Μεταγλωττιστές 2016
Προγραμματιστική Εργασία #2

Ζητούμενο

Ο στόχος της άσκησης είναι **με πρόγραμμα Python3 και regular expressions** να φιλτράρετε μια ιστοσελίδα, κρατώντας μόνο τους συνδέσμους (URL links, από a tags), καθώς και το κείμενο κάθε συνδέσμου.

Υποθέστε ότι αυτό χρειάζεται σε μια εφαρμογή crawler· στα πλαίσια της άσκησης, όμως, δεν μας ενδιαφέρει πού θα χρησιμοποιηθούν οι σύνδεσμοι!

Διαδικασία

1. Επιλέξτε μια δοκιμαστική ιστοσελίδα με αρκετούς συνδέσμους. Φορτώστε **ολόκληρη την ιστοσελίδα** σε ένα string. Αν φορτώνετε την ιστοσελίδα από το web, μπορείτε να το κάνετε ως εξής:

```
import urllib.request
page = urllib.request.urlopen('http://www.example.com/my/page/url..')
text = page.read().decode('utf-8')
page.close()
```

Φροντίστε να χρησιμοποιήσετε το σωστό encoding, αν η σελίδα που επιλέξατε δεν είναι σε κωδικοποίηση utf-8.

2. Στη συνέχεια, χρησιμοποιήστε regular expression(s) πάνω στο string που περιέχει το κείμενο της σελίδας, για να ανακτήσετε τα ζητούμενα μέρη. Στο τέλος, τυπώστε τα URL links και το αντίστοιχο κείμενο του καθενός.

Υπόδειξη 1: *Στην κλήση της μεθόδου compile προσθέστε ως δεύτερο όρισμα το re.DOTALL, επιτρέποντας* ***το ταίριασμα της τελείας (.) και με το newline*** *(ιδιαίτερα βολικό, εφόσον έχετε όλη την ιστοσελίδα σε ένα string, το οποίο περιέχει ενδιάμεσα newlines):*

```
rexp = re.compile(restr,re.DOTALL)
```

Σχετικό είναι και το re.IGNORECASE, το οποίο επιτρέπει το ταίριασμα ανεξάρτητα κεφαλαίων-πεζών. Αν θέλετε να το συνδυάσετε με το προηγούμενο flag, χρησιμοποιήστε το | (bit-wise OR).

Υπόδειξη 2: *Ένας crawler χρειάζεται τα URLs σε απόλυτη μορφή (http://..) ενώ συχνά οι ιστοσελίδες περιέχουν σχετικά links (με βάση τη διεύθυνση της τρέχουσας ιστοσελίδας). Για να αντιμετωπίσετε την τελευταία περίπτωση χρησιμοποιήστε την urllib.parse.urljoin(), δείτε π.χ. το παράδειγμα που ακολουθεί:*

```
>>> import urllib.parse
>>> a = 'http://example.com/test/test.html'
>>> b = '/another/page.html'
>>> urllib.parse.urljoin(a,b)
'http://example.com/another/page.html'
>>> c = 'http://ex2.com/base/source.html'
>>> urllib.parse.urljoin(a,c)
'http://ex2.com/base/source.html'
```

```
>>> d = 'relative/link.html'
>>> urllib.parse.urljoin(a,d)
'http://example.com/test/relative/link.html'
```

Παραδοτέο

Θα πρέπει παραδώσετε α) το πρόγραμμά σας (αρχείο .py) και β) αναφορά με συνοπτική περιγραφή της εργασία σας (αρχείο pdf). Η αναφορά θα πρέπει να περιλαμβάνει το κείμενο εισόδου (ιστοσελίδα) και τα δεδομένα εξόδου.

1. Αποθηκεύστε τα παραδοτέα σας σε κάποια on-line υπηρεσία (π.χ. Dropbox).
2. Στείλτε με e-mail το link πρόσβασης στα αρχεία σας.
   Προσοχή: συνημμένα αρχεία μέσω e-mail **δεν θα γίνουν αποδεκτά**!

Η εργασία είναι **ατομική**.

Προθεσμία παράδοσης: Τρίτη 5/4/2016 11:00.