Corso ITS:

PROGETTISTA E SVILUPPATORE SOFTWARE:

FULL STACK DEVELOPER E CLOUD SPECIALIST

Modulo: **Programmazione in Python**

Docente: *Andrea Ribuoli*

Martedì **29 Aprile** 2025

09:00 - 14:00

In [2]:
```
html = """
<html>
 <head>
  <title>
    CORSO PYTHON
  </title>
 </head>
 <body>
  <p>
    Facciamo un primo esempio
  </p>
  <p>
    di pagina HTML
  </p>
  <p>
    composta di paragrafi.
  </p>
 </body>
</html>
"""
```

In [3]:
```
html_compatto = "".join([_.strip() for _ in html.split("\n")])
```

In [5]:
```
print(html_compatto)
```

```
<html><head><title>CORSO PYTHON</title></head><body><p>Facciamo un primo ese
mpio</p><p>di pagina HTML</p><p>composta di paragrafi.</p></body></html>
```

In [6]:
```
import bs4
doc = bs4.BeautifulSoup(html_compatto)
```

In [7]:
```
print(doc.prettify())
```

```
<html>
 <head>
  <title>
   CORSO PYTHON
  </title>
 </head>
 <body>
  <p>
   Facciamo un primo esempio
  </p>
  <p>
   di pagina HTML
  </p>
  <p>
   composta di paragrafi.
  </p>
 </body>
</html>
```

BeautifulSoup

- le pagine web sono scritte nel linguaggio **HTML**
- *HTML* = **HyperText Markup Language**
- sono composte da elementi con contenuto racchiuso tra una coppia di **tag**
- *tag* = **marcatore**
- nel corpo di un tag possono essere annidati altri tag
- **Beautiful Soup** (*zuppa meravigliosa*) trasforma una "zuppa" di marcatori
- in strutture correttamente annidate

In [10]:
```
root = doc.contents[0]  ### HTML
print(type(root))
print(root.name.upper())
```

```
<class 'bs4.element.Tag'>
HTML
```

In [11]:
```
print(type(root.contents))
```

```
<class 'list'>
```

In [12]:
```
print(type(root.contents[0])) ### HEAD
```

```
<class 'bs4.element.Tag'>
```

In [13]:
```
head = root.contents[0]
print(head.name.upper())
```

```
HEAD
```

In [14]:
```
for e in root.contents:
    print(e.name.upper())
```

```
HEAD
BODY
```

- cenni alle funzioni ricorsive

In [15]:
```
def fattoriale(n) :
    if n == 1:
        return 1
    else:
        return n * fattoriale(n - 1)
```

In [16]:
```
print(fattoriale(4))
```

```
24
```

- applichiamo il concetto (funzione ricorsiva)
- all'attraversamento dell'albero costituito dal nesting dei tag HTML

In [17]:
```
def naviga(tag):
    print(tag.name.upper())
    for stag in tag.contents:
        if type(stag) == bs4.element.Tag :
            naviga(stag)
```

In [18]:
```
naviga(root)
```

```
HTML
HEAD
TITLE
BODY
P
P
P
```

- aggiungiamo una indentazione conseguente alla profondità dell'elemento

In [20]:
```
def naviga2(tag, indent) :
    print(indent + tag.name.upper())
    for stag in tag.contents:
        if type(stag) == bs4.element.Tag :
            naviga2(stag, indent + "  ")
```

In [21]:
```
naviga2(root, "")
```

```
HTML
  HEAD
    TITLE
  BODY
    P
    P
    P
```

---

casi concreti: il web

- il modulo **urllib** è parte della *libreria standard* di Python
- è necessario conoscere l'**URL**: *Uniform Resource Locator*
- e passarlo alla funzione **urlopen()** (del modulo **urllib.request**)
- ci viene restituito un oggetto di tipo **response**
- questo ci offre diversi metodi tra cui **read()**
- che ci restituisce un tipo **bytes** (che già conosciamo)
- se abbiamo passato l'url di una pagina HTML sappiamo di poter
- interpretare tali bytes come testo (default **utf-8**)
- a tal scopo si usa il metodo **decode()**

In [22]:
```
import urllib.request
url = "https://www.andrearibuoli.it"
risultato = urllib.request.urlopen(url)
theBytes = risultato.read()
text = theBytes.decode()
import bs4
doc = bs4.BeautifulSoup(text)
print(doc.prettify())
```

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w
3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
 <head>
  <meta content="text/html; charset=utf-8" http-equiv="Content-Type"/>
  <script crossorigin="anonymous" src="https://kit.fontawesome.com/dacb4f40d
7.js">
  </script>
  <title>
   www.andrearibuoli.it
  </title>
  <style type="text/css">
   /*<![CDATA[*/
 body {
  text-align: center;
  color: #336699;
  font-family: Arial, Helvetica, sans-serif;
  font-size: 48px;
  font-weight: bold;
}
 h2.c1 {
  color: #CC6600;
  font-size: 24px;
  font-weight: lighter;
}
 p.c2 {
  color: #CC6600;
  font-size: 13px;
  font-weight: lighter;
}
/*]]>*/
  </style>
 </head>
 <body>
  www.andrearibuoli.it
  <div class="collapse navbar-collapse" id="social-icons">
   <ul>
    <li>
     <a href="https://www.andrearibuoli.it/wp">
      <span class="fa fa-wordpress" data-placement="bottom" data-toggle="too
ltip" title="WordPress">
       <span class="visible-xs-inline">
        WordPress
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://it.linkedin.com/in/andrea-ribuoli-5b37403b">
      <span class="fa fa-linkedin-square" data-placement="bottom" data-toggl
e="tooltip" title="LinkedIn">
       <span class="visible-xs-inline">
        LinkedIn
       </span>
      </span>
     </a>
```

```
    </li>
    <li>
     <a href="https://github.com/AndreaRibuoli">
      <span class="fa fa-github" data-placement="bottom" data-toggle="toolti
p" title="GitHub">
       <span class="visible-xs-inline">
        GitHub
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/PD_29_03_2025.zip">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
="tooltip" title="ITPASS Meeting 29 Marzo 2025">
       <span class="visible-xs-inline">
        ITPASS Meeting 29/03/2025
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/PU_30_11_2024.zip">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
="tooltip" title="ITPASS Meeting 30 Novembre 2024">
       <span class="visible-xs-inline">
        ITPASS Meeting 30/11/2024
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/PU_20_4_2024.zip">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
="tooltip" title="ITPASS Meeting 20 Aprile 2024">
       <span class="visible-xs-inline">
        ITPASS Meeting 20/04/2024
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/ITPASS_11_novembre_2023.t
ar.gz">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
="tooltip" title="ITPASS Meeting 11 Novembre 2023">
       <span class="visible-xs-inline">
        ITPASS Meeting 11/11/2023
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/ITPASS_6_maggio_2023.ta
r.gz">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
```

```
="tooltip" title="ITPASS Meeting 6 Maggio 2023">
       <span class="visible-xs-inline">
        ITPASS Meeting 06/05/2023
       </span>
      </span>
     </a>
    </li>
    <li>
     <a href="https://www.andrearibuoli.it/Meeting/Meeting%20ITPASS%2016%20O
ttobre%202021.zip">
      <span class="fa fa-file-archive-o" data-placement="bottom" data-toggle
="tooltip" title="ITPASS Meeting 16 Ottobre 2021">
       <span class="visible-xs-inline">
        ITPASS Meeting 16/10/2021
       </span>
      </span>
     </a>
    </li>
   </ul>
  </div>
  <div data-iframe-height="270" data-iframe-width="180" data-share-badge-hos
t="https://www.youracclaim.com" data-share-badge-id="a295ff45-8185-4d62-a28
a-b485f1c4b581">
  </div>
  <script async="" src="//cdn.youracclaim.com/assets/utilities/embed.js" typ
e="text/javascript">
  </script>
  <div data-iframe-height="270" data-iframe-width="180" data-share-badge-hos
t="https://www.credly.com" data-share-badge-id="cdca0969-2f1a-418d-a57c-6f50
da9e0c5d">
  </div>
  <script async="" src="//cdn.credly.com/assets/utilities/embed.js" type="te
xt/javascript">
  </script>
 </body>
</html>
```