2.17 在字符串中处理html和xml¶

问题¶

你想将HTML或者XML实体如 `&entity;` 或 `&#code;` 替换为对应的文本。 再者，你需要转换文本中特定的字符(比如<, >, 或 &)。

解决方案¶

如果你想替换文本字符串中的 '<' 或者 '>'，使用 `html.escape()` 函数可以很容易的完成。比如：

```
>>> 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 &quot;&lt;tag&gt;text&lt;/tag&gt;&quot;.

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

如果你正在处理的是ASCII文本，并且想将非ASCII文本对应的编码实体嵌入进去， 可以给某些I/O函数传递参数 `errors='xmlcharrefreplace'` 来达到这个目。比如：

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

为了替换文本中的编码实体，你需要使用另外一种方法。 如果你正在处理HTML或者XML文本，试着先使用一个合适的HTML或者XML解析器。 通常情况下，这些工具会自动替换这些编码值，你无需担心。

有时候，如果你接收到了一些含有编码值的原始文本，需要手动去做替换， 通常你只需要使用HTML或者XML解析器的一些相关工具函数/方法即可。比如：

```
>>> s = 'Spicy &quot;Jalape&#241;o&quot.'
>>> from html.parser import HTMLParser
>>> p = HTMLParser()
>>> p.unescape(s)
'Spicy "Jalapeño".'
>>>
>>> t = 'The prompt is &gt;&gt;&gt;'
>>> from xml.sax.saxutils import unescape
>>> unescape(t)
'The prompt is >>>'
>>>
```

讨论¶

在生成HTML或者XML文本的时候，如果正确的转换特殊标记字符是一个很容易被忽视的细节。 特别是当你使用 `print()` 函数或者其他字符串格式化来产生输出的时候。 使用像 `html.escape()` 的工具函数可以很容易的解决这类问题。

如果你想以其他方式处理文本，还有一些其他的工具函数比如 `xml.sax.saxutils.unescapge()` 可以帮助你。 然而，你应该先调研清楚怎样使用一个合适的解析器。 比如，如果你在处理HTML或XML文本， 使用某个解析模块比如 `html.parse` 或 `xml.etree.ElementTree` 已经帮你自动处理了相关的替换细节。