ΑΛΕΞΑΝΔΡΕΙΟ Τ.Ε.Ι ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επισκόπηση των γλωσσών προγραμματισμού για εξόρυξη δεδομένων και ανάπτυξη τέτοιου προγράμματος για τον ιστοχώρο του χρυσού οδηγού.

Θεμελιώδεις έννοιες

Τι είναι η εξόρυξη δεδομένων από το διαδίκτυο;

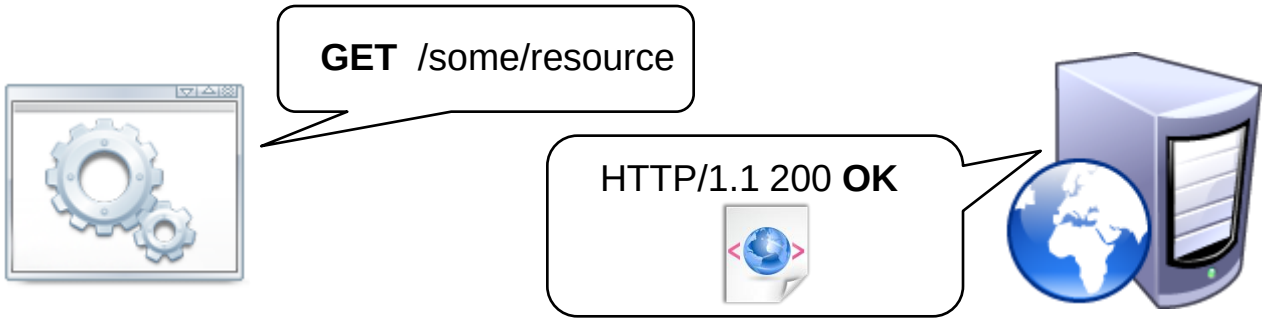
Είναι μια αυτοματοποιημένη διαδικασία άντλησης πληροφοριών και δεδομένων από μια ή περισσότερες ιστοσελίδες.

Θεμελιώδεις έννοιες

Τι είναι η εξόρυξη δεδομένων από το διαδίκτυο;

Είναι μια αυτοματοποιημένη διαδικασία άντλησης πληροφοριών και δεδομένων από μια ή περισσότερες ιστοσελίδες.

Στάδιο 1ο'
(Ανάκτηση του περιεχομένου των ιστοσελίδων)

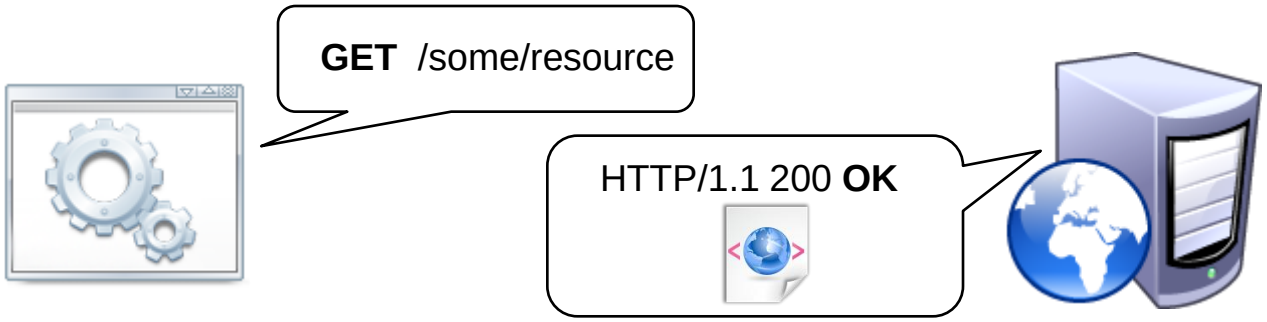


Θεμελιώδεις έννοιες

Τι είναι η εξόρυξη δεδομένων από το διαδίκτυο;

Είναι μια αυτοματοποιημένη διαδικασία άντλησης πληροφοριών και δεδομένων από μια ή περισσότερες ιστοσελίδες.

Στάδιο 1ο' (Ανάκτηση του περιεχομένου των ιστοσελίδων)



Στάδιο 2ο' (Εξαγωγή και αποθήκευση των χρήσιμων δεδομένων)

Συγκεντρωτικός πίνακας…

Χρήσιμες βιβλιοθήκες, κλάσεις και συναρτήσεις σε γνωστές γλώσσες προγραμματισμού

	URL fetching	HTML Parsing	Regexp
Ruby	open-uri rest-open-uri	Htree/ReXML Hpricot RubyfulSoup WWW::Mechanize ScRUBYt! Watir	Regexp
Perl	LWP::Simple	HTML::TreeBuilder WWW::Mechanize Web::Scraper	//, =~, …
Python	urllib urllib2	HTMLParser BeautifulSoup lxml Mechanize scrape.py	re
PHP	fopen loadHTMLFile	Simple HTML DOM Parser htmlSQL DOMDocument+Xath	preg_match preg_match_all …
Java	java.net.URL	JSoup	Java.util.regexp
.NET	System.Net.HTTPWeb Request	HTMLAgilityPack	System.Text.RegularExpressions

Κανονικές Εκφράσεις

```
<td>Current <strong>UTC</strong> (or GMT/Zulu)-time used: <strong
id="ctu">Saturday, March 5, 2011 at 21:54:12</strong><br>
<span class="small">UTC is Coordinated Universal Time, GMT is
Greenwich Mean Time.</span></td>
```

```
perl -MLWP::Simple -le
'$c = get("http://timeanddate.com/worldclock/");
$c =~ m@<strong id="ctu">(.*?)</strong>@ and print $1'

#Saturday, March 5, 2011 at 21:54:12
```

Μειωνεκτήματα:

- Εκφράσεις 'εύθραυστες' ακόμα και με μικρές αλλαγές στον HTML κώδικα.
- Εκφράσεις δύσκολες στη κατανόηση
- Δύσκολη διαχείριση ειδικών χαρακτήρων της HTML όπως οι ©, & κ.τ.λ.

XPath

```
<td>Current <strong>UTC</strong> (or GMT/Zulu)-time used: <strong id="ctu">Saturday, March 5, 2011 at 21:54:12</strong><br>
<span class="small">UTC is Coordinated Universal Time, GMT is Greenwich Mean Time.</span></td>
```

```
use HTML::TreeBuilder::XPath;

my $tree = HTML::TreeBuilder::XPath->new_from_content($content);
print $tree->findnodes ('//strong[@id="ctu"]') ->shift->as_text;

#Saturday, March 5, 2011 at 21:54:12
```

Πλεονεκτήματα έναντι των κανονικών εκφράσεων:

- Λιγότερο ‘εύθραυστη’ προσέγγιση
- Πιο κατανοητές εκφράσεις και πιο εύκολη συντήρηση του κώδικα

CSS Selectors

```
<td>Current <strong>UTC</strong> (or GMT/Zulu)-time used: <strong
id="ctu">Saturday, March 5, 2011 at 21:54:12</strong><br>
<span class="small">UTC is Coordinated Universal Time, GMT is
Greenwich Mean Time.</span></td>
```

```
use HTML::TreeBuilder::XPath;
use HTML::Selector::XPath qw(selector_to_xpath);

my $tree = HTML::TreeBuilder::XPath->new_from_content($content);
my $xpath = selector_to_xpath "strong#ctu";
print $tree->findnodes($xpath)->shift->as_text;

#Saturday, March 5, 2011 at 21:54:12
```

Xpath :
//strong[@id="ctu"]
CSS Selector :
strong#ctu

```
require 'open-uri'

url = "http://www.google.com/search?q=ATEI+of+Thessaloniki"
open(url) {
    |page| page_content = page.read()
    links = page_content.scan(/<a class=l.*?href=\"(.*?)\"/).flatten
    links.each {|link| puts link}
}
```

- Υποβολή αιτήσεων μέσω της build-in βιβλιοθήκης ‘open-uri’
- Εξαγωγή δεδομένων με χρήση κανονικών εκφράσεων
- Βλέπουμε την ιστοσελίδα ως αρχείο κειμένου

Ruby

- open-uri
- HTree / REXML
- Hpricot
- Nokogiri
- Watir

```
require 'rubygems'
require 'open-uri'
require 'htree'
require 'rexml/document'

open("http://www.google.com.com") do |page|
    page_content = page.read()
    doc = HTree(page_content).to_rexml
    doc.root.each_element('//img') {|elem| puts elem.attribute('src').value }
end
```

- Μετατροπή του κώδικα σε REXML (build-in XML parser της Ruby)
- Εξαγωγή δεδομένων με XPATH εκφράσεις
- Βλέπουμε την ιστοσελίδα ως DOM δέντρο και όχι σαν αρχείο κειμένου.

- open-uri
- HTree / REXML
- Hpricot
- Nokogiri
- Watir

```
require 'rubygems'
require 'hpricot'
require 'open-uri'

doc = Hpricot(open('http://www.google.com/search?q=ruby'))
links = doc/"//a[@class=l]"
links.map.each {|link| puts link.attributes['href']}
```

- Πιο γρήγορος και πιο εύχρηστος HTML parser από το συνδιασμό Htree/REXML
- Υποστήριξη XPATH εκφράσεων

- open-uri
- HTree / REXML
- Hpricot
- Nokogiri
- Watir

```ruby
require 'nokogiri'
require 'open-uri'

doc = Nokogiri::HTML(open('http://www.google.com/search?q=tenderlove'))

#  css
doc.css('h3.r a.l').each do |link|
  puts link.content
end

#  xpath
doc.xpath('//h3/a[@class="l"]').each do |link|
  puts link.content
end

# συνδιασμός και των δύο.
doc.search('h3.r a.l', '//h3/a[@class="l"]').each do |link|
  puts link.content
end
```

- Πολύ γρήγορος HTML/XML parser
- Υποστήριξη και XPATH και CSS εκφράσεων

- open-uri
- HTree / REXML
- Hpricot
- Nokogiri
- Watir

```ruby
require 'rubygems'

# Καθοδήγηση του MSIE σε Windows
require 'watir'

# Καθοδήγηση του Firefox σε Windows/Mac/Linux
# require 'firewatir'

browser = Watir::Browser.new
browser.goto("http://www.example.com")

# Συμπλήρωση ενός text field
browser.text_field(:name => "text_field").set "Watir"

# Συμπλήρωση και καθαρισμός ενός radio button
browser.radio(:value => "Watir").set
browser.radio(:value => "Watir").clear

# Συμπλήρωση και καθαρισμός ενός checkbox
browser.checkbox(:value => "Ruby").set
browser.checkbox(:value => "Ruby").clear

# Κλικ σε ένα κουμπί
browser.button(:name => "submit").click
```

- urllib + re
- BeautifulSoup
- Mechanize

```
from urllib import urlopen
import re

p = re.compile('<a .*? href="(.*?)">(.*?)</a>')
text = urlopen('http://www.xo.gr/').read()
for url, name in p.findall(text):
    print '%s (%s)' % (name, url)
```

Αποτέλεσμα:

Επικοινωνία (default_076.html)
Βοήθεια (default_077.html)
Συνήθεις Ερωτήσεις (default_078.html)
Σχόλια για "Δικός μου Χρυσός Οδηγός" (default_079.html)
...

urllib + re

BeautifulSoup

Mechanize

```
from BeautifulSoup import BeautifulSoup
import re
import urllib2

url = 'http://blogsearch.google.com/blogsearch?q=python'
response = urllib2.urlopen(url)
html = response.read()

soup = BeautifulSoup(html)
links = soup.findAll('a', id=re.compile("^p-"))
for link in links:
    print link['href']
```

Αποτέλεσμα:
http://www.daniweb.com/forums/thread331122.html
http://geert.vanderkelen.org/post/435/

…

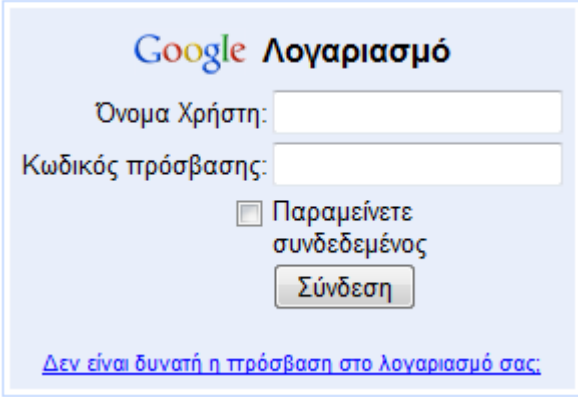
- Ανάλυση no-valid html εγγράφων
- Αυτόματη εύρεση του encoding της σελίδας
- Εύκολη εύρεση στοιχείων της σελίδας (NextSibling, PreviousSibling, next, previous, regex, …

urllib + re

BeautifulSoup

Mechanize

```
import mechanize

br = mechanize.Browser()

#Άνοιγμα της σελίδας
br.open('http://gmail.com')

# Επιλογή της πρώτης φόρμας
br.select_form(nr=0)

# Συμπλήρωση στοιχείων
br.form['Email'] = 'yanis.potamitis'
br.form['Passwd'] = '*********'

# Login
br.submit()
```



- Υποβολή φορμών
- Αποθήκευση ιστορικού επισκέψεων.
- Συνήθως χρησιμοποιείται σε συνδιασμό με το BeautifulSoup

- Simple HTML DOM Parser
- htmlSQL
- loadHTMLFile + XPath

```
$url = 'http://news.google.com/news?pz=1&cf=all&ned=el_gr
@$dom->loadHTMLFile($url);
$xpath = new domxpath($dom);

$xNodes = $xpath->query('//div[@class="title"]');

foreach ($xNodes as $xNode) {
        $sLinktext = $xNode->firstChild->firstChild->nodeValue;
        $sLinkurl = $xNode->firstChild->getAttribute('href');
}
```

Simple HTML DOM Parser

htmlSQL

loadHTMLFile + XPath

SELECT href,title **FROM** a **WHERE** $class == "class_name"

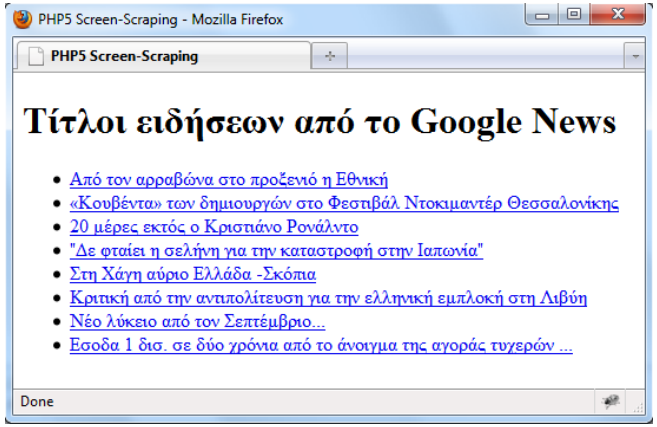
- Σύνταξη παρόμοια με της SQL
- Εύρεση στοιχείων με σύντομες και κατανοητές εκφράσεις

Simple HTML DOM Parser

htmlSQL

loadHTMLFile + XPath

```
$url = 'http://news.google.com/news?pz=1&cf=all&ned=el_gr
@$dom->loadHTMLFile($url);
$xpath = new domxpath($dom);

$xNodes = $xpath->query('//div[@class="title"]');

foreach ($xNodes as $xNode) {
        $sLinktext = $xNode->firstChild->firstChild->nodeValue;
        $sLinkurl = $xNode->firstChild->getAttribute('href');
}
```



Δημιουργία της εφαρμογής

Γιατί επιλέξαμε Python;

- Συνοπτική
- Επεκτάσιμη
- Κατανοήσιμη
- Διαδραστική
- Multiparadigm
- Multiplatform
- Διανέμεται δωρεάν