**Μεταγλωτιστές 2020**
**Προγραμματιστική Εργασία #2**

**Ονοματεπώνυμο:** ΠΑΠΑΠΟΣΤΟΛΟΥ ΜΥΡΤΩ

**ΑΜ:** Π2016030

*Το πρόγραμμα της δεύτερης εργασίας αρχίζει με την εισαγωγή της βιβλιοθήκης "re" και του αρχείου "testpage.txt", το οποίο διαβάστηκε και εισάχθηκε σε μία μεταβλητή, όλο του το περιεχόμενο. Πριν όμως ανοιχθεί το αρχείο δημιουργήθηκαν οι κανονικές εκφράσεις και οι έλεγχοι της callback συνάρτησης.*

*Κατόπιν εφαρμόστηκαν τα ζητούμενα της άσκησης,*

1. *Εξαγωγή και εκτύπωση του τίτλου με την κανονική έκφραση, '<title>(.+)</title>',*
2. *Απαλοιφή των σχολίων , '<!--.+?-->', χρησιμοποιώντας την μέθοδο sub() με «callback» συνάρτηση, η οποία αντικαθιστά το περιεχόμενο το οποίο ταίριαξε με ένα χαρακτήρα space.*
3. *Απαλοιφή των <script> και <style> tags με όλο τους το περιεχόμενο, με την κανονική έκφραση '<(script|style).*?>.*?</\1>', γίνεται χρήση backreference έτσι ώστε στο σημείο \1 να ταιριάζει ότι ταιριάζει στο group(1).*
4. *Εξαγωγή και εκτύπωση του συνδέσμου από <a> tags και του κειμένου τους με την έκφραση '<a.*?href="(.+?)".*?>(.*?)</a>' .*
5. *Απαλοιφή όλων των tags από το κείμενο, με την έκφραση '<.+?>' .*
6. *Μετατροπή των ειδικών HTML entities που υπάρχουν στο κείμενο σύμφωνα με τον δοθέντα πίνακα. Αυτή η μετατροπή γίνεται με την έκφραση, '&|>|<|  ' και με την χρήση της συνάρτησης function που καλείται από τη μέθοδο sub(), μετατρέπονται σε χαρακτήρες &,>,<, κενό (space) αντίστοιχα.*
7. *Μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό, r'\s+' .*

*Τέλος, εκτυπώνονται όλα τα παραπάνω και τα αποθήκευσα σε ένα νέο αρχείο με τίτλο, «output.txt».*

*Όλα τα ερωτήματα έγιναν με την μέθοδο sub() με συνάρτηση "callback", η οποία αντικαθιστά το περιεχόμενο το οποίο ταίριαξε με ένα χαρακτήρα space.*
*Έγινε επίσης η χρήση της re.DOTALL στα 1,2,3,4,5, η οποία κάνει τον ειδικό χαρακτήρα '.' να ταιριάζει με οποιονδήποτε χαρακτήρα, συμπεριλαμβανομένης μιας νέας γραμμής.*

***Οι πηγές*** *που χρησιμοποιήθηκαν,*

*Για την μέθοδο sub*, https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s050-sub-callback-py

Για την μετατροπή ακολουθιών συνεχόμενων χαρακτήρων whitespace σε ένα ακριβώς κενό

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

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

Για το *re.DOTALL*, http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#id8

Για την απαλοιφή των <script> και <style> tags, http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#sub