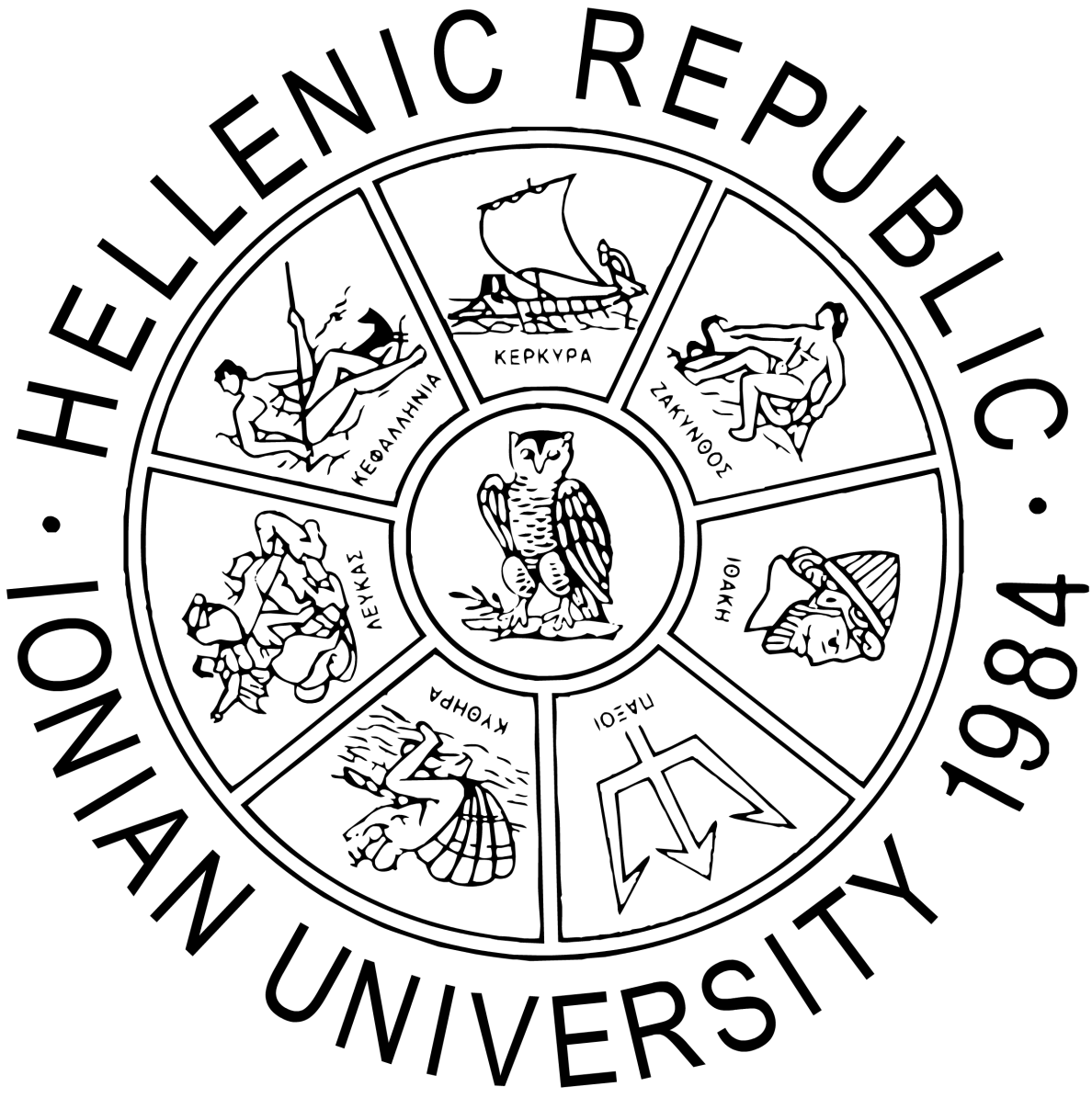
Μεταγλωτιστές 2020

Προγραμματιστική εργασία 2

Ονοματεπώνυμο :Διακουμάκος Βασίλειος

ΑΜ:Π2015088



1.Περίληψη εργασίας

Αρχικά υλοποιήθηκε η συνάρτηση μετατροπής των
html entities (htmlEntities) η οποία καλείτε μέσω τηςμεθόδου sub.Στόχος της συγκεκριμένης συνάρτησης
είναι επιστροφή ενός string για κάθε περίπτωση με βάση των πίνακα που μας δόθηκε.Στην συνέχεια με την βοήθεια της βιβλιοθήκης-module re Δημιούργησα τις κανονικές εκφράσεις για κάθε ερώτημα δημιουργώντας μια μηχανή ταιριάσματος για το καθένα .Τέλος έκανα read το testpage.txt και με
τον συνδυασμό της μεθόδου sub αλά και των κανονικών εκφράσεων που δημιουργήθηκαν παραπάνω έγινε η κατάλληλη επεξεργασία.

2. Ερωτήματα - Κανονικές εκφράσεις

Ζήτούμενο 1.

(r’<title>(.+?)</title>’)

Χρησιμοποείται για να κάνει match το οποιοδήποτε περιεχόμενο ανάμεσα στα html tags <title> </title>.Με την χρήση των παρενθέσεων () γίνεται η εξαγωγή μόνο του κειμένου ,δηλαδή οτι υπάρχει μεσα στα tags (<title> </title>)

Ζητούμενο 2.

(r'<!--.*?-->',re.DOTALL)

Χρησιμοποείται για να κάνει match όλα τα σχόλια.Πιο συγκεκριμένα ανάμεσα<!-- - -> . Η αφαίρεση τους έγινε με την χρήση της sub.

Ζητούμενο 3.

(r’<(script | style).*?>.*?</(script | style)>',re.DOTALL)

Χρησιμοποείται για να κάνει match όλα τα περιεχόμενα στα tags <style></style> kai <script> </script>.Η αφαίρεση τους έγινε με την χρήση της sub.

Ζητούμενο 4.

(r'<a.+?href="(.*?)".*?>(.*?)',re.DOTALL)

Χρησιμοποείται για να κάνει match όλα τα περιεχόμενα <href> tags (συνδέσμων των κειμένων)και <a> τίτλων των συνδέσμων.

```
 for m in rexp4.finditer(text):
    print('{} {}'.format(m.group(1),m.group(2)))
```

Κάνουμε print τα links

Ζητούμενο 5.1 & 5.2

1.(r'<.+?>|</.+?>',re.DOTALL) Χρησιμοποείται για να κάνει match όλα τα περιεχόμενα στα tags της μορφής <> </>
2.(r'<.+?/>',re.DOTALL)Χρησιμοποείται για να κάνει match όλα τα περιεχόμενα στα tags της μορφής </> η αλλιώς self closing tags

```
text = rexp51.sub(' ',text)
  text = rexp52.sub(' ',text)
```

Η Απαλοιφή ολων των διπλών tags και των self closing εγινε με την χρήση της sub

Ζητούμενο 6.

```
(r'&(amp|gt|lt|nbsp);')
```

Χρησιμοποιήθηκε για να κάνει match ολα τα html entities (& > <)

text = rexp6.sub(htmlEntities,text)

Μετατροπη Html entities με την χρήση της sub

Ζητούμενο 7.

(r'\s+')
Matching whitespace χαρακτήρων μέσα στο κείμενο.
text = rexp7.sub(' ',text)
Μετατροπή whitespaces σε ενα κενό με την χρήση της sub

3.Πηγές

Διαφάνειες μαθήματος.