Μεταγλωτιστές 2020

Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Χριστόδουλος Καλογιάννης

ΑΜ: Π2017077

Αρχικά, κάναμε import το module **re** της python, αξιοποιώντας κατά αυτό τον τρόπο πλήθος έτοιμων εργαλείων όπως για παράδειγμα την δημιουργία μηχανών ταιριάσματος για την αναζήτηση ταιριασμάτων βάση ποικίλων **regular expressions**. Στην συνέχεια, πριν το άνοιγμα του αρχείου **testpage.txt** για την αναζήτηση ταιριασμάτων, δημιουργήσαμε τις μηχανές ταιριάσματος και αντίστοιχα τις κανονικές εκφράσεις βάση των οποίων θα γίνει η αναζήτηση αργότερα. Εκτός όμως από την δημιουργία των παραπάνω, υπάρχει και η υλοποίηση μιας συνάρτησης με όνομα **myfunction**, στόχος της οποίας είναι η επιστροφή ενός string ανά περίπτωση. Η συνάρτηση αυτή, θα καλείται μέσω της μεθόδου **sub** αργότερα, όταν θα εκτελείται κώδικας για την αντικατάσταση των χαρακτήρων **& > < ** του **testpage.txt.** Μετά την υλοποίηση των κανονικών εκφράσεων και της συνάρτησης **myfunction**, ανοίξαμε το αρχείο **testpage.txt** αποθηκεύοντας όλο του το περιεχόμενο μέσα στην μεταβλητή με όνομα **text**. Έπειτα, κάναμε χρήση συναρτήσεων των μηχανών ταιριάσματος έτσι ώστε με πρότυπο τις κανονικές εκφράσεις που δημιουργήσαμε προηγουμένως να γίνει η επιθυμητή επεξεργασία του κειμένου. Παραδείγματα συναρτήσεων που χρησιμοποιήθηκαν αποτελούν οι **finditer()**, **sub()**, **search()**. Τέλος, εκτυπώνουμε την μεταβλητή **text**, παίρνοντας ως output το κείμενο που προέκυψε μετά την επιθυμητή επεξεργασία.

Παρακάτω, περιγράφονται οι κανονικές εκφράσεις που σχηματίσαμε στα πλαίσια διεκπεραίωσης της εργασίας.

Ερώτημα 1

Για το ερώτημα 1, η κανονική έκφραση που σχηματίσαμε είναι η: **('<title>(.+?)</title>')**.

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα με τον **html** τίτλο του κειμένου. Η χρήση της **.** και της lazy μορφής του τελεστή επανάληψης **+** χρησιμοποιούνται για το ταίριασμα οποιοδήποτε κειμένου βρίσκεται μεταξύ **<title>** και **</title>**. Παρατηρούμε την χρήση της παρένθεσης, στόχος της οποίας είναι η εξαγωγή μόνο του τίτλου(χωρίς τα **<title></title>**) έτσι ώστε να μπορέσουμε να τον χειριστούμε με χρήση του αριθμό του **group** στο οποίο ανήκει (στην περίπτωση μας **group(1)**).

Ερώτημα 2

Για το ερώτημα 2, η κανονική έκφραση που σχηματίσαμε είναι η: **('<!--.*?-->',re.DOTALL)**.

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα με οποιοδήποτε σχόλιο υπάρχει μέσα στο κείμενο το οποίο βρίσκεται μεταξύ **<!--** και **-->**. Παρατηρούμε την χρήση της **.** και της lazy μορφής του τελεστή επανάληψης ***** στόχος των οποίων αποτελεί το ταίριασμα του εκάστοτε κειμένου μεταξύ των **<!-**

- και **-->.** Θα μπορούσε να χρησιμοποιηθεί η lazy μορφή του τελεστή **+** αντί για τον ***** με την διαφορά ότι θα ήταν απαραίτητο να περιέχεται έστω και ένας χαρακτήρας μεταξύ των **<!--** και **-->.** Τέλος, να σημειωθεί ότι δεν γίνεται χρήση παρενθέσεων αφού η φύση του ερωτήματος απαιτεί την απαλοιφή ολόκληρου του σχολίου (μαζί με τα **<!--** και **-->**), κάτι το οποίο μπορεί να γίνει αξιοποιώντας το **group(0)** που έχει ως περιεχόμενο του ολόκληρο το ταίριασμα.

Ερώτημα 3

Για το ερώτημα 3, η κανονική έκφραση που σχηματίσαμε είναι η:
(r'<(s(?:cript|tyle)).*?>.*?</\1>',re.DOTALL).

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα των tags **<script></script>** και **<style></style>** με όλο τους το περιεχόμενο. Παρατηρούμε, ότι το αρχικό γράμμα **s** είναι κοινό και για τα δύο tags. Για τον λόγο αυτό, χρησιμοποιούμε το σύμβολο **|** (εναλλαγή), στόχος του οποίου είναι η επιλογή κάθε φορά της αντίστοιχης ακολουθίας γραμμάτων μετά το **s** (**cript** ή **tyle**), κάνοντας έτσι το ταίριασμα με το όνομα του tag. Ο συνδυασμός των χαρακτήρων **?:** χρησιμοποιείται για να μην οριστεί ένα νέο group που να έχει ως περιεχόμενο την ακολουθία **cript** ή **tyle**. Βλέπουμε όμως, την ύπαρξη μιας παρένθεσης που ξεκινάει πριν το **s**. Στόχος της συγκεκριμένης, είναι η αποθήκευση του εκάστοτε ονόματος του tag (μέσα στο **group(1)**). Αυτό, συμβαίνει με στόχο την αξιοποίηση backreference όταν θέλουμε να ταιριάξουμε το ίδιο όνομα tag, αλλά αυτή την φορά μετά τον χαρακτήρα / που δηλώνει ότι κλείνει η ετικέτα. Έτσι λοιπόν, βλέπουμε την χρήση του **\1** το οποίο θα χρησιμοποιηθεί για να ταιριάξει ότι έχει ήδη βρεθεί νωρίτερα και είναι αποθηκευμένο στο **group(1)**. Τέλος, γίνεται χρήση της **.** και της lazy μορφής του τελεστή επανάληψης ***** στόχος της οποίας είναι το ταίριασμα οποιουδήποτε χαρακτήρα μετά το **<script** ή **<style** μέχρι τον χαρακτήρα **>** και αντίστοιχα ότι περιλαμβάνεται μεταξύ των **<script></script>** ή **<style></style>**.

Ερώτημα 4

Για το ερώτημα 4, η κανονική έκφραση που σχηματίσαμε είναι η:
(r'<a.+?href="(.*?)".*?>(.*?)',re.DOTALL)

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα του **συνδέσμου** της ιδιότητας **href** (από **<a>** tags) καθώς επίσης και του περιεχόμενου που βρίσκεται μεταξύ **<a>** και ****. Βλέπουμε την χρήση της **.** και της lazy μορφής των τελεστών επανάληψης **?** και ***** με στόχο το ταίριασμα του κειμένου αρχικά από το **<a** μέχρι την ιδιότητα **href**, στην συνέχεια σε ότι βρίσκεται εντός των **“”** (αποτελεί τον σύνδεσμο), έπειτα μέχρι το **>** (ολοκληρώνεται το **<a>** tag) και τέλος μεταξύ των **<a>** και ****. Παρατηρούμε ότι υπάρχουν παρενθέσεις εντός των **“”** για την αποθήκευση του συνδέσμου στο **group(1)** καθώς και εντός των **<a>** και **** αποθηκεύοντας το αντίστοιχο επιθυμητό περιεχόμενο μέσα στο **group(2)**.

Ερώτημα 5

Για το ερώτημα 5, σχηματίσαμε δύο κανονικές εκφράσεις:

1. **(r'<.+?>|</.+?>',re.DOTALL)**
2. **(r'<.+?/>',re.DOTALL)**

Ρόλος των συγκεκριμένων εκφράσεων, αποτελεί το ταίριασμα με κάθε **html** tag του κειμένου. Όπως γνωρίζουμε από την **html**, υπάρχουν δύο κατηγορίες tag. Στην πρώτη, περιέχονται τα tag που ξεκινάνε π.χ με **<a>** και ολοκληρώνονται με ****. Στην δεύτερη, περιλαμβάνονται τα γνωστά **self-closing** tags όπως για παράδειγμα το **<meta />** (παρατηρούμε ότι ολοκληρώνεται από το πρώτο set **<>** χωρίς την ύπαρξη ενός **</meta>**). Η 1. κανονική έκφραση, αφορά την εύρεση tag της πρώτης κατηγορίας. Βλέπουμε ότι χρησιμοποιείται το σύμβολο | (εναλλαγή), έτσι ώστε κάθε φορά να ταιριάζει ή το tag με μορφή π.χ <a>, ή το tag με μορφή π.χ . Η 2. κανονική έκφραση, αφορά την εύρεση tag της δεύτερης κατηγορίας όπως για παράδειγμα **<meta charset="utf-8" />**. Βλέπουμε ότι και στις δύο κανονικές εκφράσεις, γίνεται χρήση της **.** και της lazy μορφής του τελεστή **+** με στόχο το ταίριασμα οποιουδήποτε κειμένου εντός του εκάστοτε tag.

Ερώτημα 6

Για το ερώτημα 6, η κανονική έκφραση που σχηματίσαμε είναι η: **(r'&(amp|gt|lt|nbsp);')**

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα όλων των **& > < ** του κειμένου. Παρατηρούμε την χρήση του συμβόλου **|** (εναλλαγή), στόχος του οποίου είναι το ταίριασμα κάθε φορά μιας από τις τέσσερεις πιθανές επιλογές.

Ερώτημα 7

Για το ερώτημα 7, η κανονική έκφραση που σχηματίσαμε είναι η: **(r'\s+')**

Ρόλος της συγκεκριμένης έκφρασης, αποτελεί το ταίριασμα ακολουθιών συνεχόμενων χαρακτήρων **whitespace**. Παρατηρούμε, την χρήση του συνδυασμού χαρακτήρων **\s** ο οποίος αντιπροσωπεύει τον χαρακτήρα **whitespace**. Επίσης, παρατηρούμε την χρήση του τελεστή επανάληψης **+** στόχος της οποίας αποτελεί το ταίριασμα ενός ή περισσοτέρων **whitespace**.