# Μεταγλωττιστές 2020

# Προγραμματιστική Εργασία #2


## Ονοματεπώνυμο: Κοτρώτσιος Χρήστος

## Α.Μ.: Π2017109


**1. Συνοπτική περιγραφή της σειράς βημάτων επεξεργασίας στον κώδικα σας**

Αρχικά, πριν ανοίξω το αρχείο **testpage.txt** έκανα import το **re module**. Έπειτα κατασκεύασα μια συνάρτηση **func1** η οποία θα καλείται μέσω του **sub** στο άνοιγμα του αρχείου testpage.txt και θα αντικαθιστά τα **> & <  ** και θα επιστρέφει κάποιο string. Στη συνέχεια, δημιούργησα τις μηχανές ταιριάσματος καθώς και τις κανονικές εκφράσεις. Μετά απο αυτό, άνοιξα το αρχείο **testpage.txt** και χρησιμοποίησα τις μηχανές ταιριάσματος ώστε να μπορέσω να επεξεργαστώ το κείμενο. Τέλος εκτύπωσα τα αποτελέσματα σε Linux Terminal και τα έκανα copy.

**2. Περιγραφή της κανονικής έκφρασης που χρησιμοποιήσατε σε κάθε βήμα**

**1) ('<title>(.+?)</title>')**

Η χρηση της έκφρασης αυτής έχει ως ρόλο την επιλογή των χαρακτήρων που βρίσκονται μέσα στο html title tag. Χρησιμοποιώ τον τελεστή **.** για οποιονδήποτε χαρακτήρα και τον τελεστή επανάληψης **+** για οποιονδήποτε χαρακτήρα εμφανίζεται 1 ή περισσότερες φορές

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

Η χρηση της έκφρασης αυτής έχει ως ρόλο την επιλογή σχόλιων. Επιπλέον αφού μου ζητείται να κάνω απαλοιφή ολόκληρων των σχόλιων , δεν χρησιμοποιώ παρενθέσεις. Επίσης χρησιμοποιώ re.DOTALL καθώς τα σχόλια μπορεί να είναι περισσότερες απο μια σειρές.

**3) (r'<(script|style).*?>.*?</(script|style)>',re.DOTALL)**

Ρόλος αυτής της έκφρασης είναι η επιλογή και απαλοιφή των script και style. Χρησιμοποιούμε τον τελεστή | για να επιλέγουμε και τα 2. Ομοίως αφού ζητείται απαλοιφή, δεν χρησιμοποιούμε παρενθέσεις.

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

Με την έκφραση αυτή επιλέγουμε τα περιεχόμενα του href καθώς και τα περιεχόμενα ανάμεσα στα <a> </a> tags . Ταιριάζουμ αρχικά ότι βρίσκεται μεταξύ του <a και του href, στην συνέχεια ότι βρίσκεται μέσα στα "  " που αποτελέι τον σύνδεσμο, μετά ταιριάζουμε οτι βρίσκεται μεταξύ " και > και τέλος ότι βρίσκεται ανάμεσα στα <a> </a>.

**5) r'<.+?>|</.+?>',re.DOTALL  ,  r'<.+?/>',re.DOTALL**

Χρησιμοποιούμε 2 εκφράσεις καθώς, πέρα απο τα tags που ξεκινάνε ως < > και κλείνουν ως < /> υπάρχουν και τα self closing tags </>. Επομένως χρησιμοποιούμε 2 εκφράσεις για να είμαστε καλυμμένοι.

**6) (r'&(amp|gt|lt|nbsp);')**

Ρόλος αυτής της έκφρασης είναι το ταίριασμα με τα **> & <  ** entities.

**7) (r'\s+')**

Ρόλος αυτής της έκφρασης είναι το ταίριασμα με τα whitespaces , όπου \s αντιπροσωπεύει τον χαρακτήρα whitespace

**3. Αναφορά σε πηγές που πιθανόν χρησιμοποιήσατε**

Σημειώσεις του μαθήματος