爬取唯美女生照片（正则，bs4综合自我练习

笔记本： reptile_draft
创建时间： 2021/7/3 9:27 **更新时间：** 2021/7/3 15:45
作者： 134exetj717
URL： about:blank

- 爬取的过程中，出现错误了并不一定是程序出错了，而是有可能超时了，因为我们可能添加了timeout

```
picContent = requests.get(item[2],headers=headers,timeout = 1)
```

此时，我们可以删除timeout就可以正常执行了

```
# -*- codeing = utf-8 -*-
# @Time : 2021/6/13 19:23
# @Author : 希杰
# @file : reptile_wmgirls.py
# @software : PyCharm


import os
import re
import time
import urllib.request


import requests
from bs4 import BeautifulSoup


#寻找不同图片网站地址
findPicWeb = re.compile('<a href=(.*?.html) class=".*?".*?>\n(.*?) </a>')      #
其中.*?不包含\n符号，需要重点关注
#寻找图片
findPic = re.compile('<a href="(.*?)" alt=".*?" title=".*?"><img alt=".*?"
src=".*?" alt=".*?"/></a>')


headers = {
    'User-Agent': 'Mozilla / 5.0(WindowsNT10.0;  Win64;   x64;   rv: 89.0) Gecko
/ 20100101  Firefox / 89.0'
}
#获取网页
def askWeb(baseUrl):
    askObject = urllib.request.Request(url=baseUrl, headers=headers)
    response = urllib.request.urlopen(askObject)
    html = response.read().decode('utf-8')
    return html


#寻找不同图片网站
def seekPicWeb(baseUrl,html):
    url = re.findall(findPicWeb,str(html))
    picGroup = []
    for itemUrl in url:
        #time.sleep(0.1)
        askObject = urllib.request.Request(url=baseUrl +
itemUrl[0],headers=headers)
        response = urllib.request.urlopen(askObject)
```

```
        childHtml = response.read().decode('utf-8')
        picUrl = re.findall(findPic,childHtml)
        i = 1
        for pic in picUrl:
            picGGroup = []
            picGGroup.append('./' + itemUrl[1])
            picGGroup.append(str(i))
            picGGroup.append('https:' + pic)
            picGroup.append(picGGroup)
            i = i + 1
    print(picGroup)
    return picGroup


#保存图片
def savePic(picGroup):
    for item in picGroup:
        if not os.path.exists(item[0]):
            os.mkdir(item[0])
        picContent = requests.get(item[2],headers=headers,timeout = 1)
        file = open(item[0] + '/' + item[1] + '.jpeg','wb')
        file.write(picContent.content)
        file.close()
        # if len(item) != 0:
        #     print(item[0] + '/' + item[1] + '.jpeg')



if __name__ == '__main__':
    baseUrl = 'https://www.vmgirls.com/'
    html = askWeb(baseUrl)
    picGroup = seekPicWeb(baseUrl, html)
    savePic(picGroup)
```