Μεταγλωτιστές 2020
Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Ορφέας Γεωργίου
ΑΜ: Π2015033

Έκανα import το re και άνοιξα το αρχείο. Χρησιμοποίησα encoding utf 8 για να μη βγάζει κάποιο σφάλμα. Έπειτα ξεκίνησα να υλοποιώ με τη σειρά που μας δώθηκαν τα ερωτήματα και να ελέγχω για κάθε ένα απο αυτά τα αποτελέσματα στο τερματικό. Κατα τη διάρκια του ερωτήματος 6 δημιουργήθηκε και η απαραίτητη συνάρτηση. Τέλος, εκτύπωσα το τελικό αρχείο.

1.
re.compile(r'<title>(.+?)</title>')
Ένας ή περισσότεροι χαρακτήρες που βρίσκονται μεσα στα tags.
Το (?) μπαίνει ώστε να μην συνεχίζει να ελέγχει χαρακτήρες μέχρι τέλους.
2.
re.compile(r'<!---.*?--->',re.DOTALL)
Επιλογή των συμβόλων των comments και των περιεχομένων τους.
Το (.*) γιατί μπορεί να μην υπάρχει τίποτα.
3.
re.compile(r'<script.*?>.*?</script>',re.DOTALL)
re.compile(r'<style.*?>.*?</style>',re.DOTALL)
Επέλεξα να τα κάνω χωριστά για σιγουριά. Σε περίπτωση που υπήρχε για κάποιο περίεργο λόγο script και μετά style ίσως να θέλαμε να το κρατήσουμε.
Επιλογή του πρώτου tag με ότι περιέχει (.*?), επιλογή ότιδήποτε βρίσκεται ανάμεσα στα tags(.*?) και επιλογή του 2ου tag
4.
re.compile(r'<a.*?href="(.*?)".*?>(.*?)',re.DOTALL)
Έχουν μπει οι 2 παρενθέσεις ώστε να να γίνει μετά η εμφάνιση τόσο των περιεχομένων της href οσο και του ότι βρίσκεται μέσα στα tags με τη μέθοδο των groups.
5.
re.compile(r'</.+?>',re.DOTALL)
re.compile(r'<.+?/>',re.DOTALL)
re.compile(r'<.+?>',re.DOTALL)
Για κάθε περίπτωση tag που θέλουμε να αφαιρεθεί μια κανονική έκφραση.
(.+?) μια η περισσότερες φορές αυτό που θα υπάρχει μεσα
6.
re.compile(r'&(amp|gt|lt|nbsp);',re.DOTALL)
Ταίριασμα όπου υπάρχει &και μετά μια απο αυτές τις λέξεις
7.
re.compile(r'\s+')
Ταίριασμα τα κενά μια ή περισσότερες φορές