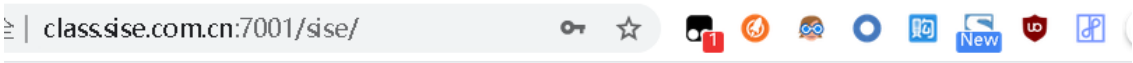
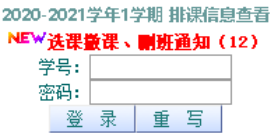
关于华软学院课程表爬虫

一、网址的作用

1. 身份认证首页：http://class.sise.com.cn:7001/sise





网络报修系统

(注：网络报修系统，由网络中心维护，有疑问可联系网络中心，电话87818122)

2. 登录验证：http://class.sise.com.cn:7001/sise/login_check_login.jsp
    - 验证成功返回的结果是：

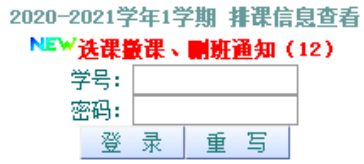
```
result:


<script>top.location.href='/sise/index.jsp'</script>
```

    - 验证失败放回的结果是：

```
<script>alert('警告！ 你不是本站登录的用户');parent.window.opener = 'xxx';parent.window.close();</script>
```

3. 华软学生课表页：http://class.sise.com.cn:7001/sise/module/student_schedular/student_schedular.jsp
    - 当第二步成功的话返回的结果是：

- 当第二步失败的话返回结果是：



这是因为没有通过验证，所以后端将页面重定向到第一步。

二、身份认证首页：http://class.sise.com.cn:7001/sise

1. 首页源代码（有将无关紧要的去除）

```
<html>
<head>
    <title>身份认证 </title>
    <meta http-equiv="Content-Type" content="text/html; charset=gb2312">
    <link rel="stylesheet" href="/sise/css/style.css" type="text/css">
    <script type="text/javascript" src="/sise/js/jquery-1.4.2.min.js">
</script>
    <script type="text/javascript" src="/sise/js/jquery.cookie.min.js">
</script>
    <script type="text/javascript" src="/sise/js/encode.js?0"></script>
</head>
<script>
function loginwithpwd() {
```

```
    if (document.all.username.value.length == 0) {
        alert("请输入用户名！");
        document.all.username.focus();
        return;
    }
    if (document.all.password.value.length == 0) {
        alert("请输入密码！");
        document.all.password.focus();
        return;
    }
    if ((document.all.password.value.length >0) &&
(document.all.username.value.length > 0)) {
        document.getElementById("Submit").disabled  = true;
        document.getElementById("Submit2").disabled  = true;
        form1.submit();
    }
}
function resetWin() {
    document.all.username.value = "";
    document.all.password.value = "";

    document.all.username.focus();
}
function Check_Nums() {
    if (event.keyCode == 13) {
        loginwithpwd();
    }
}
function goNext() {
    if (event.keyCode == 13) {
        form1.password.focus();
    }
}
</script>
<body  text="#000000" topmargin="0" leftmargin="0">
    <div align="center">
        <form name="form1" method="post" action="login_check_login.jsp">
            <input type="hidden" name="2df3c01c64e7817ec5529945fe5052e3"
 value="3d3c43af87065e433e5c973059400829">
            <input id="random"   type="hidden"  value="1603527440833"
 name="random" />
            <input id="token"  type="hidden" name="token" />
             <div> <a href="/sise/coursetemp/courseInfo.html"
target="_blank"><b>2020-2021学年1学期 排课信息查看</b></a></div>
            <div><img alt="有新班" src="/sise/images/new.gif"><a
href="addclass.txt" target="_blank"><b><font color="red">选课撤课、删班通知
（12）</font></b></a></div>
        <div><font size="2" color="#006666">学号：</font><input
name="username" id="username" type="text" size="15" class="notnull"
onkeypress="goNext()" ></div>
            <div><font size="2" color="#006666">密码：</font><input
name="password" id="password" type="password" size="15" class="notnull"
onkeypress="Check_Nums()" ></div>
```

```
            <div><input type="button" id="Submit" name="Submit" value=" 登
录 " class="button" onclick="loginwithpwd();"
onmouseover="this.style.color='red'" onmouseout="this.style.color='#1e7977'">
<input type="button" id="Submit2" name="Submit2"  value=" 重  写 "
class="button" onclick="resetWin();" onmouseover="this.style.color='red'"
onmouseout="this.style.color='#1e7977'"></div>
        </form>
    <div><a href="http://service.scse.com.cn/user/login" target="_blank"
style="font-size:20px;">网络报修系统<br>(注：网络报修系统，由网络中心维护，有疑问可联
系网络中心，电话87818122)</a></div>
    </div>
</body>
</html>
```

- ```
  <form name="form1" method="post" action="login_check_login.jsp"></form>
  //这里明显是一个form表单，将表单里面的参数提交的login_check_login.jsp页面
  //在.jsp页面中可以使用<%request.getParameter("xxx")%>的到xxx属性的值，并验证和
  校验数据。
  ```

- ```
  <div align="center">
      <form name="form1" method="post" action="login_check_login.jsp">
          <input type="hidden" name="2df3c01c64e7817ec5529945fe5052e3"  value="3d3c43af87065e433e5c973059400829
          <input id="random"   type="hidden"  value="1603524979616"  name="random" />
          <input id="token"  type="hidden" name="token" />
          <br>
  ```

  通过form里面的html，可以看出除了requestbody需要提交username的值和password的值之外，还需要提交2df3c01c64e7817ec5529945fe5052e3、random、token的值，所以这里提交的时候，Body中的值为：

  Headers | TextView | SyntaxView | WebForms | HexView | Auth | Cookies | Raw | JSON | XML

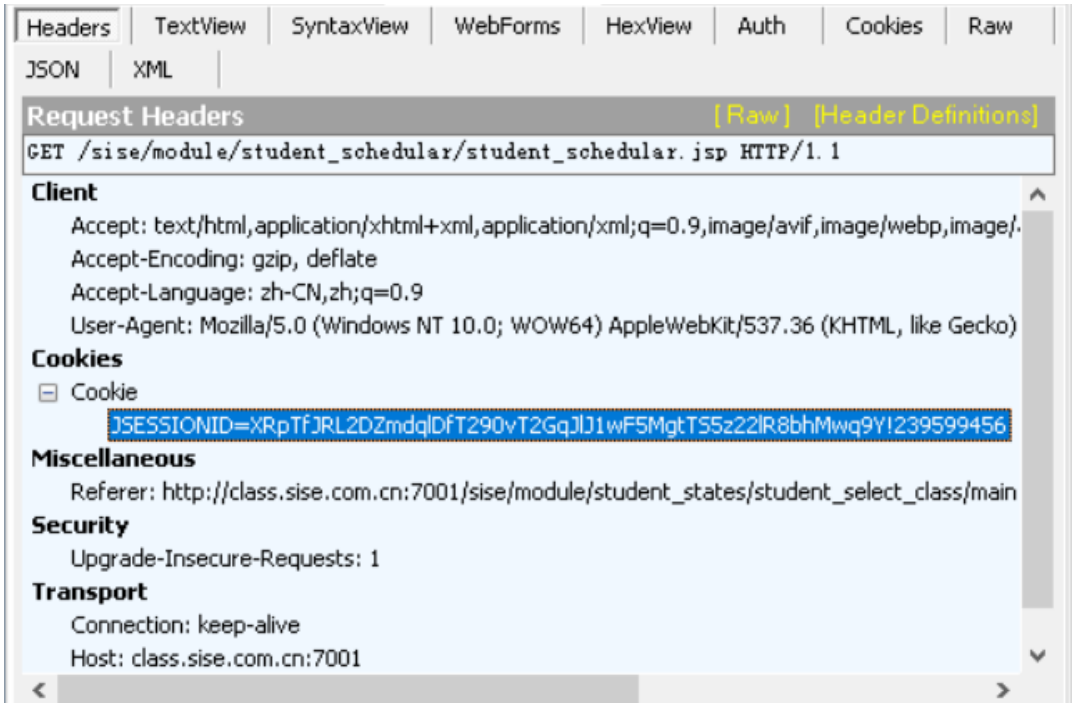
  QueryString

Name	Value

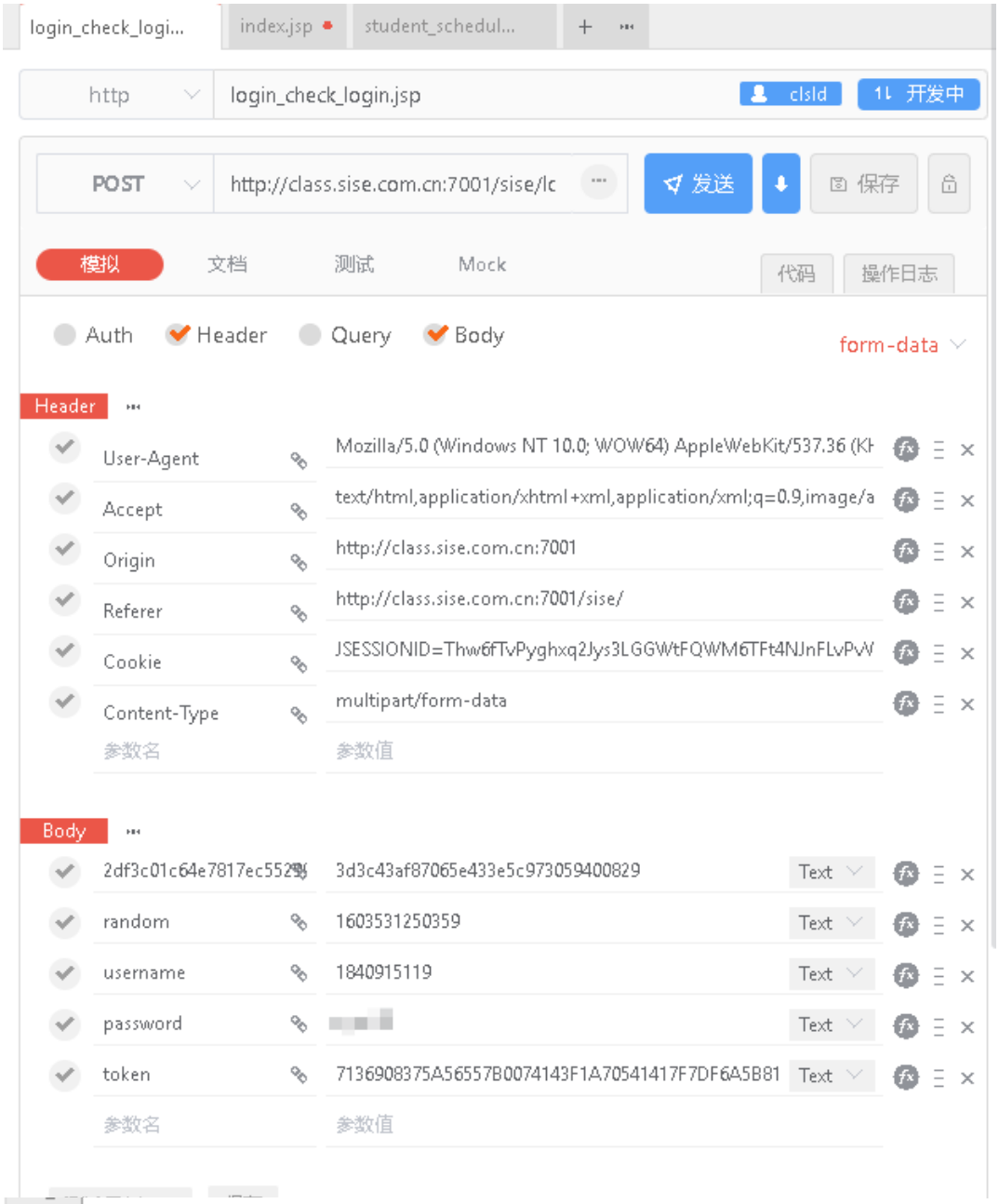
  Body

Name	Value
2df3c01c64e7817ec5529945fe5052e3	3d3c43af87065e433e5c973059400829
random	1603527440833
token	A11640A3553257B4D440C863433D5E67C96D278658028
username	1840915119
password	110120

- 最重要的是JSESSIONID，服务器通过用户第一次访问登录首页时服务端返回set-SESSIONID设置cookie，可以猜测当用户提交表单的时候，服务器通过提交的cookie作为i唯一d把数据存到服务器中，这个保存在浏览器的cookie会有自动清除的时间，保存在服务端的cookie也会在一段时间后删除。

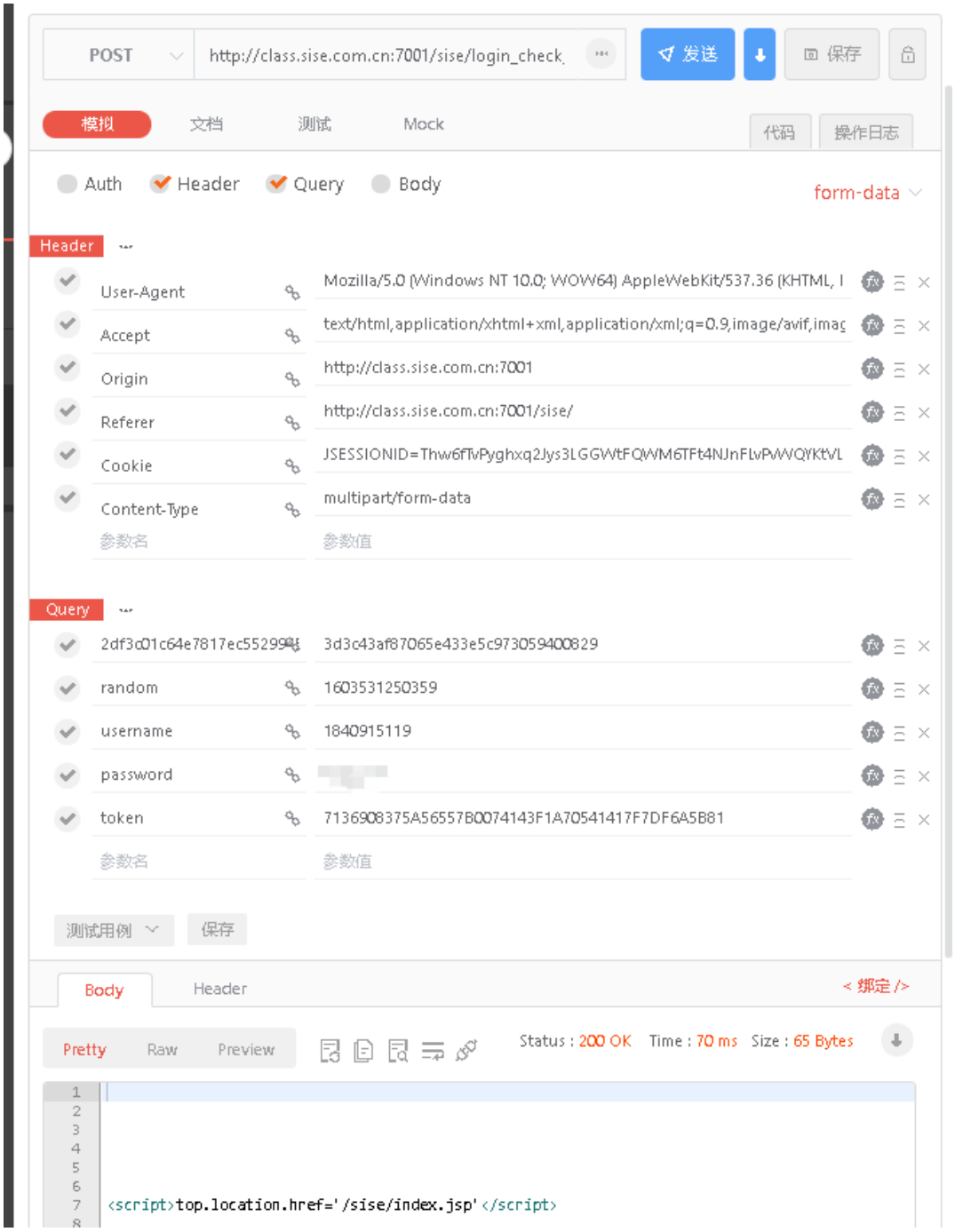


2. 在模拟发送请求软件中模拟发送请求看能不能得到第二步的验证成功



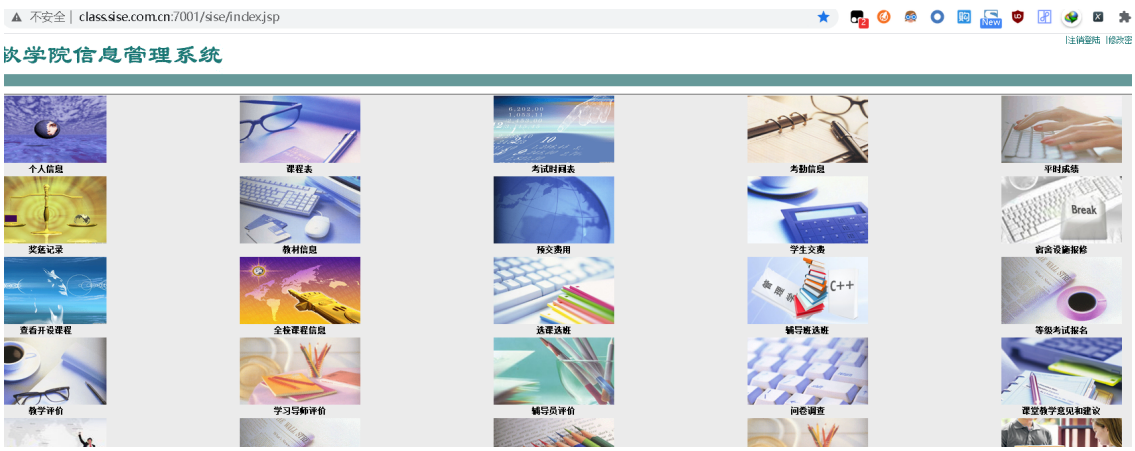

```
3
4
5
6
7  <script>top.location.href='/sise/index.jsp'</script>
8
9
```

- 从返回结果看，第二步验证失败，其实这里出现失败的原因是表单提交的数据不是放在body中，而是放在query中，把参数放在query后结果：



eeee,从放回结果看，说明是验证成功了。

3. index.jsp是这个页面：



三、华软学生课表页：http://class.sise.com.cn:7001/sise/module/student_schedular/student_schedular.jsp

1. 因为上一步已近验证登录了，服务器已近保留了我们的cookie对应的信息，即包括用户名和密码，所以现在直接访问这个网址，并且伪装好浏览器的标识符Accept和User-Agent，最重要的是上一步的cookie，服务器通过这个cookie验证我们是否是前面的登录用户。

通过返回可以看出我们已近获得了课程表的信息

```
title>华软学院信息管理系统</title>
link rel="stylesheet" href="/sise/css/style.css" type="text/css">
script language="javascript" src="/sise/js/componentdiv.js"></script>
style type="text/css">
style15 {color: #339999}
style16 {color: #006766}
style17 {font-size: 16px;color: #339999;}
/style>
d>
pt>
unction initialize(){
   var a = new xWin("1",120,300,800,150,"<font><b>作息时间表</b></font>","07

indow.onload = initialize;
ipt>
 bgcolor="#FFFFFF" text="#000000" topmargin="0" leftmargin="5" rightmargin='
 name="form1">
e width="95%" border="0" class="table1" cellspacing="1" align="center" heigh
>
 class="tablehead" height="32" align="center" valign="top">
div align="left"></div>
 <table width="100%" border="0" cellspacing="2"  cellpadding="0"  align="lef
     <tr>
         <td width="47"  class="td_right"> <div align="right">学 年: </div><
       <td width="63"  height="26" class="td_left"> <div align="left">
         <select name="schoolyear">
             <option value=2020>2020</option><option value=2019>2019</optior
           </select>
       </div></td>
         <td width="46"  height="26" class="td_right"> <div align="right">

       </div></td>
       <td width="96"  height="26" class="td_left"> <div align="left">
         <select name="semester" >

               <option value="1" selected>第一学期</option>
               <option value="2" >第二学期</option>

           </select>
       </div></td>
         <td width="301"  height="26" class="td_right">
           <div align="left">
             <input  type="submit" class="button" value=" 确  定 " onmouse
         </div></td>
         <td width="296" height="26" class="td_left">
         <div align="left"></div></td>
         <td width="82" >
     <div align="center"><span class="td_right">
     <input  type="button" class="button" id="act_stu_charge" value=" 返  回
     </span></div></td>
   </tr>
 </table>
d>

le>
e width="95%" border="0" cellspacing="0" cellpadding="0" align="center" heig
tr height="50">
<td>
    <div align="center" class="style15">
    <span class="style17">
```

四、python网络爬虫的实现

```
import requests
import re

url = 'http://class.sise.com.cn:7001/sise/'
toLogin_url = 'http://class.sise.com.cn:7001/sise/login_check_login.jsp'
JSESSIONID = ''
random = ""
post_key = 333
post_value = ''
username=''
```

```
password=''

def get_values():
    global JSESSIONID
    global random
    global post_key
    global post_value
    request = requests.get(url)
    html = request.content.decode('GBK')
    print(html)
    JSESSIONID = request.cookies.get('JSESSIONID')
    print("self.JSESSIONID:"+JSESSIONID)
    random = re.findall('<input id="random"   type="hidden"  value="(.*?)"
name="random" />',html,re.S)[0]
    values = re.findall('<input type="hidden"(.*?)>',html,re.S)[0]
    print("values:"+values)
    post_key = re.findall('name="(.*?)"',values,re.S)[0]
    print("self.post_key:"+post_key)
    post_value = re.findall('value="(.*?)"',values,re.S)[0]
    print("self.post_value:"+post_value)


def to_login():
    global JSESSIONID
    global random
    global post_key
    global post_value
    global username
    global password
    get_values()
    headers = {
        'Host': 'class.sise.com.cn:7001',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:76.0)
Gecko/20100101 Firefox/76.0',
        'Accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
        'Accept-Language': 'zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-
US;q=0.3,en;q=0.2',
        'Accept-Encoding': 'gzip, deflate',
        'Content-Type': 'application/x-www-form-urlencoded',
        'Content-Length': '172',
        'Origin': 'http://class.sise.com.cn:7001',
        'Connection': 'close',
        'Referer': 'http://class.sise.com.cn:7001/sise/',
        'Cookie': 'JSESSIONID='+JSESSIONID,
        'Upgrade-Insecure-Requests': '1',
    }
    data = {
        post_key:post_value,
        'random':random,
        'username':username,
        'password':password
    }
    print("post_key:"+data[post_key])
    print("random:" + data['random'])
    print("username:" + data['username'])
    print("password:" + data['password'])
```

```
    result =
requests.post(toLogin_url,headers=headers,data=data).content.decode('GBK')
    print("result:"+result)
    if result.find('<script>top.location.href=\'/sise/index.jsp\'</script>') ==
-1:
        return False
    else:
        return True
def student_schedular():
    global JSESSIONID
    headers = {
        'Host': 'class.sise.com.cn:7001',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:76.0)
Gecko/20100101 Firefox/76.0',
        'Accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
        'Accept-Language': 'zh-CN,zh;q=0.8,zh-TW;q=0.7,zh-HK;q=0.5,en-
US;q=0.3,en;q=0.2',
        'Accept-Encoding': 'gzip, deflate',
        'Connection': 'close',
        'Referer':
'http://class.sise.com.cn:7001/sise/module/student_states/student_select_class/ma
in.jsp',
        'Cookie': 'JSESSIONID='+JSESSIONID,
        'Upgrade-Insecure-Requests': '1',
    }
    url =
'http://class.sise.com.cn:7001/sise/module/student_schedular/student_schedular.js
p'
    student_class_html = requests.post(url,headers=headers).content.decode('GBK')
    print(student_class_html)
    MySchedular_dict = re.findall("class='font12'>(.*?)
</td>",student_class_html,re.S)
    print(MySchedular_dict)

if __name__ == '__main__':
        username = input("学号：")
        password = input("密码：")
        if not to_login():
           print('获取失败！！！')
        student_schedular()
```

1. 首先先通过getvalue()得到JSESSIONID和将浏览器信息保存到服务器，然后通过tologin()函数将jsession和form表单参数传给服务器，伪装成用户登录访问，服务器记录信息，最后通过student_schedular()将Jsessionid传给服务器，服务器返回课程表信息。

```
</table>


</form>
</body>
</html>

['1 - 2 节<br>09:00 - 10:20', '微信小程序应用开发(BTD 杨微 1 2 3 4 5 6 7 8 9 10 11 12 1

Process finished with exit code 0
```