμα με τα &>< sp.

Ερώτημα 7:

```
(r'\s+')
```

Έγινε η χρήση της παραπάνω έκφρασης για το ταίριασμα με τα whitespaces, όπου \s αντιπροσωπεύει τον χαρακτήρα whitespace.