《Python 网络爬虫程序技术》

测试 2 参考答案

题目 1

查找文档中所有<p>超级链接包含的文本值

```
from bs4 import BeautifulSoup
doc='''
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>
<p class="story">
Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.
</p>
<p class="story">...</p>
</body>
</html>
'''
soup=BeautifulSoup(doc,"lxml")
_____________________________________
for tag in tags:
    _____________________
```

缺失的语句是：

A.

tags=soup.find_all("p"); print(tag.text)

(正确答案)

B.

tags=soup.find("p"); print(tag.text)

C.

tags=soup.find("p"); print(tag["text"])

D.

tags=soup.find_all("p"); print(tag["text"])

题目 2

找出文档中 The Dormouse's story 的元素节点的所有父节点的名称。

```
from bs4 import BeautifulSoup
doc='''
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>
<p class="story">
Once upon a time there were three little sisters; and their names were
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.
</p>
<p class="story">...</p>
</body>
</html>
'''
soup=BeautifulSoup(doc,"lxml")
print(soup.name)
```

_____

```
while tag:
    print(tag.name)
```

_____

缺失的语句是：

A.

tag=soup.find("b"); tag=tag.parent

(正确答案)

B.

tag=soup.find("b"); tag=tag["parent"]

C.

tag=soup.find_all("b"); tag=tag.parent

D.

tag=soup.find_all("b"); tag=tag["parent"]

题目 3

获取 元素的所有直接子元素节点

```
from bs4 import BeautifulSoup
doc='''
<html><head><title>The Dormouse's story</title></head>
<body>
```

The Dormouse's story Once upon a time ...
</body>
</html>
'''
soup=BeautifulSoup(doc,"lxml")

_____
for x in _____:
    print(x)
缺失的语句是：

A.
tag=soup.find("p"); tag.children
(正确答案)
B.
tag=soup.find("p"); tag.child
C.
tag=soup.find_all("p"); tag.children
D.
tag=soup.find_all("p"); tag.child

题目 4

获取 元素的所有子孙元素节点
from bs4 import BeautifulSoup
doc='''
<html><head><title>The Dormouse's story</title></head>
<body>
The Dormouse's story Once upon a time ...
</body>
</html>
'''
soup=BeautifulSoup(doc,"lxml")

_____
for x in _____:
    print(x)
缺失的语句是：

A.
tag=soup.find("p"); tag.descendants
(正确答案)
B.
tag=soup.find("p"); tag.children
C.

tag=soup.find_all("p"); tag.children

D.

tag=soup.find_all("p"); tag.descendants

题目 5

soup.select("a") 查找文档中所有<a>元素节点；

A. 正确(正确答案)

B. 错误

题目 6

soup.select("p a") 查找文档中所有<p>节点下的所有<a>元素节点；

A. 正确(正确答案)

B. 错误

题目 7

soup.select("p[class='story'] a") 查找文档中所有属性 class="story"的<p>节点下的所有<a>元素节点；

A. 正确(正确答案)

B. 错误

题目 8

soup.select("p[class] a") 查找文档中所有具有 class 属性的<p>节点下的所有<a>元素节点；

A. 正确(正确答案)

B. 错误

题目 9

soup.select("a[id='link1']") 查找属性 id="link1"的<a>节点；

A. 正确(正确答案)

B. 错误

题目 10

soup.select("body head title") 查找<body>下面<head>下面的<title>节点；

A. 正确(正确答案)

B. 错误