链滴

Python-CookBook：36、在文本中处理 HTML 和 XML 实体

作者：zhaolixiang
原文链接：https://ld246.com/article/1566384916435
来源网站：链滴

问题

我们想将&entity或&#code这样的HTML或XML实体替换为它们相对应的文本。或者，我们需要生文本，但是要对特定的字符（比如<,>或&）做转义处理。

解决方案

如果要生成文本，使用html.escape()函数来完成替换<or>这样的特殊字符相对来说是比较容易的。如：

```
>>> s = 'Elements are written as "<tag>text</tag>".'
>>> import html
>>> print(s)
Elements are written as "<tag>text</tag>".
>>> print(html.escape(s))
Elements are written as "<tag>text</tag>".

>>> # Disable escaping of quotes
>>> print(html.escape(s, quote=False))
Elements are written as "<tag>text</tag>".
>>>
```

如果要生成ASCII文本，并且想针对非ASCII字符将它们对应的字符编码实体嵌入到文本中，可以在各同I/O相关的函数中使用errors='xmlcharrefreplace'参数来实现。示例如下：

```
>>> s = 'Spicy Jalapen~o'
>>> s.encode('ascii', errors='xmlcharrefreplace')
b'Spicy Jalape&#241;o'
>>>
```

要替换文本中的实体，那就需要不同的方法。如果实际上是在处理HTML或XML，首先应该尝试使用个合适的HTML或XML解析器。一般来说，这些工具在解析的过程中会自动处理相关值的替换，而我完全无需为此操心。

如果由于某种原因在得到的文本中带有一些实体，而我们想手工将它们替换掉，通常可以利用各种HTL或XML解析器自带的功能函数和方法来完成。示例如下：

```
>>> s = 'Spicy "Jalapeño&quot.'
>>> from html.parser import HTMLParser
>>> p = HTMLParser()
>>> p.unescape(s)
'Spicy "Jalapen~o".'
>>>

>>> t = 'The prompt is >>>'
>>> from xml.sax.saxutils import unescape
>>> unescape(t)
'The prompt is >>>'
>>>
```

讨论

在生成HTML或XML文档时，适当地对特殊字符做转义处理常常是个容易被忽视的细节。尤其是当自用print()或其他一些基本的字符串格式化函数来产生这类输出时更是如此。简单的解决方案是使用像html.escape()这样的工具函数。

如果需要反过来处理文本（即，将HTML或XML实体转换成对应的字符），有许多像xml.sax.saxutils.nescape()这样的工具函数能帮上忙。但是，我们需要仔细考察一个合适的解析器应该如何使用。例，如果是处理HTML或XML，像html.parser或xml.etree.ElementTree这样的解析模块应该已经解决有关替换文本中实体的细节问题。