

Formal Languages and Compilers

Lab6

Web Scraping

- ▶ An automated technique for gathering huge quantities of information from websites. (from unstructured data in HTML format to structured data in a database or spreadsheet so that it can be used in multiple applications)
- ▶ Usage:
 - **Price Monitoring**
 - **2. Market Research**
 - **3. News Monitoring**
 - **4. Sentiment Analysis** (e.g., collect data from social media to see how a product is viewed by consumers)
 - **5. Email Marketing**

HTML Tags

Tag	Description
<html> ... </html>	Declares the Web page to be written in HTML
<head> ... </head>	Delimits the page's head
<title> ... </title>	Defines the title (not displayed on the page)
<body> ... </body>	Delimits the page's body
<h <i>n</i> > ... </h <i>n</i> >	Delimits a level <i>n</i> heading
 ... 	Set ... in boldface
<i> ... </i>	Set ... in italics
<center> ... </center>	Center ... on the page horizontally
 ... 	Brackets an unordered (bulleted) list
 ... 	Brackets a numbered list
 ... 	Brackets an item in an ordered or numbered list
 	Forces a line break here
<p>	Starts a paragraph
<hr>	Inserts a horizontal rule
	Displays an image here
 ... 	Defines a hyperlink

HTTP

- ▶ Hyper Text Transfer Protocol
- ▶ Communication between client computers and web servers is done by sending **HTTP Requests** and receiving **HTTP Responses**

Method	Description
GET	Request for resource from server
POST	Submit data to the server
HEAD	Same as GET but does not return the body
PUT	The data within the request must be stored at the URL supplied, replacing any existing data.
DELETE	Delete a resource
OPTIONS	Return the HTTP methods supported by the server

HTTP Status Codes



Inspect a web page

The Master's program **Business Administration and Engineering** at the Faculty of Engineering in Foreign Languages include:

- **Business Administration and Engineering**
Presentation, Study plan, Admission, Facebook
- Software Engineering
Presentation, Study plan, Admission, Facebook

Business Administration and Engineering

Inspector Console Debugger Network Style Editor Memory Performance Storage Accessibility Application

Search HTML

```
<h2>Master's Degrees in English</h2>
<p>The Master's program at the Faculty of Engineering in Foreign Languages include:</p>
<ul>
  <li>
    <strong>Business Administration and Engineering</strong>
    <p style="padding-left: 60px;">
      <a href="http://ing.pub.ro/en/education/master/mbae/" target="_blank"> Presentation</a>
      <a href="http://ing.pub.ro/wp-content/uploads/2014/11/2.-Plan-de-invatamant-MBAE-FILS.xlsx.pdf">Study plan</a>
      <a href="http://ing.pub.ro/admission/master/masterat-business-administration-and-engineering/" target="_blank">Admission</a>
      <a href="https://www.facebook.com/MBAE.UPB" target="_blank">Facebook</a>
    </p>
  </li>
  <li>
    <strong>Software Engineering</strong>
    <p style="padding-left: 60px;">
      <a href="http://ing.pub.ro/en/education/master/mbae/" target="_blank"> Presentation</a>
      <a href="http://ing.pub.ro/wp-content/uploads/2014/11/2.-Plan-de-invatamant-MBAE-FILS.xlsx.pdf">Study plan</a>
      <a href="http://ing.pub.ro/admission/master/masterat-business-administration-and-engineering/" target="_blank">Admission</a>
      <a href="https://www.facebook.com/MBAE.UPB" target="_blank">Facebook</a>
    </p>
  </li>
</ul>
```

Filter Styles

This Element

```
element {
  display: inline-block;
}
* {
  -webkit-box-sizing: border-box;
  -moz-box-sizing: border-box;
  box-sizing: border-box;
}
body {
  color: #666;
  font-size: 14px;
  line-height: 1.75em;
}
body {
  font-family: sans-serif;
}
body {
  font-family: "Helvetica Neue", Helvetica, Arial, sans-serif;
  font-size: 14px;
  line-height: 1.42857143;
  color: #333;
}
```

Layout Computed Changes Compatibility

Selected Element

No compatibility issues found.

All Issues

- font-smooth (prefix needed)
-webkit-font-smoothing
10 occurrences
- webkit-tap-highlight-color
html.sb-init
- text-size-adjust (experimental, prefix needed)
-webkit-text-size-adjust
html.sb-init
- clip (deprecated)
span.sr-only
- webkit-overflow-scrolling
div#top-navbar-collapse.collapse.navbar-collapse

Settings

Web Scraping using Python

► Libraries:

- [Requests](#) (pip install requests) - allows to write HTTP requests easily
- [BeautifulSoup](#) (pip install beautifulsoup4) - used for scraping information from web pages
- [Scrapy](#) (open source and collaborative framework for extracting data from websites, written in Python)

Web Scraping using Python

- ▶ Import necessary packages
- ▶ Get content of an URL (content or text)
- ▶ Create a soup object with the URL content

```
import requests
from bs4 import BeautifulSoup
import re
```

```
#make GET request for a URL
req = requests.get("http://ing.pub.ro/en/education/master/")
```

```
#extract content
a=req.content
print(a)
```

```
#extract text
```

```
c=req.text
print(c)
```

```
#create soup object, parsing data as html
soup = BeautifulSoup(a, "html.parser")
print(soup)
```

[illegible]

```
In [2]: runfile('D:/Facultate/Predat/FormalLanguagesAndCompilers/2022/Lab6.py', wdir='D:/Facultate/Predat/FormalLanguagesAndCompilers/2022')
```

```
<!DOCTYPE html>
<html lang="en">
```

```
<head>
  <meta charset="UTF-8">
  <meta name="viewport" content="width=device-width, initial-scale=1">
  <meta name="viewport" content="width=device-width, initial-scale=1.0, minimum-scale=1.0, maximum-scale=1.0, user-scalable=no">
```

```
<title>Master &raquo; FILS</title>
```

```
<link rel="shortcut icon" href="../assets/img/favicon.png">
```

```
<meta name="description" content="">
```

```
<!-- CSS -->
```

```
<link href="//maxcdn.bootstrapcdn.com/bootstrap/3.3.2/css/bootstrap.min.css" rel="stylesheet">
<link rel="stylesheet" href="//maxcdn.bootstrapcdn.com/font-awesome/4.3.0/css/font-awesome.min.css">
<link href="http://ing.pub.ro/wp-content/themes/files/assets/css/files.min.css" rel="stylesheet" media="screen">
```

```
<!-- HTML5 shim and Respond.js IE8 support of HTML5 elements and media queries -->
```

```
<!--[if lt IE 9]>
  <script src="https://cdnjs.cloudflare.com/ajax/libs/html5shiv/3.7.2/html5shiv.min.js"></script>
  <script src="https://cdnjs.cloudflare.com/ajax/libs/respond.js/1.4.2/respond.min.js"></script>
<[/endif-->
```

</head>

```
<a href="#" In [3]: runfile('D:/Facultate/Predat/FormalLanguagesAndCompilers/2022/Lab6.py', wdir='D:/Facultate/Predat/FormalLanguagesAndCompilers/2022')
```

```
search-box" data-bbox="111 100 888 140" style="background-color: #2e3436; color: #eeeeec; padding: 5px;">
  ut-group">
    <!DOCTYPE html>
    <html lang="en">
      <head>
        <meta charset="utf-8"/>

```

```
<meta content="width=device-width, initial-scale=1" name="viewport"/>
<meta content="width=device-width, initial-scale=1.0, minimum-scale=1.0, maximum-scale=1.0, user-scalable=no" name="viewport"/>
```

```
<title>master » FILS</title>
<link href="../../../assets/img/favicon.png" rel="shortcut icon"/>
<meta content="" name="description"/>
<!-- CSS -->
<link href="//maxcdn.bootstrapcdn.com/bootstrap/3.3.2/css/bootstrap.min.css" rel="stylesheet"/>
<link href="//maxcdn.bootstrapcdn.com/font-awesome/4.3.0/css/font-awesome.min.css" rel="stylesheet"/>
<link href="http://ing.pub.ro/wp-content/themes/fils/assets/css/fils.min.css" media="screen" rel="stylesheet"/>
<!-- HTML5 shim and Respond.js IEB support of HTML5 elements and media queries -->
<!--[if lt IE 9]>
    <script src="https://cdnjs.cloudflare.com/ajax/libs/html5shiv/3.7.2/html5shiv.min.js"></script>
    <script src="https://cdnjs.cloudflare.com/ajax/libs/respond.js/1.4.2/respond.min.js"></script>
<![endif]-->
```

```
</head>
<!-- Preloader -->
<body><div id="preloader">
<div id="status"> </div>
</div>
<div id="sb-site">
<div class="boxed">
```

```
#find content between h1 tags
header1=soup.find('h1').text
print("Header1:")
print(header1, "\n")
```

```
#find content between h2 tags
header2=soup.find('h2').text
print("Header2:")
print(header2, "\n")
```

```
#find the content between paragraph tags
paragraph=soup.find("p").text
print("Paragraphs:")
print(paragraph, "\n")
```

```
#find all hyperlinks with specific HTML attribute
x=soup.find_all("a", attrs={"class":"animated fadeIn animation-delay-7"})
print("A:")
print(x)
```

Header1:
Master

Header2:
Master's Degrees in English

Paragraphs:

Master's Degrees in English

The Master's programs given in English at the Faculty of Engineering in Foreign Languages include:

Business Administration and Engineering

Presentation, Study plan, Admission, Facebook

Software Engineering

Presentation, Study plan, Admission, Facebook

Energy Engineering, in cooperation with T. U. Darmstadt (Germany)

Presentation, Study plan, Admission

Biomaterials for Tissue Engineering (BioTE) - Biomateriale pentru ingineria tesuturilor (în limba engleză)

Presentation, Study plan, Admission

A:

```
[<a class="animated fadeIn animation-delay-7" href="http://ing.pub.ro/en/education/master/">

</a>]
```



a.animated.fadeIn.animation-delay-7 | 44 x 31

The screenshot shows the Fils website header. The Fils logo is on the left. To its right is a navigation bar with flags for Romania, UK, France, and Germany. A search bar is on the far right. Below the navigation bar is a blue button with a white menu icon. The Chrome DevTools Inspector is open at the bottom, showing the HTML structure of the page. The selected element is the link to the Master's Degrees in English page, which contains an image of the English flag.


```
#get the upload links from wensite
#1 from the content of the request
pattern=re.compile(r'ing.pub.ro/wp-content/uploads/.+?(?=">')
pdfs=pattern.findall(c)
print(pdfs)

#2 from the content of a specific attribute
pdf=soup.find_all(attrs={"style":"padding-left: 60px;"})
pdfs2=pattern.findall(str(pdf))
print(pdfs2)


#from the soup (must be casted to string)
pdfs3=pattern.findall(str(soup))
print(pdfs3)

#get the names of the links from the UI
pattern2=re.compile(r'pdf">(.*?)</a>')
name=pattern2.findall(str(soup))
print('\n', name)

#zip is used in order to join 2 tuples
with open("output.txt", "w") as f:
    for a, b in zip(pdfs3, name):
        f.write(b + "\t" + a + "\n")
```

```
['ing.pub.ro/wp-content/uploads/2014/11/2.-Plan-de-invatamant-MBAE-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/
uploads/2022/04/1.-Plan-de-invatamant-ISP-FILS.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/Prezentare_EE.pdf',
'ing.pub.ro/wp-content/uploads/2014/11/6.-Plan-Invatamant-EE-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/uploads/
2014/11/Descriere-BioTE.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/7.-Plan-Invatamant-BIOTE-FILS.xlsx.pdf',
'ing.pub.ro/wp-content/uploads/2019/06/Prezentare-master.pdf" target="_blank', 'ing.pub.ro/wp-content/uploads/
2014/11/8.-Plan-Invatamant-AMPD-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/Flyer_AdvMatProc.pdf']
['ing.pub.ro/wp-content/uploads/2014/11/2.-Plan-de-invatamant-MBAE-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/
uploads/2022/04/1.-Plan-de-invatamant-ISP-FILS.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/Prezentare_EE.pdf',
'ing.pub.ro/wp-content/uploads/2014/11/6.-Plan-Invatamant-EE-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/uploads/
2014/11/Descriere-BioTE.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/7.-Plan-Invatamant-BIOTE-FILS.xlsx.pdf',
'ing.pub.ro/wp-content/uploads/2019/06/Prezentare-master.pdf" target="_blank', 'ing.pub.ro/wp-content/uploads/
2014/11/8.-Plan-Invatamant-AMPD-FILS.xlsx.pdf', 'ing.pub.ro/wp-content/uploads/2014/11/Flyer_AdvMatProc.pdf']

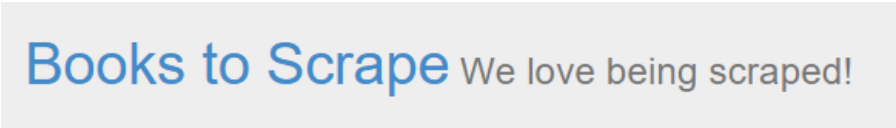

['Study plan', 'Study plan', 'Presentation', 'Study plan', 'Presentation', 'Study plan', 'Study plan',
'Poster']
```

 output.txt - Notepad

File Edit Format View Help

Study plan	ing.pub.ro/wp-content/uploads/2014/11/2.-Plan-de-invatamant-MBAE-FILS.xlsx.pdf
Study plan	ing.pub.ro/wp-content/uploads/2022/04/1.-Plan-de-invatamant-ISP-FILS.pdf
Presentation	ing.pub.ro/wp-content/uploads/2014/11/Prezentare_EE.pdf
Study plan	ing.pub.ro/wp-content/uploads/2014/11/6.-Plan-Invatamant-EE-FILS.xlsx.pdf
Presentation	ing.pub.ro/wp-content/uploads/2014/11/Descriere-BioTE.pdf
Study plan	ing.pub.ro/wp-content/uploads/2014/11/7.-Plan-Invatamant-BIOTE-FILS.xlsx.pdf
Study plan	ing.pub.ro/wp-content/uploads/2019/06/Prezentare-master.pdf" target="_blank
Poster	ing.pub.ro/wp-content/uploads/2014/11/8.-Plan-Invatamant-AMPD-FILS.xlsx.pdf

Exercises

- ▶ Use <http://books.toscrape.com/>
- ▶ Extract  We love being scraped!
- ▶ Extract  Page 1 of 50
- ▶ Extract the content of the button tag
- ▶ Use the name of the class (product_pod) in order to extract the products title and price
 - Create a regular expression in order to get the title from it.
 - Create a regular expression in order to get the price from it.
- ▶ Extract all the book categories (hint: find all class, iterate through the result and find li)
- ▶ Extract the content of img alt="" (hint: find all img, regex for alt content)

Homework

- ▶ Use <https://www.scrapethissite.com/pages/simple/>
- ▶ Extract:
 - Country name
 - Capital
 - Population
 - Area
- ▶ Create an excel file with columns "CountryName", "Capital", "Population", "Area" and populate it with the information extracted from the website

- ▶ Extract the links from

 [Scrape This Site](#)

 [Sandbox](#)


 [Lessons](#)

 [FAQ](#)

[Login](#)

- ▶ Extract [Lessons and Videos](#) © Hartley Brody 2018

- ▶ Extract

 There are 4 [video lessons](#) that show you how to scrape this page.

Data via <http://peric.github.io/GetCountries/>