作业

做一个xpath和xslt、html之间的文档转换，主要体现转换过程。

答：xslt本是以xpath作为基础的转换

如：<xsl:template match=”price[@unit=’USD’]”>

说明一个带unit属性的元素price，而且unit的值必须为”USD”。

XML代码：

<price>

<unit>USD</unit>

<amount>100</amount>

</price>

<price>

<unit>RMB</unit>

<amount>300</amount>

</price>

与html的转换：

HTML代码：

<html>

<body>

<form id=”loginForm”>

<input name=”username” type=”text”/>

<input name=”password” type=”password”/>

<input name=”continue” type=”submit” value=”Login”/>

<input name=”continue” type=”button” value=”Clear”/>

</form>

</body>

</html>

XPath语法：

查找页面根元素：//

查找页面上所有的input元素：//input

查找页面上第一个form元素内的直接子input元素（即只包括form元素的下一级input元素，使用绝对路径表示，单/号）：//form[1]/input

查找页面上第一个form元素内的所有input子input元素（只要在form元素内的input都算，不管嵌套了多少个其他标签，使用相对路径表示，双//号）：//form[1]//input

查找页面上第一个form元素：//form[1]

查找页面上id为loginForm的form元素：//form[@id=’loginForm’]

查找页面上具有name属性为username的input元素：//input[@name=’username’]

查找页面上id为loginForm的form元素下的第一个input元素：//form[@id=’loginForm’]/input[1]

查找页面上具有name属性为continue并且type属性为button的input元素：//input[@name=’continue’][@type=’button’]

查找页面上id为loginForm的form元素下的第四个input元素：//form[@id=’loginForm’]/input[4]

html = '''

<div id="class3">

我左青龙,

<span id='tiger'>

右白虎,

<ul>上朱雀,

<li>下玄武.</li>

</ul>

老牛在当中,

</span>

龙头在胸口.

</div>

'''

from lxml import etree

selector = etree.HTML(html)

data = selector.xpath('//div[@id="class3"]')[0]

info = data.xpath('string(.)')#实际上是去除了div中间的其他多余标签

print(info)

content2=info.replace('\n','').replace(' ','')#将换行与空格分别取代

print(content2)

# 输出结果:我左青龙,右白虎,上朱雀,下玄武.老牛在当中,龙头在胸口.