# Μεταγλωττιστες 2020

**Προγραμματιστική εργασία #2**

**Γεώργιος Ντότσιος**
**Π2014148**

Αρχικά γίνεται η εισαγωγή της βιβλιοθήκης re απαραίτητη για την υλοποίηση της εργασίας με κανονικές εκφράσεις.Στην πορεία όπως γράφω και στα σχόλια δημιουργώ μια συνάρτηση με όνομα cb, η οποία πραγματοποιεί αντικατάσταση των χαρακτήρων &, >, < και nbsp;. Αφού τελείωσα το παραπάνω βήμα άρχισα να φτιάχνω μια-μια με την σειρά της εκφώνησης, τις κανονικές εκφράσεις, οι οποίες και ελέγχθηκαν μέσα στην with σε ένα δοκιμαστικό κείμενο.

Ερώτημα 1: (r'(<title>)(.+?)</title>'): Εντοπίζει οτι ξεκινάει με <title> και τελειώνει με </title>. Με αυτόν τον τρόπο γίνεται η εξαγωγή του τίτλου αφου απομονώνει το ενδιάμεσο κομμάτι.

Ερώτημα 2: `('<!--.*?-->',re.DOTALL)` Αποτελεί μια κανονική έκφραση με την οποία πραγματοποιείται η απαλοιφή των σχολίων που βρίσκονται μεταξύ <!-- και -->. Με την εντολή re.DOTALL επιτρέπουμε το πρόγραμμα να διαβάζει σχόλια που είναι πάνω απο 1 σειρά.

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

Με αυτήν την κανονική έκφραση πραγματοποιούμε απαλοιφή των script και style tags καθώς και όλο το ενδιάμεσο περιεχόμενο.Ξαναχρησιμοποιήθηκε και η DOTALL εντολή.

Ερώτημα 4 `(r'<a.+?href="(.*?)".*?>(.*?)</a>',re.DOTALL)` :Σε αυτό το βήμα πραγματοποιείται εξαγωγή και εκτύπωσε συνδέσμων href που βρίσκονται μετξαξύ <a> και </α>

Ερώτημα 5 `(r'<.+?>|</.+?>',re.DOTALL)` και `(r'<.+?/>',re.DOTALL)`: Στην πρώτη περίπτωση επιτυγχάνεται απαλοιφή των tags που βρίσκονται μεταξύ <> και </> ενω στη δεύτερη σε αυτά της μορφής </>.

Ερώτημα 6 `(r'&(amp|gt|lt|nbsp);')` : Κάνει μετατροπη των html entities.

Ερώτημα 7 `(r'\s+')` : Κάνει απαλοιφή των whitespace μια ή και περισσότερες φορές.