

Python Library: lxml

DSCI 551

Wensheng Wu

Installing Python library lxml

- `sudo apt install python3-lxml`

```
▼<bib>
  <cd>abc</cd>
  ▼<book>
    <publisher>Addison-Wesley</publisher>
    <author>Serge Abiteboul</author>
    ▼<author>
      <first-name>Rick</first-name>
      <last-name>Hull</last-name>
    </author>
    <author age="20">Victor Vianu</author>
    <title>Foundations of Databases</title>
    <year>1995</year>
    <price>38.8</price>
  </book>
  ▼<book price="55">
    <publisher>Freeman</publisher>
    <author>Jeffrey D. Ullman</author>
    <title>Principles of Database and Knowledge Base Systems</title>
    <year>1998</year>
  </book>
  ▼<book>
    <title>xyz</title>
    <author/>
  </book>
</bib>
```

Example

- `from lxml import etree`
- `f = open('bibs.xml')`
- `tree = etree.parse(f)`
- `print(etree.tostring(tree).decode('utf-8'))`

Example

- for element in tree.xpath("//author"):
 print(etree.tostring(element))



```
<author>Serge Abiteboul</author>
```

```
<author><first-name>Rick</first-name><last-name>Hull</last-name></author>
```

```
<author age="20">Victor Vianu</author>
```

```
<author>Jeffrey D. Ullman</author>
```

```
<author/>
```

Example

- `for element in tree.xpath("//author"):`
 `print element.tag, element.text`

=>

author Serge Abiteboul

author None

author Victor Vianu

author Jeffrey D. Ullman

author None

Example

- for element in tree.xpath('//author[first-name="Rick"]'):

```
    print(etree.tostring(element))
```

=>

```
<author><first-name>Rick</first-name><last-name>Hull</last-name></author>
```

Helper function

```
def printf(elems):  
    if (isinstance(elems, list)):  
        for elem in elems:  
            if isinstance(elem, str):  
                print(elem)  
            else:  
                print(etree.tostring(elem).decode('utf-8'))  
    else: # just a single element  
        print(etree.tostring(elems).decode('utf-8'))
```

- `printf(tree.xpath('//author[first-name="Rick"]'))`

Work with HTML document

```
from lxml import html
```

```
myfile = open('Express.html')
```

```
htree = html.parse(myfile)
```

```
▼ <table border="1"> == $0
  ▼ <thead>
    ▼ <tr>
      <td>Account number</td>
      <td>First name</td>
      <td>Last name</td>
      <td>Address</td>
      <td>Balance</td>
    </tr>
  </thead>
  ▼ <tbody>
    ▼ <tr>
      <td>136</td>
      <td>Winnie</td>
      <td>Holland</td>
      <td>198 Mill Lane</td>
      <td>45801</td>
    </tr>
    ► <tr>...</tr>
    ► <tr>...</tr>
    ► <tr>...</tr>
```

Work with HTML document

```
print(html.tostring(htree, pretty_print=True))
```

```
htree.xpath('//tbody/tr[1]/td[1]/text()')
```

```
htree.xpath('//tbody/tr[1]/td[2]/text()')
```

```
htree.xpath('//tbody/tr[1]/td[3]/text()')
```

Resources

- Lxml - XML and HTML with Python
 - <https://lxml.de/>