

EVENTO  
ITALIA  
CON  
R

# Scraping Techniques in Python

Stefano Cotta Ramusino  
<[whitone@gmail.com](mailto:whitone@gmail.com)>  
2009/05/09



# What's the scraping?

## Origin:

scraping data from mainframes  
from green texts on black screens  
to new data structures or API

## Nowadays:

forcing data from old websites in  
something new (web  $\geq$  2.0)



# Why to scrape web pages?

To have online resources available in data structures and files you want, such as:

XML, db, PDF and so on...



# How to scrape?

Necessary elements:

Fuzzy logic

Pattern recognition

This is true hacking technique



# Why Python?

A lot of libraries

Simple regexp, but powerful

Not only an unique  
technique available



# Some books

Atom and RSS - Leslie Orchard  
Wiley Publishing, 2005

Python in a nutshell - Alex Martelli  
O'Reilly, 2006

Beginning Python - Magnus Lie Hetland  
Apress, 2008



# Libraries inside Python

HTMLParser

re



# HTMLParser

```
class FormParser(HTMLParser):  
    """Basic XHTML/HTML form parser"""  
  
    def handle_starttag(self, tag, attrs):  
        if tag == "form":  
            self.handle_startform(attrs)  
        if tag == "input":  
            self.handle_input(attrs)  
  
    def handle_endform(self):  
        if (self._password):  
            # stop parsing  
            pass
```



# HTMLParser

```
def handle_input(self, attrs):  
  
    name = value = ""  
  
    for attr_name, attr_value in attrs:  
        # password input found  
        if attr_name == "type":  
            if attr_value == "password":  
                self._password = True  
  
        if attr_name == "name":  
            name = attr_value  
  
        if attr_name == "value":  
            value = attr_value  
  
    self.inputs[name] = value
```



# Third libraries

Beautiful Soup  
mechanize  
lxml  
html5lib  
scrapemark  
pyquery  
scrapy



# Third libraries

	Pros	Cons
Beautiful Soup	pure	some errors
mechanize	simple	parsing
lxml	speed	unusual
scrapemark	template	no flexibility



# Beautiful Soup

[www.crummy.com/software/BeautifulSoup](http://www.crummy.com/software/BeautifulSoup)

```
soup = BeautifulSoup(webpage)

form = soup.find(type="password").findPrevious("form")

tag_input = form.findAll('input')

for tag in tag_input:
    name = tag['name']
    value = tag['value']

    inputs[name] = value
```



# mechanize

[wwwsearch.sourceforge.net/mechanize](http://wwwsearch.sourceforge.net/mechanize)

```
from mechanize import Browser
```

```
br = Browser()
```

```
br.open(uri)
```

```
assert br.viewing_html()
```

```
br.select_form(name="login")
```

```
br["username"] = "utente"
```

```
br["password"] = "segreto"
```

```
br.submit()
```



# lxml

[codespeak.net/lxml](http://codespeak.net/lxml)

```
for form in page.forms:
    for input in form.inputs:
        if input.type == "password":
            break

form.fields = dict(
    username = "utente",
    password = "segreto"
)

submit_form(form)
```



# scrapemark

[arshaw.com/scrapemark](http://arshaw.com/scrapemark)

```
scrape("""  
    {*  
        <form name='{{ form }}' action='{{ [form].method }}'>  
        {*  
            <input name = '{{ [form].[name] }}'  
                value = '{{ [form].[name].value }}'  
        {*  
    {*  
""", uri)
```



# Questions and answers

[www.whitone.tk](http://www.whitone.tk)

[whitone@gmail.com](mailto:whitone@gmail.com)