Εισαγωγή στη γλώσσα προγραμματισμού

Ενότητα 6. Διαχείριση δεδομένων

Μάθημα 20.

Ανάκτηση δεδομένων από το διαδίκτυο

https://docs.python.org/3/howto/urllib2.html

Νίκος Αβούρης, Πανεπιστήμιο Πατρών. hci.ece.upatras.gr/avouris

[idle]

20. Βιβλιοθήκη urllib.request

```
req_object = urllib.request.Request(διαδ. πόρος)
mypage  = urllib.request.urlopen(req_object)
html = mypage.read().decode()
```

20. Βιβλιοθήκη urllib.request

```
import urllib.request
url = urllib.request.Request('https://www.upatras.gr/el')
with urllib.request.urlopen(url) as response:
    char_set = response.headers.get_content_charset()
    html = response.read().decode(char_set)
with open("www_upatras_gr.html", "w", encoding = char_set) as p:
    p.write(html)
```

Η μεταβλητή html περιέχει τον κώδικα της σελίδας, τον οποίο μπορούμε να επεξεργαστούμε, πχ με regex

[idle]

20. Ανάκτηση ιστοσελίδων: έλεγχος σφάλματος

```
import urllib.request
import urllib.error
url = urllib.request.Request('https://www.upatras3.gr/el')
try:
    with urllib.request.urlopen(url) as response:
        char_set = response.headers.get_content_charset()
        html = response.read().decode(char_set)
    with open("www_upatras_gr.html", "w", encoding = char_set) as p:
        p.write(html)
except urllib.error.HTTPError as e:
    print(e.code)
    print('Σφάλμα HTTP')
except urllib.error.URLError as e:
    if hasattr(e, 'reason'):
        print('Αποτυχία σύνδεσης στον server')
        print('Αιτία: ', e.reason)
else:
    print('τέλος προγράμματος')
```

#https://docs.python.org/3/howto/urllib2.html#urlerror

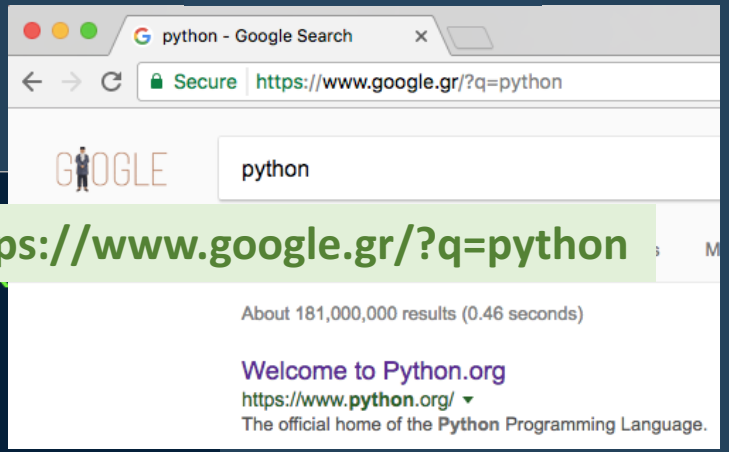
[idle]

20. Ανάκτηση ιστοσελίδων: timeout

```
import urllib.request
import urllib.error
import socket
# timeout : χρόνος αναμονής σε δευτερόλεπτα
timeout = 10
socket.setdefaulttimeout(timeout)
# Η κλήση στο urllib.request.urlopen χρησιμοποιεί το timeout
# που ορίστηκε στη βιβλιοθήκη socket
url = urllib.request.Request('https://www.upatras.gr/el')
try:
    with urllib.request.urlopen(url) as response:
        char_set = response.headers.get_content_charset()
        html = response.read().decode(char_set)
    with open("www_upatras_gr.html", "w", encoding = char_set) as p:
        p.write(html)
except urllib.error.HTTPError as e:
```

20. Ανάκτηση ιστοσελίδων: δεδομένα GET

```
import urllib.request
import urllib.error
my_UA = "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12_1) Apple
url = 'https://www.google.gr/#q=python&*'
try:
    headers = {}
    headers['User-Agent'] = my_UA
    req = urllib.request.Request(url, headers = headers)
    with urllib.request.urlopen(req) as response:
        char_set = response.headers.get_content_charset()
        html = response.read().decode(char_set)
    with open("www_google_com_q_python.html", "w", encoding = char_set) as p:
        p.write(html)
except urllib.error.HTTPError as e:
    print('Σφάλμα HTTP:', e.code)
except urllib.error.URLError as e:
        print('Αποτυχία σύνδεσης στον server')
        print('Αιτία: ', e.reason)
else:
    print('τέλος προγράμματος')
```



Ορισμός User-Agent

[idle]

20. urllib Ασκήσεις

Να αναπτύξετε τις εξής εφαρμογές με χρήση των βιβλιοθηκών urllib και re

20.1 Ξεκινώντας από τη σελίδα https://service.eudoxus.gr/public/departments διαλέξτε ένα Πανεπιστημιακό Τμήμα και να αναζητήσετε τα μαθήματα του και τους κωδικούς τους

20.2 Ξεκινώντας από τη σελίδα καθηγητών του πανεπιστημιακού τμήματος : http://www.ece.upatras.gr/gr/personnel/faculty.html να αναζητήσετε τα τηλέφωνα των μελών του Τμήματος

[idle]