

Re: Processing XML that's embedded in HTML

Source: <http://coding.derkeiler.com/Archive/Python/comp.lang.python/2008-01/msg02654.html>

- *From:* Mike Driscoll <kyosohma@xxxxxxxxxx>
 - *Date:* Tue, 22 Jan 2008 12:48:03 -0800 (PST)
-

On Jan 22, 11:32 am, Paul Boddie <p...@xxxxxxxxxxxxxxxx> wrote:

The rest of the document is html, javascript div tags, etc. I need the information only from the row where the Relationship tag = Owner and the Priority tag = 1. The rest I can ignore. When I tried parsing it with minidom, I get an ExpatError: mismatched tag: line 1, column 357 so I think the HTML is probably malformed.

Or that it isn't well-formed XML, at least.

I probably should have posted that I got the error on the first line of the file, which is why I think it's the HTML. But I wouldn't be surprised if it was the XML that's behaving badly.

I looked at BeautifulSoup, but it seems to separate its HTML processing from its XML processing. Can someone give me some pointers?

With libxml2dom [1] I'd do something like this:

```
import libxml2dom
d = libxml2dom.parse(filename, html=1)
# or: d = parseURI(uri, html=1)
rows = d.xpath("//XML/BoundData/Row")
# or: rows = d.xpath("//XML[@id='grdRegistrationInquiryCustomers']/BoundData/Row")
```

Even though the document is interpreted as HTML, you should get a DOM containing the elements as libxml2 interprets them.

I am currently using Python 2.5 on Windows XP. I will be using Internet Explorer 6 since the document will not display correctly in

Re: Processing XML that's embedded in HTML

Firefox.

That shouldn't be much of a surprise, it must be said: it isn't XHTML, where you might be able to extend the document via XML, so the whole document has to be "proper" HTML.

Paul

[1] <http://www.python.org/pypi/libxml2dom>

I must have tried this module quite a while ago since I already have it installed. I see you're the author of the module, so you can probably tell me what's what. When I do the above, I get an empty list either way. See my code below:

```
import libxml2dom
d = libxml2dom.parse(filename, html=1)
rows = d.xpath('//XML[@id="grdRegistrationInquiryCustomers"]/BoundData/
Row')
# rows = d.xpath("//XML/BoundData/Row")
print rows
```

I'm not sure what is wrong here...but I got lxml to create a tree from by doing the following:

```
<code>
from lxml import etree
from StringIO import StringIO

parser = etree.HTMLParser()
tree = etree.parse(filename, parser)
xml_string = etree.tostring(tree)
context = etree.iterparse(StringIO(xml_string))
</code>
```

However, when I iterate over the contents of "context", I can't figure out how to nab the row's contents:

```
for action, elem in context:
if action == 'end' and elem.tag == 'relationship':
# do something...but what!?!
# this if statement probably isn't even right
```

Thanks for the quick response, though! Any other ideas?

Mike

.

Re: Processing XML that's embedded in HTML