How to extract just the text from html page articles

October 27, 2012

One of the reasons I keep going back to Python is because of the lxml library.

Not only is it terrific in terms of handling xml, it can do wonders with html of all flavors, even badly-formed and specification-invalid html data.

A common task I have these days is to grab the text from an html page or article (e.g., in curating content for Macaronics).

As this gist shows, lxml makes this dead simple, using xpath and the "descendant-or-self::" axis selector.

The only real work is understanding the page structure and creating the correct xpath expression for each site (the readability algorithm is essentially a collection of these rules), and monitoring their changes over time so that the xpath expression can be updated accordingly.

Another bonus is that it works with foreign language sites, too, provided the parser is passed the same encoding as defined in the target page's Content-Type meta tag.

Here's an example of grabbing the text from a web article by Facta, a Japanese business magazine, and saving it as a text file, so I can add it to the list of articles in Macaronics:

```
>>> import urllib, text_grabber
>>> data=urllib.urlopen('http://facta.co.jp/article/201211043-print.html').read()
>>> t=text_grabber.facta_print(data)
>>> import codecs
>>> f=codecs.open('facta-201211043-print.txt', 'w', 'utf-8'); f.write(t); f.close()
```
