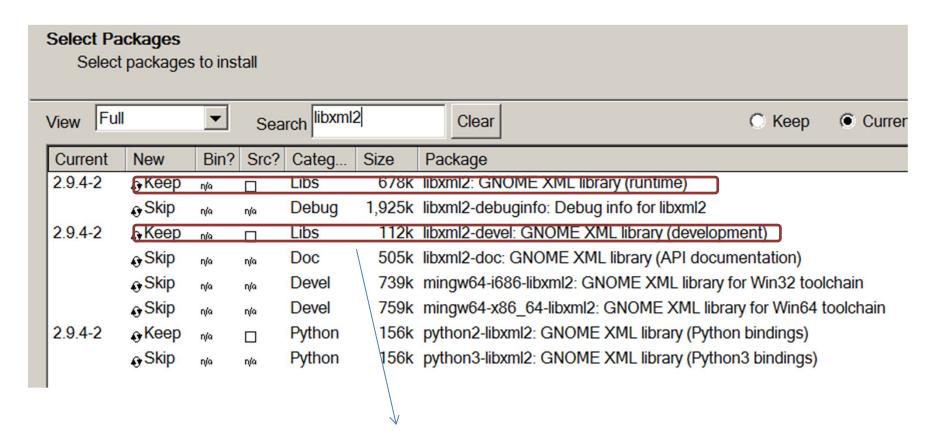
Python Library: Ixml

DSCI 55x Wensheng Wu

Extraction using Python library lxml

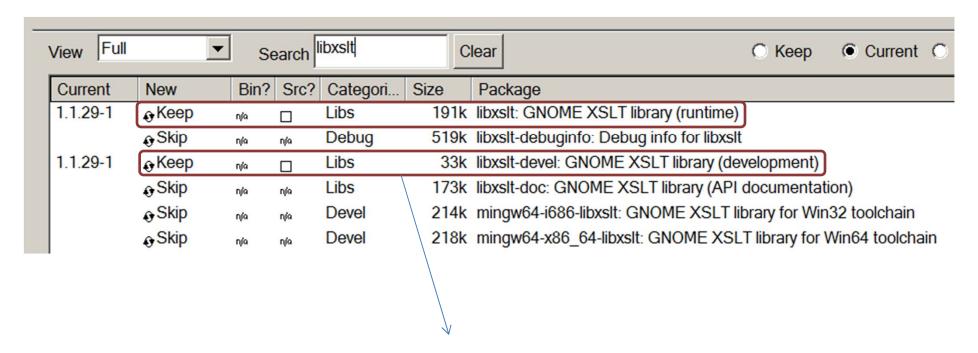
- Ixml should be installed by default in your Amazon EC2 AMI Linux instance
- If not
 - sudo pip install lxml (this is for python 2)
 - sudo python3 -m pip install lxml (python 3)
- If you are using Cygwin
 - First install libxml2, libxslt (see next slide)
 - Then "pip install lxml"

Cygwin



Install libxml2 and libxml2-devel

Cygwin



Install libxslt and libxslt-devel

```
▼ <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>
```

- from lxml import etree
- f = open('bibs.xml')
- tree = etree.parse(f)

print(etree.tostring(tree, pretty_print=True))

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



```
<author>Serge Abiteboul</author>
<author>Serge Abiteboul</author>
<author><first-name>Rick</first-name><last-name>Hull</last-name></author>
<author>ge="20">Victor Vianu</author>
<author>Jeffrey D. Ullman</author>
<author/>
```

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

```
    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, pretty_print=True))
    else: # just a single element
        print(etree.tostring(elems))
```

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

Work with HTML document

from lxml import html

myfile = open('Express.html')

htree = html.parse(myfile)

```
▼ == $0
▼<thead>
 ▼
   Account number
   First name
   Last name
   Address
   Balance
  </thead>
▼
 ▼
   136
   Winnie
   Holland
   198 Mill Lane
   45801
  ▶ ...
 ▶ ...
```

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/