ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΜΕΤΑΓΛΩΤΤΙΣΤΕΣ

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΕΡΓΑΣΙΑ #2

ΔΙΔΑΣΚΩΝ: ΜΙΧΑΗΛ ΣΤΕΦΑΝΙΔΑΚΗΣ

ΓΙΩΡΓΟΣ ΜΑΡΓΑΡΙΤΗΣ (Π2015139)

ΜΑΪΟΣ 2020

Αρχικά πραγματοποιείται η εισαγωγή του re module για τις regular expressions. Στη συνέχεια ακολουθεί η callback function της sub() με όνομα cb η οποία είναι υπεύθυνη για την αντικατάσταση των HTML entities. Έπειτα υπάρχουν οι απαραίτητες κανονικές εκφράσεις για την εκπλήρωση των ζητημάτων. Γίνεται διάβασμα του αρχείου testpage.txt και εφαρμογή των κανονικών εκφράσεων, ενώ, ως τελευταίο βήμα εκτυπώνεται η τελική μορφή του κειμένου.

Ζητούμενο 1

Εξαγωγή και εκτύπωση του τίτλου (οτιδήποτε βρίσκεται μεταξύ <title> και </title>).

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

Επιλογή οποιουδήποτε χαρακτήρα `(.)` βρίσκεται ανάμεσα από τα `<title> </title>` 1 ή περισσότερες φορές `(+?)`.

Ζητούμενο 2

Απαλοιφή των σχολίων (οτιδήποτε βρίσκεται μεταξύ <!-- και -->).

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

Επιλογή οποιουδήποτε χαρακτήρα `(.)` βρίσκεται ανάμεσα από τα `<!-- -->` 0 ή περισσότερες φορές `(*?)` διότι μπορεί να υπάρχει κενό σχόλιο.

Έγινε χρήση του `re.DOTALL` γιατί ενδέχεται να υπάρχουν σχόλια πολλαπλών γραμμών.

Ζητούμενο 3

Απαλοιφή των <script> και <style> tags με όλο τους το περιεχόμενο, μέχρι δηλαδή νασυναντήσετε το αντίστοιχο </script> ή </style> (και τα τελευταία).

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

Επιλογή ολόκληρων των `<script>`, `<style>` tags με τη χρήση του μη-άπληστου τελεστή `(*?)` αλλά και backreferencing στο `group(1)`.

Χρήση του `re.DOTALL` γιατί ενδέχεται τα tags να επεκτείνονται σε πολλές γραμμές.

Ζητούμενο 4

Εξαγωγή και εκτύπωση του συνδέσμου (ιδιότητα href) από <a> tags και του κειμένου τους(ό,τι βρίσκεται δηλαδή μεταξύ των <a> και).

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

Γίνεται ταίριασμα του `href` καθώς και των περιεχομένων των `<a> </a>` tags. Ταίριαγμα χαρακτήρων ανάμεσα από το `<a` και του `href` , στη συνέχεια ανάμεσα από τα `""`, μεταξύ του `"` και του `>` και τέλος ανάμεσα στα `<a> </a>`.

Χρησιμοποιείται το `re.DOTALL` γιατί ενδέχεται τα tags να επεκτείνονται σε πολλές γραμμές.

Ζητούμενο 5

Απαλοιφή όλων των tags από το κείμενο.

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

Επιλογή οποιουδήποτε χαρακτήρα `(.)` βρίσκεται ανάμεσα από τα `< >` 0 ή περισσότερες φορές `(*?)` διότι μπορεί να υπάρχει κενό tag. Έγινε χρήση του `re.DOTALL` γιατί ενδέχεται τα tags να επεκτείνονται σε πολλές γραμμές.

Ζητούμενο 6

Μετατροπή των ειδικών HTML entities που υπάρχουν στο κείμενο σύμφωνα με τον πίνακα που δίνεται.

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

Επιλογή των HTML entities χρησιμοποιόντας το σύμβολο είτε `(|)`.

Ζητούμενο 7

Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό.

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

Εξαγωγή των whitespaces `(\s)` μία ή περισσότερες φορές `(+)`.

Ζητούμενο 8

Τυπώστε το κείμενο, όπως έχει διαμορφωθεί μετά τις προηγούμενες μετατροπές.

```
print(text)
```

Εκτύπωση του κειμένου στη τελική του μορφή.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s050-sub-callback-py

2. https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s010-hint-keep-only-words-py

3. http://mixstef.github.io/courses/compilers/lecturedoc/appendix-python/module1.html#id5

4. http://mixstef.githourub.io/cses/compilers/lecturedoc/unit2/module1.html#id8

5. http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#sub