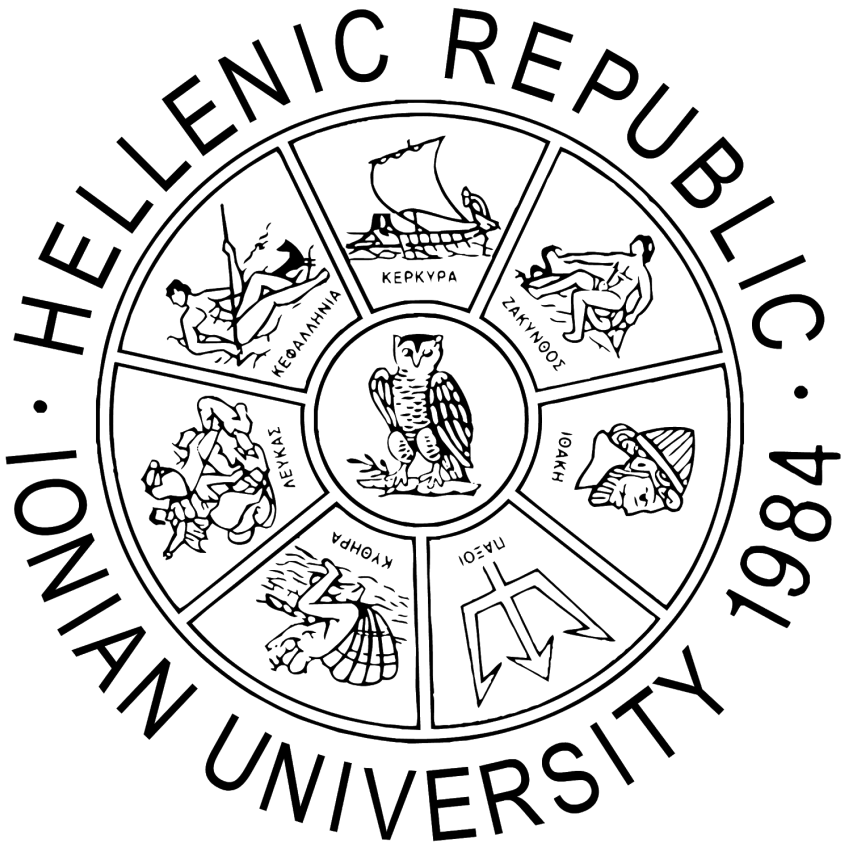

Μεταγλωττιστές 2020

Προγραμματιστική Εργασία #2

Nemanja Jevtic
Π2017182

Βήματα επεξεργασίας και οι κανονικές εκφράσεις

1. Εξαγωγή και εκτύπωση του τίτλου.
   Με την χρήση της κανονικής έκφρασης ‘<title>.+?</title>’ γινεται η αναζήτηση στο αρχείο για τον τίτλο. Στην συνέχεια χρησιμοποιείται η βοηθητική κανονική έκφραση '<.+?>' στην συνάρτηση sub() για να γίνει η απαλοιφή των tags πριν την εκτύπωση του τίτλου.
2. Απαλοιφή των σχολίων.
   Με την χρήση της κανονικής έκφρασης ‘<!--.+?-->’ και πάλι της συνάρτησης sub αντικαθίστανται τα σχόλια με κενό χαρακτήρα.
3. Απαλοιφή των script/style tags.
   Με την χρήση της έκφρασης ‘<.+?>.+?</(.+?)>’, της sub() και της callback συνάρτησης remove_tags(). Η έκφραση βρίσκει πρακτικά ολα τα ζεύγη tags, και στην συνέχεια η sub καλώντας την remove_tags() ελέγχει εάν το match είναι ένα script ή ένα style tag και τα αντικαθιστά με έναν χαρακτήρα κενού.
4. Εξαγωγή και εκτύπωση των συνδέσμων, μαζί με το συνοδευτικό κείμενο.
   Η έκφραση ‘<a.+?>.+?’ αναγνωρίζει όλα τα anchor tags. Με την χρήση της findall μπορούμε να τα βρούμε. Ύστερα χρησιμοποιώντας την έκφραση ’href="(.+?)"’ και την συνάρτηση search() μπορουμε να βρούμε τον σύνδεσμο που περιέχεται στο tag. Με μία matchObject.group(1) είμαστε έτοιμοι να εκτυπώσουμε τον σύνδεσμο. Με την χρήση της ίδη γνωστής έκφρασης (<.+?>) απαλοίφουμε και το tag με αποτέλεσμα να έχουμε μείνει μόνο με το συνοδευτικό κείμενο (στην περίπτωση που υπάρχει).
5. Απαλοιφή όλων των tags από το κείμενο
   Εδώ χρησιμοποιούμε για τρίτη και τελευταία φορά το ‘<.+?>’ για να αντικαταστήσουμε όλα τα tags με από έναν χαρακτήρα κενού.
6. Μετάφραση των ειδικών χαρακτήρων HTML.
   Παρόμοια υλοποίηση με το (3). Η ‘&.+?;’ αναγνωρίζει όλους τους ειδικούς χαρακτήρες, και η callback συνάρτηση replace_special() ελέγχει για ποιά περίπτωση πρόκειται και τα αντικαθιστά με σωστό χαρακτήρα μέσω της sub. Σημειώνεται ότι επειδή η εκφώνηση ζήταγε τα βήματα επεξεργασίας να

γίνουν το ένα μετά το άλλο, χωρίς να διευκρινίζεται εάν είναι επιτρεπτή ή όχι η αλλαγή της σειράς των βημάτων της εκφώνησης, τα άφησα ως έχουν. Αυτό είχε ως αποτέλεσμα στο βήμα (4) μερικά από τα συνοδευτικά κείμενα να εκτυπωθούν με χαρακτήρες html. Αν γίνει η πολύ απλή αλλαγή σειράς εκτέλεσης των βημάτων και το βήμα (6) εκτελεσθεί πρίν το βήμα (4) το πρόβλημα λύνεται.

Πηγές :

The Python Standard Library Documentation : re -- Regular expression operations
https://docs.python.org/3/library/re.html