Μεταγλωττιστές 2020
Προγραμματιστική Εργασία #2

(Προσοχή: η παράδοση της άσκησης θα γίνει μέσω github. Διαβάστε τις οδηγίες στο τέλος της εκφώνησης)

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.

Τα βήματα επεξεργασίας που ζητούνται είναι τα ακόλουθα:

1. Εξαγωγή και *εκτύπωση* του τίτλου (οτιδήποτε βρίσκεται μεταξύ `<title>` και `</title>`).
2. *Απαλοιφή* των σχολίων (οτιδήποτε βρίσκεται μεταξύ `<!--` και `-->`).
3. *Απαλοιφή* των `<script>` και `<style>` tags με όλο τους το περιεχόμενο, μέχρι δηλαδή να συναντήσετε το αντίστοιχο `</script>` ή `</style>` (και τα τελευταία).
4. Εξαγωγή και *εκτύπωση* **του συνδέσμου** (ιδιότητα `href`) από `<a>` tags και **του κειμένου** τους (ό,τι βρίσκεται δηλαδή μεταξύ των `<a>` και `</a>`).
5. *Απαλοιφή* **όλων των tags** από το κείμενο.
6. *Μετατροπή* των ειδικών HTML entities που υπάρχουν στο κείμενο σύμφωνα με τον παρακάτω πίνακα:

HTML entities	Αντικαταστήστε με
`&amp;`	`&`
`&gt;`	`>`
`&lt;`	`<`
` `	`χαρακτήρα space`

   Για τη μετατροπή χρησιμοποιήστε τη μέθοδο `sub()` με «callback» συνάρτηση, βλ. παράδειγμα στο https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s050-sub-callback-py.
7. *Μετατροπή* ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό, βλ. και https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s010-hint-keep-only-words-py (εδώ όμως *διατηρούμε* τα σημεία στίξης!).
8. Στο τέλος *τυπώστε* και το κείμενο, όπως έχει διαμορφωθεί μετά τις προηγούμενες μετατροπές. **Η έξοδος του προγράμματός σας περιλαμβάνεται στα παραδοτέα της εργασίας.**

Τα βήματα επεξεργασίας πρέπει να γίνουν το ένα μετά το άλλο, με ξεχωριστή κανονική έκφραση (ή εκφράσεις) το κάθε βήμα.

Ως κείμενο εισόδου (ιστοσελίδα HTML) θα πρέπει να χρησιμοποιήσετε το αρχείο **testpage.txt**, το οποίο θα βρείτε στο repository της άσκησης (βλ. ενότητα «Παραδοτέο» στο τέλος της άσκησης) ή στο http://mixstef.github.io/courses/compilers/testpage.txt).

Υποδείξεις

- Χρησιμοποιήστε τον κώδικα από τις σημειώσεις του εργαστηρίου

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

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

- Διερευνήστε αν και πότε πρέπει να χρησιμοποιήσετε το flag `re.DOTALL` (βλ. http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#id8 στις σημειώσεις) σε κάποιες από τις κανονικές εκφράσεις σας.
- Όπου ζητείται *απαλοιφή*, αντικαταστήστε με έναν κενό χαρακτήρα (space).
- Μπορείτε να επιτύχετε την απαλοιφή των `<script>` και `<style>` tags με *μία μόνο* κανονική έκφραση; Δοκιμάστε τη χρήση backreferences στην κανονική έκφραση, βλ. http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#sub

Παραδοτέο

Η παράδοση θα γίνει μέσω github. Οδηγίες:

1. Αντιγράψτε (**fork**) το repository **https://github.com/mixstef/compilers1920a2** στο δικό σας repository. Βεβαιωθείτε ότι δουλεύετε αποκλειστικά στο **master branch**.
2. Τροποποιήστε κατάλληλα τα αρχεία που περιέχονται στο repository σας με το δικό σας περιεχόμενο:
   - Συμπληρώστε τα στοιχεία σας στο αρχείο **README.md** .
   - Γράψτε τον κώδικά σας στο αρχείο **html-processor.py** .
   - Τοποθετήστε την έξοδο του προγράμματός σας στο αρχείο **output.txt** .
   - Προσθέστε την αναφορά σας ως **report.pdf** .
   - **Προσοχή: πρέπει να διατηρήσετε τα ονόματα των παραπάνω αρχείων!**
3. Ενημερώστε το repository σας στο github εντός προθεσμίας. **Μην κάνετε pull request!**

Η εργασία είναι αυστηρά ατομική. Για την εγκυρότητα της υποβολής σας θα χρησιμοποιηθεί η χρονοσήμανση των αλλαγών (commits) των αρχείων σας.

Προθεσμία παράδοσης: Παρασκευή 29/5/2020.