Μεταγλωττιστές 2023
Προγραμματιστική Εργασία #2

Ζητούμενο

Ο στόχος της άσκησης είναι να κατασκευάσετε πρόγραμμα Python3, το οποίο θα χρησιμοποιεί (**αποκλειστικά και μόνον**) τη βιβλιοθήκη `re` των κανονικών εκφράσεων για να επεξεργαστεί κείμενο HTML ιστοσελίδας.

Η εργασία **πρέπει γίνει σε ένα jupyter notebook**, το οποίο θα παραδοθεί στο opencourses.

Προσοχή: χρησιμοποιήστε ένα **διαφορετικό κελί** για **κάθε ένα από τα βήματα** που ζητούνται στη συνέχεια. Τα βήματα επεξεργασίας πρέπει να γίνουν το ένα μετά το άλλο, με ξεχωριστή κανονική έκφραση για το κάθε βήμα.

Βήματα υλοποίησης

Τα βήματα επεξεργασίας που ζητούνται είναι τα ακόλουθα:

1. **Ανάγνωση του αρχείου εισόδου σε ένα string**

   Θα βρείτε το αρχείο εισόδου που θα χρησιμοποιήσετε στη διεύθυνση

   http://mixstef.github.io/courses/compilers/testpage.txt

   Χρησιμοποιήστε τον κώδικα από τις σημειώσεις του εργαστηρίου

   http://mixstef.github.io/courses/compilers/lecturedoc/appendix-python/module1.html#id13

   για την ανάγνωση **όλου του κειμένου της ιστοσελίδας** από το αρχείο εισόδου **testpage.txt** σε μια μεταβλητή string πριν ξεκινήσετε την επεξεργασία.

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

   Η εκτέλεση του αντίστοιχου κελιού θα πρέπει να δίνει τον τίτλο της ιστοσελίδας χωρίς τα `<title>` και `</title>`.

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

   Αντικαταστήστε τα σχόλια με έναν κενό χαρακτήρα (space). Το νέο string που προκύπτει θα χρησιμοποιηθεί ως είσοδος στο επόμενο βήμα.

   Υπόδειξη: σκεφτείτε αν πρέπει να χρησιμοποιήσετε το flag `re.DOTALL` (ισχύει και για τα άλλα βήματα).

4. **Απαλοιφή των `<script>` και `<style>` tags**

   μαζί με όλο τους το περιεχόμενο, μέχρι δηλαδή να συναντήσετε το αντίστοιχο `</script>` ή `</style>` (και τα τελευταία). **Χρησιμοποιήστε backreference** στην κανονική έκφραση. Αντικαταστήστε με έναν κενό χαρακτήρα (space). Το νέο string που προκύπτει θα χρησιμοποιηθεί ως είσοδος στο επόμενο βήμα.

5. **Εξαγωγή και εκτύπωση του συνδέσμου (ιδιότητα `href`) από `<a>` tags και του κειμένου**

τους (ό,τι βρίσκεται δηλαδή μεταξύ των <a> και)

Η εκτέλεση του κελιού θα πρέπει να τυπώνει σε ένα loop το link και το κείμενο του συνδέσμου για κάθε http/https υπερσύνδεσμο της σελίδας. Η έξοδος θα πρέπει να φαίνεται στο notebook που θα παραδώσετε.

6. **Απαλοιφή όλων των tags από το κείμενο**

   Αντικαταστήστε οποιοδήποτε tag υπάρχει ακόμα (είτε tag αρχής είτε tag τέλους) με έναν κενό χαρακτήρα (space). Το νέο string που προκύπτει θα χρησιμοποιηθεί ως είσοδος στο επόμενο βήμα.

7. **Μετατροπή των ειδικών HTML entities που υπάρχουν στο κείμενο**

   σύμφωνα με τον παρακάτω πίνακα:

HTML entities	Αντικαταστήστε με
`&`	`&`
`>`	`>`
`<`	`<`
` `	`χαρακτήρα space`

   Για τη μετατροπή χρησιμοποιήστε τη μέθοδο `sub()` με «**callback**» συνάρτηση. Το νέο string που προκύπτει θα χρησιμοποιηθεί ως είσοδος στο επόμενο βήμα.

8. **Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό**

   Το νέο string που προκύπτει θα χρησιμοποιηθεί ως είσοδος στο επόμενο βήμα.

9. **Τυπώστε το κείμενο όπως έχει διαμορφωθεί μετά τις προηγούμενες μετατροπές (με print)**

   Στο notebook που θα παραδώσετε **πρέπει να φαίνεται πλήρως η έξοδος** της εκτέλεσης αυτού του κελιού.

<u>Παράδοση εργασίας</u>

Ανεβάστε το τελικό σας notebook στο opencourses (**Εργασία 2**) έως και τη **Δευτέρα 3/4**.
(Ίσως χρειαστεί να κάνετε zip το αρχείο ipynb για να γίνει αποδεκτό από το σύστημα)