Μεταγλωτιστές 2020
Προγραμματιστική Εργασία #2

Ονοματεπώνυμο: Φοίβος Αργυρίδης
ΑΜ: Π2013041

Συνοπτική περιγραφή της σειράς βημάτων επεξεργασίας στον κώδικά σας

Έγινε import η βιβλιοθήκη re και στη συνέχεια δομήθηκαν τα expressions για να εφαρμοστούν στο κείμενο. Η λογική ήταν να επαναχρησιμοποιηθούν κάποια από τα expressions όπως και έγινε με την περίπτωση των urls.

Πριν αρχίσει η εφαρμογή των expressions στο string της σελίδας δημιουργήθηκε callback function που κάνει replace τα special chars όπως αναφέρεται στην εκφώνηση της άσκησης.

Τέλος διαβάστηκε το αρχείο της ιστοσελίδας και έγιναν οι απαραίτητες μετατροπές όπως θα δούμε παρακάτω ώστε να παραχθεί το τελικό αποτέλεσμα.

Περιγραφή της κανονικής έκφρασης που χρησιμοποιήσατε σε κάθε βήμα.

Η πρώτη έκφραση αφαιρεί το title tag από τον τίτλο της σελίδας έτσι ώστε να τυπώσουμε σκέτο τον τίτλο. Αυτό επιτυγχάνεται έτσι και αλλιώς στα παρακάτω βήματα που αφαιρούνται όλα τα tags (μαζί και ο τίτλος)

```
removeTitle = re.compile('<title>(.*?)</title>')
```

Η δεύτερη έκφραση αφαιρεί όλα τα σχόλια από τον κώδικα.

```
removeComments = re.compile('<!--.*?-->',re.DOTALL)
```

Η Τρίτη έκφραση αφαιρεί ότι υπάρχει από script kai style και ανάμεσα σε αυτά στον κώδικα της σελίδας

```
removeCode = re.compile(r'<(script|style).*?>.*?</(scri
pt|style)>',re.DOTALL)
```

Η Τρίτη έκφραση αφαιρεί στην ουσία τα a tags κρατώντας το url μέσα στο href και ότι υπάρχει ανάμεσα στα <a>..

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

Επειδή όμως υπάρχουν και άλλα html tags μέσα στα a tags όπως images <img src=”” alt=””> το αποτέλεσμα της έκφρασης αποθηκεύτηκε σε νέο string όπου εφαρμόστηκαν μερικές από τις παρακάτω εκφράσεις πριν τυπωθεί στο τέλος του κώδικα μετά το τελικό κείμενο. Η ιδέα ήταν όλα τα urls με το κείμενό τους να βρίσκονται στο τέλος του output.

Η επόμενη έκφραση βρίσκει τα special chars στον κώδικα και συνδυάζεται με το replace function που κάνει replace τους χαρακτήρες αυτούς σύμφωνα με τις οδηγίες.

```
removeSpecial = re.compile(r'&(amp|gt|lt|nbsp);')
```

```
def replace(m):
    if (m.group(0)=='&amp;'):
        return '&'
    elif (m.group(0)=='&gt;'):
        return '>'
    elif (m.group(0)=='&lt;'):
        return '<'
    else:
        return ' '
```

Στις τελευταίες εκφράσεις αφαιρούνται τα παραπάνω whitespaces

```
removeWhitespaces = re.compile(r'\s+')
```

και τέλος, αφαιρούνται όλα τα html tags

```
removeAllTags = re.compile(r'<.*?>',re.DOTALL)
```

Εφαρμογή των εκφράσεων για την τελική παραγωγή του αποτελέσματος

Στο σημείο αυτό ανοίγουμε το testpage.txt και αποθηκεύουμε το αποτέλεσμα σε ένα string με την ονομασία text. Εδώ χρησιμοποιούμε την πρώτη έκφραση ώστε να τυπώσουμε τον τίτλο.

```
title = removeTitle.search(text)
print(title.group(1))
```

Ο παραπάνω κώδικας χρησιμοποιεί την έκφραση removeTitle που χρησιμοποιήσαμε πριν.

Εφαρμόζουμε όλες τις εκφράσεις στο string text στο παρακάτω παράδειγμα αφαιρούμε τα comments, scripts και styles καθώς και τα special chars.;

```
text = removeComments.sub(' ',text)
text = removeCode.sub(' ',text)
text = removeSpecial.sub(replace,text)
```

Πριν προχωρήσουμε στην αφαίρεση των tags γράφουμε όλα τα urls σε ένα νέο string

```
urls = ""
for link in removeLinks.finditer(text):
        urls = urls + '{} {}'.format(link.group(1),link
.group(2))
```

Στη συνέχεια επειδή τα urls περιέχουν tags όπως <img ή άλλα εφαρμόζουμε τις εκφράσεις για να αφαιρέσουμε τα tags, καθώς και τα white spaces που μπορεί να περιέχουν.

```
urls = removeAllTags.sub(' ',urls)
 urls = removeWhitespaces.sub(' ',urls)
```

Εφαρμόζουμε τις ίδιες εκφράσεις και στο string του text.

```
text = removeAllTags.sub(' ',text)
text = removeWhitespaces.sub(' ',text)
```

και στη συνέχεια τυπώνουμε τα δύο string

```
print(text)
print(urls)
```

Πηγες

https://stackoverflow.com/questions/4435169/how-do-i-append-one-string-to-another-in-python

https://stackoverflow.com/questions/3398852/using-python-remove-html-tags-formatting-from-a-string

https://stackoverflow.com/questions/28208186/how-to-remove-html-comments-using-regex-in-python

https://stackoverflow.com/questions/9662346/python-code-to-remove-html-tags-from-a-string

https://gist.github.com/mixstef/39d5257c7498dceac1aa6428e33f2003#file-s050-sub-callback-py

http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#sub

http://mixstef.github.io/courses/compilers/lecturedoc/unit2/module1.html#id8

http://mixstef.github.io/courses/compilers/lecturedoc/appendix-python/module1.html#id5