

Re: convert xhtml back to html

Source: <http://coding.derkeiler.com/Archive/Python/comp.lang.python/2008-04/msg03407.html>

- *From:* Stefan Behnel <stefan_ml@xxxxxxxxxx>
 - *Date:* Fri, 25 Apr 2008 08:16:57 +0200
-

bryan rasmussen top-posted:

On Thu, Apr 24, 2008 at 9:55 PM, Stefan Behnel <stefan_ml@xxxxxxxxxx> wrote:

```
from lxml import etree

tree = etree.parse("thefile.xhtml")
tree.write("thefile.html", method="html")
```

<http://codespeak.net/lxml>

wow, that's pretty nice there.

Just to know: what's the performance like on XML instances of 1 GB?

That's a pretty big file, although you didn't mention what kind of XML language you want to handle and what you want to do with it.

lxml is pretty conservative in terms of memory:

<http://blog.ianbicking.org/2008/03/30/python-html-parser-performance/>

But the exact numbers depend on your data. lxml holds the XML tree in memory, which is a lot bigger than the serialised data. So, for example, if you have 2GB of RAM and want to parse a serialised 1GB XML file full of little one-element integers into an in-memory tree, get prepared for lunch. With a lot of long text string content instead, it might still fit.

However, lxml also has a couple of step-by-step and stream parsing APIs:

<http://codespeak.net/lxml/parsing.html#the-target-parser-interface>
<http://codespeak.net/lxml/parsing.html#the-feed-parser-interface>
<http://codespeak.net/lxml/parsing.html#iterparse-and-iterwalk>

They might do what you want.

Stefan

Re: convert xhtml back to html