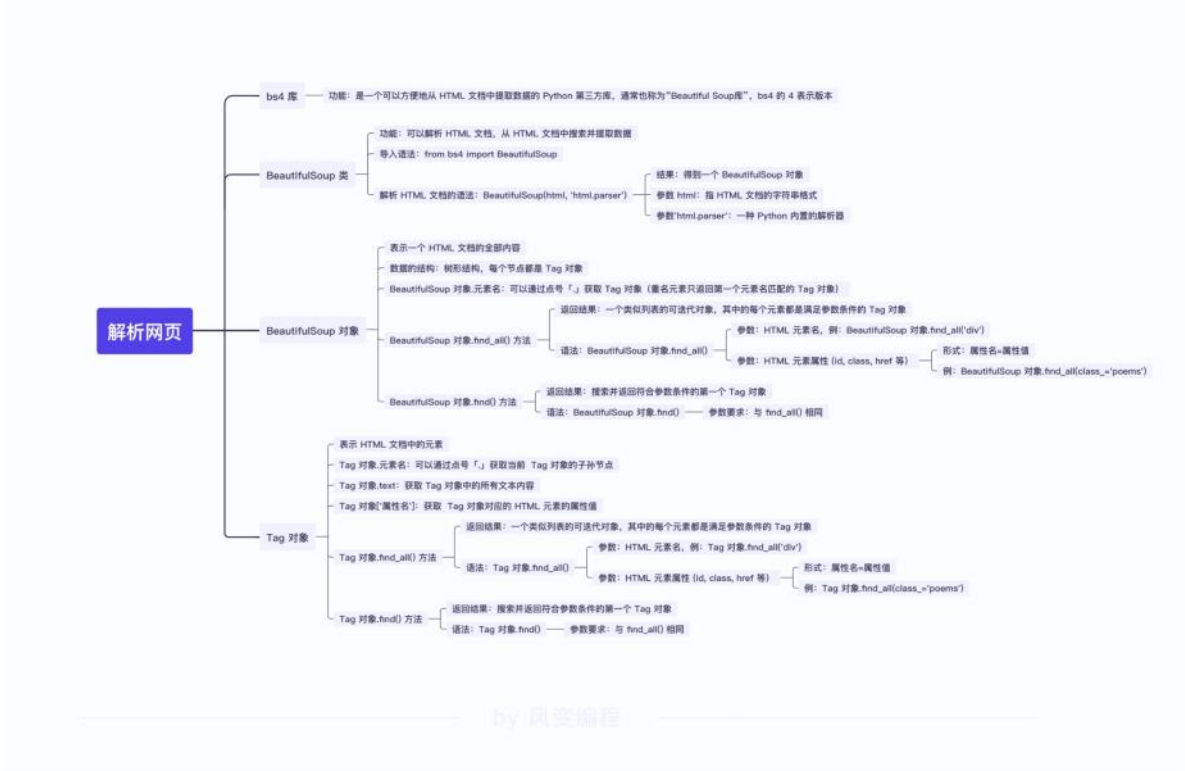
第四关 第四关练习

2022年12月2日 16:48

1. 知识回顾

本节课为爬虫第四关的练习课。在正式开始练习前，让我们通过下面的思维导图回顾一下第四关的重要知识点：



为了让你能更好的巩固第四关的知识点，做好学习下一关的准备。请你试着完成下方的练习吧：

2. 牛刀小试

选择题 1 - BeautifulSoup 类

单选题

以下选项中，哪一项对于 BeautifulSoup 类的实例化语法 BeautifulSoup(html, 'html.parser')描述是正确的？

A.
要想使用 BeautifulSoup 类必须使用语句 import BeautifulSoup 来导入 BeautifulSoup 模块。

B.
第一个参数 html 必须是 Requests 库获取的网络请求内容。

C.
BeautifulSoup 类的实例化语法运行成功后会得到一个 BeautifulSoup 对象。

D.
'html.parser' 是唯一的 Python 解析器。

回答正确

选项 A：BeautifulSoup库的导入语句为 from bs4 import BeautifulSoup
选项 B：第一个参数为 HTML 文档的字符串格式，而 Requests 库获取的网络请求内容刚好就是这么一种格式
选项 D：'html.parser' 只是一种 Python 解析器。

本题的知识包含 BeautifulSoup 类，如果有所遗忘，课后请适当进行回顾复习。

选择题 2 - BeautifulSoup 对象

单选题

以下对 BeautifulSoup 对象的描述错误的是?

A.
BeautifulSoup 对象的数据结构是由 Tag 对象组成的树形结构，每一个节点都是 Tag 对象。

B.
BeautifulSoup 对象.find_all() 方法可以获取所有符合参数条件的 Tag 对象。

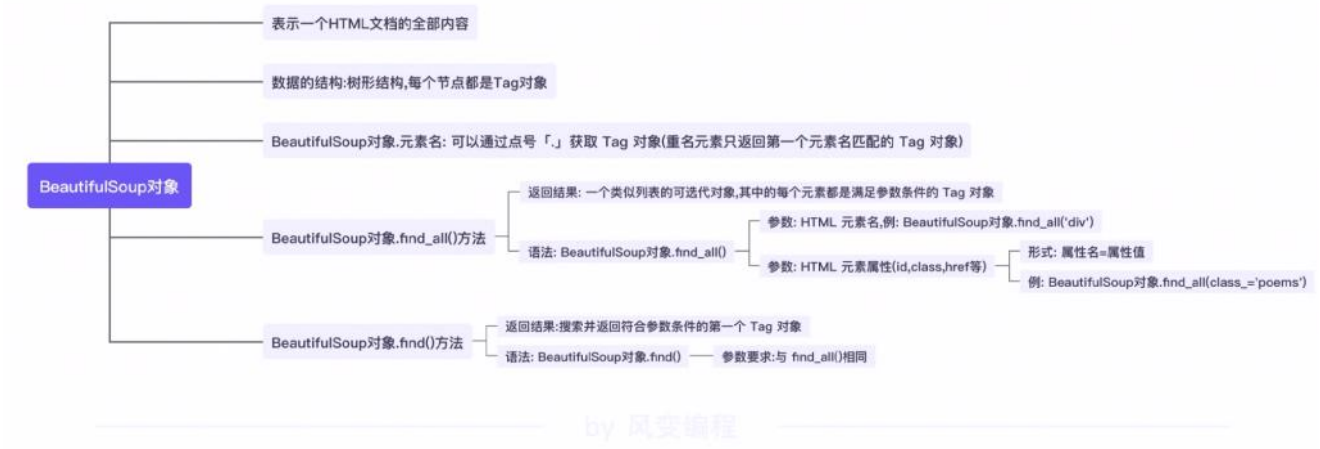
C.
BeautifulSoup 对象.find() 方法可以获取第一个符合参数条件的 Tag 对象。

D.
BeautifulSoup 对象只能通过 find_all() 以及 find() 方法来获取 Tag 对象。

回答正确

我们还可以通过点号 . 即 BeautifulSoup对象.元素名 语句来获取 Tag 对象。

本题的知识包含 BeautifulSoup 对象，如果有所遗忘，课后请适当进行回顾复习。



选择题 3 - Tag 对象

单选题

以下对 Tag 对象的描述错误的是?

A.
我们可以通过 . 即 Tag 对象.属性名 语句来获取 Tag 对象对应的 HTML 元素的属性值。

B.
我们可以通过 Tag 对象.text 语句来获取 Tag 对象中的所有文本内容。

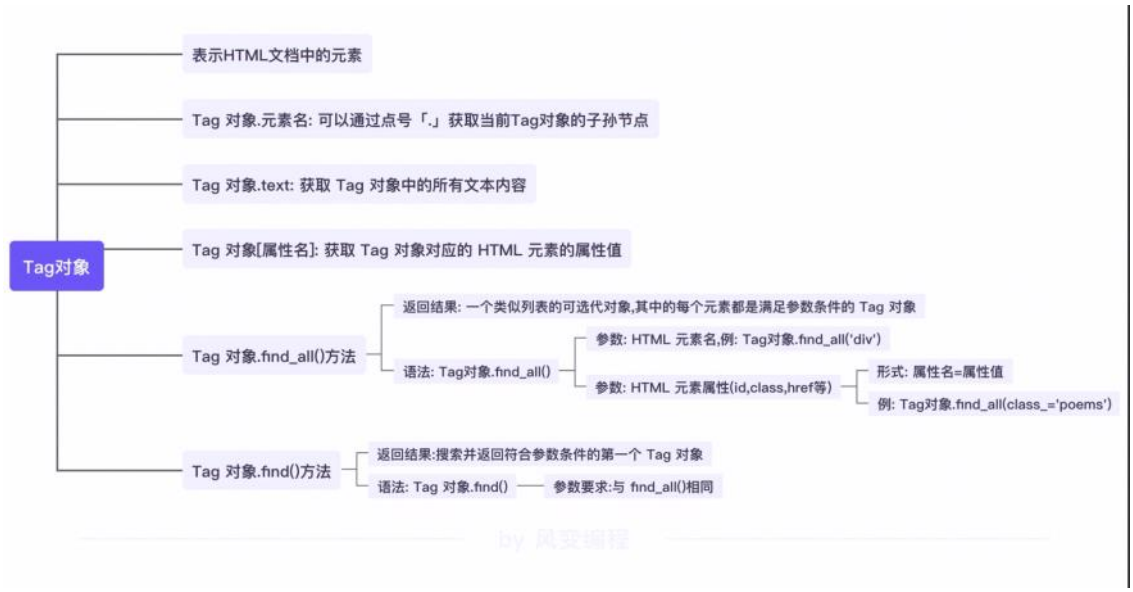
C.
Tag 对象 的 find() 和 find_all() 方法运行效果与 BeautifulSoup 对象一样。

回答错误

答案是 A

获取 Tag 对象对应的 HTML 元素的属性值的语句为**Tag 对象['属性名']**

本题的知识包含 Tag 对象，如果有所遗忘，课后请适当进行回顾复习。



选择题 4 - BeautifulSoup 简单应用

请阅读以下 HTML 代码，回答问题。

```
html = '''
    <div class="movie">
    <div id="title">
    <h2>复仇者联盟</h2>
    </div>
    <div class="director">
    <dl>
    <dd>导演</dd>
    <dt>安东尼·罗素</dt>
    <dt>乔·罗素</dt>
    </dl>
    </div>
    <div class="actor">
    <dl>
    <dd>主演</dd>
    <dt>小罗伯特·唐尼</dt>
    <dt>克里斯·埃文斯</dt>
    </dl>
    </div>'''

bs = BeautifulSoup(html, 'html.parser')
```

下面哪个语句可以获取内容为“复仇者联盟”的元素？

A.
bs.find('h2')
B.
bs.h2

C.
bs.div.h2

D.
bs.find_all('h2')

回答正确

以上选项均可获取内容为“复仇者联盟”的元素。

BeautifulSoup对象.元素名 可以获取第一个元素名匹配的 Tag 对象，bs 对象只有一个 <h2> 元素，因此 B、C 选项都能获取内容为“复仇者联盟”的元素。

本题的知识包含 BeautifulSoup 对象以及Tag 对象，如果有所遗忘，课后请适当进行回顾复习。

3. 代码实战

实战 1 - 网页请求

练习介绍

右侧代码节选了豆瓣电影《复仇者联盟4》详情页的部分代码

题目要求

请你选择合适的方法，补充右方代码的第 44、47~49、52~54 行，实现以下功能：

1）补充 movie_name 变量的定义语句，让其能获取电影名；

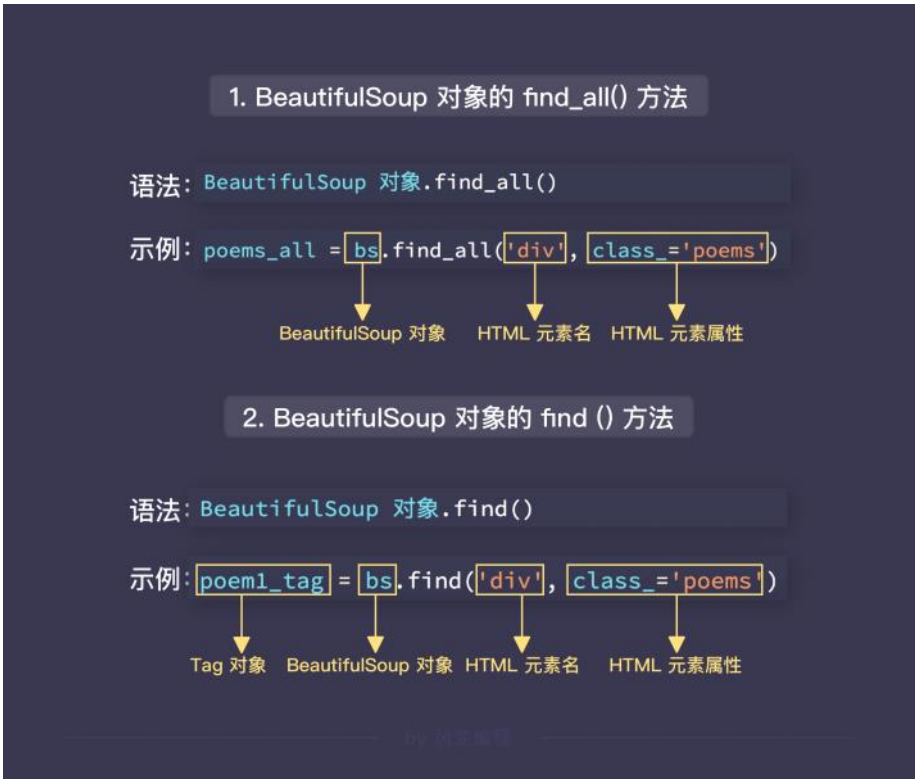
2）补充代码，爬取获取导演名并写进 director 列表；

3）补充代码，爬取主演名并写进 actor 列表，让其能获取主演名。

url 为：https://movie.douban.com/subject/26100958/，你可以点进去查看效果。

书写代码

1）知识点提示



「.元素名」操作，find_all()方法和find()方法

功能	使用方法	参数	结果与功能
.元素名	BeautifulSoup 对象.素名 Tag 对象.元素名	—	结果是一个 Tag 对象，获取第一个元素名匹配的 Tag 对象。
find_all()	BeautifulSoup 对象.find_all() Tag 对象.find_all()	HTML 元素名 HTML 元素属性	结果是一个类似列表的可迭代对象，包含所有满足参数条件的 Tag 对象。
find()	BeautifulSoup 对象.find() Tag 对象.find()	HTML 元素名 HTML 元素属性	结果是一个 Tag 对象，只返回第一个满足参数条件的Tag 对象。

那么，请你正式开始练习吧。

```
from bs4 import BeautifulSoup

html = '''
        <h1>
        <span property="v:itemreviewed">复仇者联盟4：终局之战 Avengers: Endgame</span>
            <span class="year">(2019)</span>
        </h1>
        <span><span class="pl">导演</span>
        ':'
          <span class="attrs">
            <a href="/celebrity/1321812/" rel="v:directedBy">安东尼·罗素</a>
            /
            <a href="/celebrity/1320870/" rel="v:directedBy">乔·罗素</a>
          </span>
        </span>
        <span><span class="pl">编剧</span>
        ':'
          <span class="attrs">
            <a href="/celebrity/1276125/">克里斯托弗·马库斯</a>
            /
            <a href="/celebrity/1276126/">斯蒂芬·麦克菲利</a>
            /
            <a href="/celebrity/1013888/">斯坦·李</a>
            /
            <a href="/celebrity/1050183/">杰克·科比</a>
            /
            <a href="/celebrity/1360715/">吉姆·斯特林</a>
          </span>
        </span>
        <span class="actor">
          <span class="pl">主演</span>
          ':'
          <span class="attrs">
            <span><a href="/celebrity/1016681/" rel="v:starring">小罗伯特·唐尼</a> / </span>
            <span><a href="/celebrity/1017885/" rel="v:starring">克里斯·埃文斯</a> / </span>
            <span><a href="/celebrity/1040505/" rel="v:starring">马克·鲁弗洛</a> / </span>
            <span><a href="/celebrity/1021959/" rel="v:starring">克里斯·海姆斯沃斯</a> / </span>
            <span><a href="/celebrity/1004568/" rel="v:starring">乔什·布洛林</a> / </span>
          </span>
        </span>
        '''
```

```python
bs = BeautifulSoup(html, 'html.parser')

movie_name = bs.find('span',property='v:itemreviewed').text

director = []
director_list = bs.find_all('a',rel='v:directedBy')
for i in director_list:
    director.append(i.text)

actor = []
actor_list = bs.find_all('a',rel='v:starring')
for n in actor_list:
    actor.append(n.text)


movie_dict={'电影名':movie_name,'导演':director,'主演':actor}
print(movie_dict)
```

```
{'电影名': '复仇者联盟4: 终局之战 Avengers: Endgame', '导演': ['安东尼·罗素', '乔·罗素'], '主演': ['小罗伯特·唐尼
', '克里斯·埃文斯', '马克·鲁弗洛', '克里斯·海姆斯沃斯', '乔什·布洛林']}
```

代码讲解

（1）第一步：解析 BeautifulSoup 对象，提取电影名。

电影名称在代码的第 5 行的元素中，属性property="v:itemreviewed"是 HTML 代码中的唯一值。

因此代码的第 44 行，我们只需要对 BeautifulSoup 对象使用 find() 方法，搜索这一元素、属性，提取元素内容即能获取 HTML 代码中的电影名称。

（2）第二步：解析 BeautifulSoup 对象，提取导演名单。

导演名单在代码的第 11、13 行的<a>元素中，属性rel="v:directedBy"只在这两个元素中出现。

因此代码的第 47 行，我们只需要对 BeautifulSoup 对象使用 find_all() 方法，搜索这一元素、属性，即能获取导演名所在的 HTML 元素。

由于我们 find_all() 方法返回的是一个可迭代对象，因此我们还需要使用 for 循环语句将元素名称提取出来。

（3）第三步：解析 BeautifulSoup 对象，提取主演名单。

导演名单在代码的第 34-38 行的<a>元素中，属性rel="v:starring"只在这无个元素中出现。

因此代码的第 52 行，我们只需要对 BeautifulSoup 对象使用 find_all() 方法，搜索这一元素、属性，即能获取主演名所在的 HTML 元素。

这里得到的同样是一个可迭代对象，因此我们还需要使用 for 循环语句将元素名称提取出来。

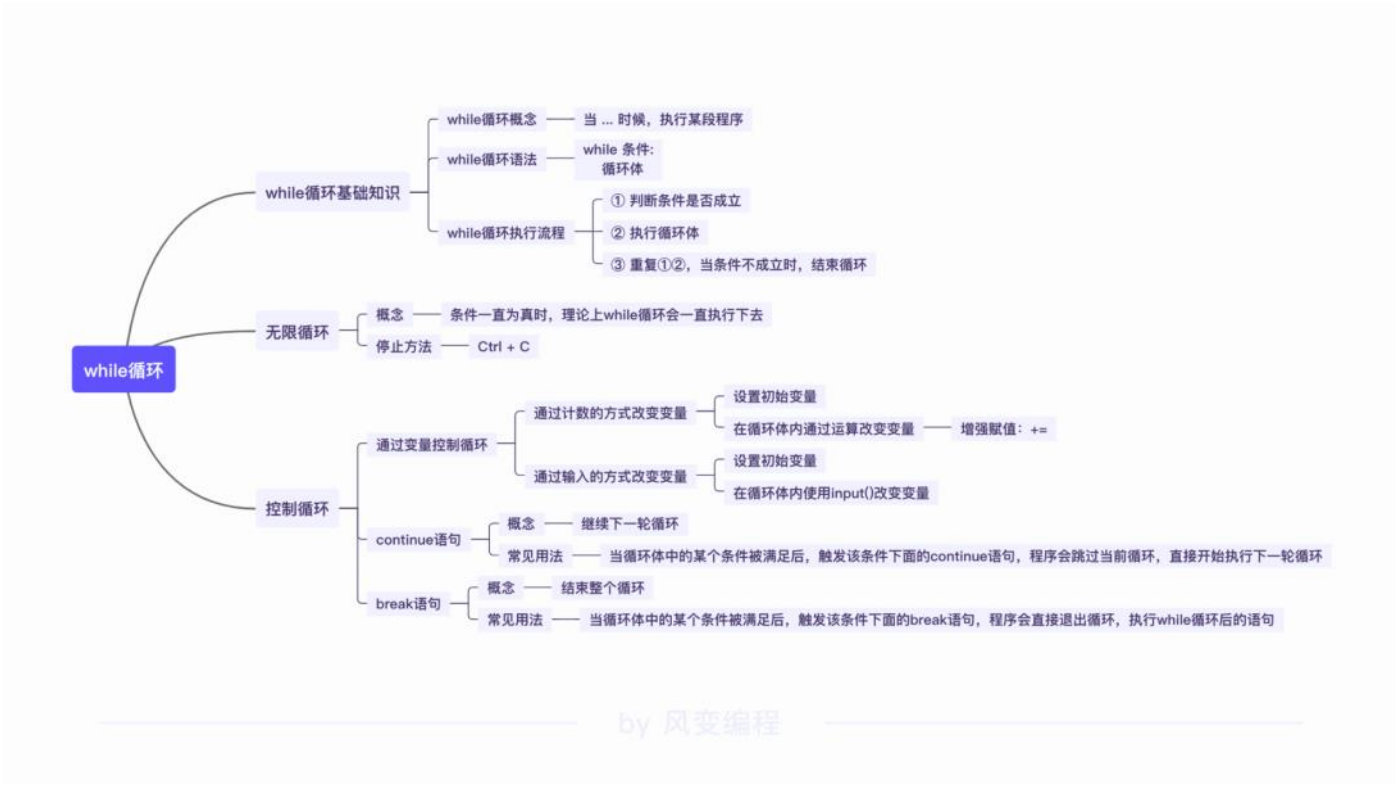
本题的主要知识点是 BeautifulSoup 对象、Tag 对象，如果有所遗忘，课后可以适当进行回顾复习。

```
frombs4importBeautifulSoup

html='''
<h1>
<spanproperty="v:itemreviewed">复仇者联盟4: 终局之战Avengers:Endgame</span>
<spanclass="year">(2019)</span>
</h1>
<span><spanclass="pl">导演</span>
':'
<spanclass="attrs">
<ahref="/celebrity/1321812/"rel="v:directedBy">安东尼·罗素</a>
/
<ahref="/celebrity/1320870/"rel="v:directedBy">乔·罗素</a>
</span>
</span>
<span><spanclass="pl">编剧</span>
':'
<spanclass="attrs">
<ahref="/celebrity/1276125/">克里斯托弗·马库斯</a>
/
<ahref="/celebrity/1276126/">斯蒂芬·麦克菲利</a>
/
<ahref="/celebrity/1013888/">斯坦·李</a>
/
<ahref="/celebrity/1050183/">杰克·科比</a>
/
<ahref="/celebrity/1360715/">吉姆·斯特林</a>
</span>
</span>
```

```
<spanclass="actor">
<spanclass="pl">主演</span>
':'
<spanclass="attrs">
<span><ahref="/celebrity/1016681/"rel="v:starring">小罗伯特·唐尼</a>/</span>
<span><ahref="/celebrity/1017885/"rel="v:starring">克里斯·埃文斯</a>/</span>
<span><ahref="/celebrity/1040505/"rel="v:starring">马克·鲁弗洛</a>/</span>
<span><ahref="/celebrity/1021959/"rel="v:starring">克里斯·海姆斯沃斯</a>/</span>
<span><ahref="/celebrity/1004568/"rel="v:starring">乔什·布洛林</a>/</span>
</span>
</span>
'''
bs=BeautifulSoup(html,'html.parser')

movie_name=bs.find('span',property='v:itemreviewed').text

director=[]
director_list=bs.find_all('a',rel='v:directedBy')
foriindirector_list:
director.append(i.text)

actor=[]
actor_list=bs.find_all('a',rel='v:starring')
fornina ctor_list:
actor.append(n.text)


movie_dict={'电影名':movie_name,'导演':director,'主演':actor}
print(movie_dict)
```

4. 下一课预习

在下节课中，你需要用到很久之前接触过的 while 循环、csv 模块知识。
为了不让你在看到这些知识点时措手不及，我带你重新回顾一下这些知识点的练习。



选择题 5 - while 循环

```
a = 5
while a < 10:
    print(a)
    a = a + 2
print("循环结束")
```

单选题

有关上面代码的描述，下列描述正确的是？

A.
第一轮循环时，先打印5，然后a的值变为7。

B.
第一轮循环时，先打印7，然后a的值不变。

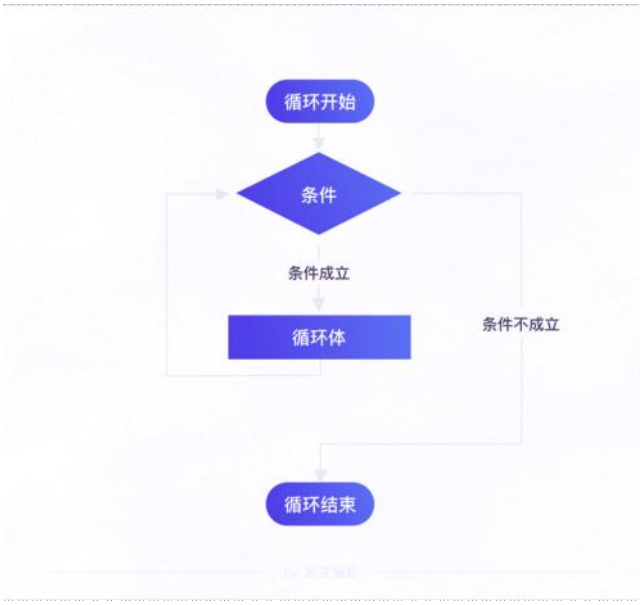
C.
第二轮循环时，先打印9，然后a的值变为11。

D.
一共会循环四轮。

回答正确

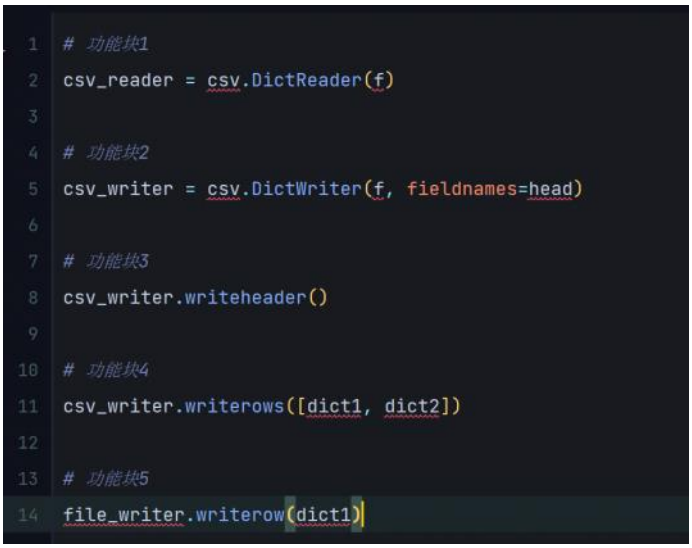
第一轮循环，程序在终端打印5；第二轮循环，程序在终端打印7；第三轮循环，程序在终端打印9，然后a的值变为11，不符合循环条件，循环结束。

本题的知识包含 while循环执行顺序，如果有所遗忘，请适当进行回顾复习。



选择题 6 - csv 模块的类

下方有 5 个功能块，是从某段代码中截取出来的，请你仔细观察并回答问题。



```
# 功能块1
csv_reader = csv.DictReader(f)

# 功能块2
csv_writer = csv.DictWriter(f, fieldnames=head)

# 功能块3
csv_writer.writeheader()

# 功能块4
csv_writer.writerows([dict1, dict2])

# 功能块5
file_writer.writerow(dict1)
```

多选题

请选出选项中正确的项：

A.
功能块1：实例化 DictReader 对象

B.
功能块2：实例化 DictWriter 对象

C.
功能块3：使用 writeheader() 方法写入表头

D.
功能块4：使用 writerows() 方法写入单行

E.
功能块5：使用 writerow() 方法写入多行

回答正确

D、E 选项混淆了 writerow() 方法和 writerows() 方法的功能，正确的答案应该是：writerow() 写入单行，而 writerows() 写入多行。

本题的知识点包含 csv.DictReader() 类、csv.DictWriter() 类、DictWriter 对象.writeheader() 方法、DictWriter 对象.writerow() 方法以及DictWriter 对象.writerows() 方法，如果有所遗忘，课后可以适当进行回顾复习。

恭喜你完成本节课的练习题。