Μεταγλωττιστές 2020

Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Πλιάφας Θεόδωρος

ΑΜ: Π2017012

- **<u>Συνοπτική περιγραφή της σειράς βημάτων επεξεργασίας στον κώδικα σας.</u>**

  Αρχικά, έγινε import το module re της python. Έπειτα, δημιουργήθηκε η συνάρτηση cb η οποία θα καλείται με τη μέθοδο sub για να αντικαθιστά τους χαρακτήρες & > < . Στη συνέχεια, δημιουργήθηκαν οι κανονικές εκφράσεις και τέλος διαβάστηκε το αρχείο και με τη χρήση των μηχανισμών ταιριάσματος έγιναν οι απαραίτητες διορθώσεις στο κείμενο και εκτυπώθηκε.

- **<u>Περιγραφή της κανονικής έκφρασης που χρησιμοποιήσατε σε κάθε βήμα.</u>**

  1. Δημιουργήθηκε η κανονική έκφραση (‘<title>(.+?)</title>’) και επιλέχτηκε η . και το + ώστε να ταιριάξουν οτιδήποτε βρίσκεται μέσα στο <title></title>.
  2. Δημιουργήθηκε η κανονική έκφραση (‘<!--.*?-->,re.DOTALL) και επιλέχτηκε η . και το * γιατί υπάρχει πιθανότητα κάποιο σχόλιο να μην περιέχει κάτι. Δεν υπάρχουν παρενθέσεις γιατί θέλουμε απαλοιφή.
  3. Δημιουργήθηκε η κανονική έκφραση (r’<s(?:cript|tyle)).*?>.*?</\1>’,re.DOTALL) και επιλέχτηκε το | για να επιλέγει το script ή το style κάθε φορά, το * και η . για να ταιριαστεί οτιδήποτε υπάρχει ανάμεσα στα <script></script> και <style></style και μετά από αυτά.

4. Δημιουργήθηκε η κανονική έκφραση (r’<a.+?href=”(.*?)”.*?>(.*?)’,re.DOTALL) και επιλέχτηκε η ., το ? και το * για το ταίριασμα οτιδήποτε περιέχει το href.
5. Δημιουργήθηκαν οι κανονικές εκφράσεις (r'<.+?>|</.+?>',re.DOTALL) και (r'<.+?/>',re.DOTALL) για τα δύο διαφορετικά είδη tags που υπάρχουν.
6. Δημιουργήθηκε η κανονική έκφραση (r'&(amp|gt|lt|nbsp);') για το ταίριασμα των & > < .
7. Δημιουργήθηκε η κανονική έκφραση (r'\s+') για την εξαγωγή των whitespaces.