Μεταγλωτιστες 2020
Προγραμματιστική Εργασία #2
ονοματεπώνυμο : Μουστάκας Ιωάννης
ΑΜ:Π2016174

Βήμα 0

Έκανα εισαγωγή την re βιβλιοθήκη, άνοιξα τα αρχεία μου και έκανα μια function που να παίρνει ως όρισμα μια λίστα και να μετατρέπει το κάθε όρισμα σε string και να το γράφει σε newline

Βήμα 1ο

Χρησιμοποιησα την έκφραση <title>(.+?)</title> όπου αναγνωρίζει οτι βρίσκετε μεταξύ του tag title(<title></title>) και έκανα εύρεση και εκτύπωση του τίτλου

Βήμα 2ο

Χρησιμοποιησα την έκφραση <!--(.*?)--> που αναγνωρίζει οτι βρίσκετε μεταξύ των σχολιων(<!-- -->) και εκανα ευρεση και αντικατάσταση με κενό χαρακτήρα(space)με την βοήθεια της sub για να γίνει η απαλοιφή.Στην έκφραση πρόσθεσα και την re.dotall για να διαβάζει περισσότερο από μια γραμμή

Βήμα 3ο

Χρησιμοποιησα την έκφραση <script(.*?)</script>|style="(.*?)"που αναγνωρίζει οτι βρίσκετε μεταξύ <script </script> και μεταξύ style=" " και έκανα εύρεση και αντικατάσταση με κενό χαρακτήρα(space)με την βοήθεια της sub για να γίνει η απαλοιφή.
Στην έκφραση πρόσθεσα και την re.dotall για να διαβάζει περισσότερο από μια γραμμή

Βήμα 4ο

Χρησιμοποιησα την έκφραση <a[^>]* href="https://([^"]*)">([^<]*)' που αναγνωρίζει τα href="https:// που βρίσκετε μεταξύ από <a και > και σχόλια(τίτλος) ανάμεσα σε<a>και και έκανα εύρεση και εκτύπωσα το link και το τίτλο του και έγραψα τα αποτελέσματα στο αρχείο αποτελεσμάτων

Βήμα 5ο
Δημιούργησα μια function που παίρνει ως όρισμα το αρχείο και κάνει απαλοιφή των tags και χρησιμοποίησα την έκφραση r'<.*?>' που αναγνωρίζει οτι βρίσκετε μεταξύ του < και > και έκανα εύρεση και αντικατάσταση με κενό χαρακτήρα(space)με την βοήθεια της sub για να γίνει η απαλοιφή.
Στην έκφραση πρόσθεσα και την re.dotall για να διαβάζει περισσότερο από μια γραμμή

Βήμα 6ο

Δημιούργησα μια function που ομαδοποίησα τις οντότητες που θέλω να αντικαταστησω με άλλες τιμές-λέξεις και χρησιμοποίησα την έκφραση r'&(.*?);' που αναγνωρίζει οτι βρίσκετε μεταξύ του & και ; και έκανα εύρεση στο κείμενο και με την βοήθεια της sub έκανα αντικατασταση των αποτελματων με την faction που έκανα προηγουμένως

Βήμα 7ο

Χρησιμοποιησα την έκφραση r'\s+' αναγνωρίζει τα κενά σύμβολα (1 η περισσότερα) και με την βοήθεια της sub τα έκανα αντικατάσταση με 1 μόνο κενό

Βήμα 8ο

Έκανα εκτύπωση το διορθωμένο κείμενο