**Μεταγλωττιστές 2020**

**Προγραμματιστική Εργασία #2**

**Ονοματεπώνυμο:** Πασχάλης Γρίβας

**ΑΜ:** Π2017082

## Περιγραφή Βημάτων Επεξεργασίας και Κανονικών Εκφράσεων

Αρχικά στον κώδικα, ανοίγεται το αρχείο εισόδου για ανάγνωση και στην συνέχεια σε μία μεταβλητή εκχωρείται όλο το περιεχόμενό του. Στην συνέχεια, γίνονται τα εξής βήματα:

1. Χρησιμοποιείται η κανονική έκφραση `'<title>(.+)</title>'` ώστε να αναγνωρίζεται ο τίτλος και να τυπώνεται το περιεχόμενό του τυπώνοντας το m.group(1).
2. Χρησιμοποιήθηκαν οι κανονικές εκφράσεις `'<!--.+?-->'` και `'<(script|style).*?>.*?</\1>'` προκειμένου να αναγνωρίζονται τα html σχόλια και τα `<script>` και `<style>` tags αντίστοιχα. Στη δεύτερη κανονική έκφραση γίνεται χρήση backreference έτσι ώστε στο σημείο \1 να ταιριάζει ότι ταιριάζει στο group(1). Οι εκφράσεις αυτές χρησιμοποιούνται από τη μέθοδο sub() η οποία αντικαθοστά το περιεχόμενο το οποίο ταίριαξε με ένα χαρακτήρα space.
3. Παρόμοια με το βήμα 1 γίνεται χρήση της κανονικής έκφρασης `'<a.*?href="(.+?)".*?>(.*?)</a>'` έτσι ώστε να αναγνωρίζονται και να τυπώνονται οι σύνδεσμοι και το κείμενο των `<a>` tags.
4. Χρησιμοποιείται η έκφραση `'<.+?>'` προκειμένου να αναγνωρίζονται όλα τα html tags του κειμένου και με τη μέθοδο sub() αντικαθίστανται με ένα χαρακτήρα space.
5. Γίνεται χρήση της έκφρασης `'&|>|<| '` έτσι ώστε να αναγνωρίζονται ειδικά html entities και στη συνέχεια με τη βοήθεια της συνάρτησης cb, η οποία καλείται από τη μέθοδο sub(), μετατρέπονται σε χαρακτήρες &,>,<, κενό (space) αντίστοιχα.
6. Με την μέθοδο sub() και την κανονική έκφραση `'\s+'` αναγνωρίζονται οι συνεχόμενοι χαρακτήρες space και μετατρέπονται σε έναν.

Τέλος, τυπώνεται το κείμενο όπως αυτό έχει διαμορφωθεί μετά από τις μετατροπές των βημάτων 2, 4, 5 και 6.

Πρέπει να σημειωθεί ότι στις κανονικές εκφράσεις των βημάτων 1 έως 4 ήταν απαραίτητο να χρησιμοποιηθεί το flag re.DOTALL προκειμένου στο σύμβολο . (τελεία) να ταιριάζει έναν οποιονδήποτε χαρακτήρα **και** το newline.