链滴

python 爬虫简介

作者：dy12138

原文链接：https://ld246.com/article/1665757835905

来源网站：链滴

re模块的使用

- 在 Python中，我们可以使用内置的 re 模块来使用正则表达式

```
import re

"""
re.findall通过正则表达式筛选出文本中所有符合条件的数据
"""
# info = re.findall('python', 'hello this is python3.8 and python2.7')
# print(info)                    # ['python', 'python']
"""
re.finditer与re.findall作用一样，但是它的结果会被处理成迭代器对象，主要用于节省内存
"""
# info = re.finditer('h', 'hello this is python3.8 and python 2.7')
# print(info)                    # <callable_iterator object at 0x0000027AF3835C70>
"""
re.search通过正则表达式，只要匹配到一个符合条件的数据就结束
"""
# info = re.search('h', 'hello this is python3.8 and python 2.7')
# print(info)                    # <re.Match object; span=(0, 1), match='h'>
# print(info.group())              # h

"""
re.match通过正则表达式从头开始匹配，如果开头不符合条件，那后面就不在继续操作了
"""
# info = re.match('e', 'hello this is python3.8 and python 2.7')
# print(info)                    # None

"""
re.compile能够提前准备好正则，之后可以反复进行使用，极大的减少了代码的冗余
"""
# info = re.compile('h')
# print(re.findall(info, 'history not is forget '))     # ['h']
# print(re.findall(info, 'hhheiheiheihei'))            # ['h', 'h', 'h', 'h', 'h', 'h']
```

re模块之分组

```
"""
re模块之分组
"""

# info = re.findall('python', 'python3.8python2.7python3.6')
# print(info)              # ['python', 'python', 'python']

# info = re.findall('pyt(h)on', 'python3.8python2.7python3.6')
# print(info)               # ['h', 'h', 'h']
"""
re.findall针对分组正则表达式匹配到的结果会优先展示，同样也可以取消分组
"""
# info = re.findall('pyt(?:h)on', 'python3.8python2.7python3.6')
# print(info)               # ['python', 'python', 'python']
```

```
"""
re.search针对分组正则表达式匹配到的结果不做任何展示，可以通过索引的方式获取分组的结果，且引的数据值必须要有对应的分组，可以自己进行分组，否则执行会报错
"""
# info = re.search('pyt(h)on', 'python3.8python2.7python3.6')
# print(info)                 # <re.Match object; span=(0, 6), match='python'>
# print(info.group())         # python
# print(info.group(0))        # python
# print(info.group(1))        # h
```

re模块之别名

```
"""
re模块之别名，针对分组正则表达式，除了使用索引获取分组的内容，还可以给分组添加别名来获取组内容，如果正则表达式过长的话，就可以给它起个别名，这样便于筛选查找
"""
# info = re.search('pyt(?P<name>.*?)on', 'python3.8python2.7python3.6')
# print(info.group())               # python
# print(info.group('name'))         # h
```

爬虫简介

什么是互联网?

● 互联网（ internet），又称国际网络，指的是网络与网络之间所串连成的庞大网络，这些网络以组通用的协议相连，形成逻辑上的单一巨大国际网络。

● 互联网始于1969年 美国的阿帕网。通常internet泛指互联网，而**Internet**则特指因特网。这种将算机网络互相联接在一起的方法可称作“网络互联”，在这基础上发展出覆盖全世界的全球性互联网称互联网，即是互相连接一起的网络结构。互联网并不等同万维网，万维网只是一建基于超文本相互接而成的全球性系统，且是互联网所能提供的服务其中之一。

互联网的作用?

● 互联网可以用来传播信息、进行电子商务、进行网络交流等多种用途，也可以实现计算机之间的资共享

网络的本质?

● 计算机网络是指将地理位置不同的具有独立功能的多台计算机及其外部设备，通过通信线路连接起，在网络操作系统，网络管理软件及网络通信协议的管理和协调下，实现资源共享和信息传递的计算系统。

网络爬虫的本质

● 网络爬虫（webcrawler）又称为网络蜘蛛（webspider）或网络机器人（webrobot），另外一些常使用的名字还有蚂蚁、自动索引、模拟程序或蠕虫，同时它也是“物联网”概念的核心之一。网络虫本质上是一段计算机程序或脚本，其按照一定的逻辑和算法规则自动地抓取和下载万维网的网页，搜索引擎的一个重要组成部分。

爬虫实战

```
import re

import requests

# 获取红牛全国分公司的信息，eg:地址、电话、邮箱等

# 1.向目标地址发送网络请求获取相应的数据(相当于在浏览器地址栏中输入地址)

get_info = requests.get('http://www.redbull.com.cn/about/branch')
# print(get_info.content)        # <Response [200]>     获取页面数据，数据类型为bytes
# print(get_info.text)           # 获取解码后的数据
"""
 为了避免每次执行程序都要发送网络请求  也可以提前保存页面数据到文件
"""
with open(r'rednull.html', 'wb') as f:
    f.write(get_info.content)

# 2.读取页面数据

with open(r'rednull.html', 'r', encoding='utf8') as f:
    data = f.read()

# 3.研究目标数据的特征，使用正则筛选数据
# 4.获取所有分公司的名称
company_name_list = re.findall('<h2>(.*?)</h2>', data)
# print(company_name_list)
# 5.获取所有分公司的地址
company_address_list = re.findall("<p class='mapIco'>(.*?)</p>", data)
# print(company_address_list)
# 6.获取所有分公司的邮箱地址
company_email_list = re.findall("<p class='mailIco'>(.*?)</p>", data)
# print(company_email_list)
# 7.获取所有分公司的电话
company_phone_list = re.findall("<p class='telIco'>(.*?)</p>", data)
# print(company_phone_list)
# 8.将上述列表中的数据按照对应的位置进行整合
info = zip(company_name_list, company_address_list, company_email_list, company_phone_lis
)
for i in info:
    print("""
    公司名称:%s
    公司地址:%s
    公司邮箱:%s
    公司电话:%s
    """ % i)
```

openpyxl模块

主要用于操作excel表格 也是pandas底层操作表格的模块

在python中能够操作excel表格的模块有很多

openpyxl属于近几年比较流行的模块
openpyxl针对03版本之前的excel文件兼容性不好
xlwt、xlrd也可以操作excel表格
兼容所有版本的excel文件 但是使用方式没有openpyxl简单

1.excel版本问题
    03版本之前 excel文件的后缀名 .xls
     03版本之后 excel文件的后缀名 .xlsx
    如果是苹果电脑excel文件的后缀 .csv
2.下载模块
    pip3.8 install openpyxl

openpyxl实战

```
# 1.创建Excel表格
# 2.导入模块
import requests
import openpyxl

from openpyxl import Workbook

wb = Workbook()
# 3.创建Excel文件
wb1 = wb.create_sheet('生死簿', 0)
wb2 = wb.create_sheet('琅琊榜', 1)
wb3 = wb.create_sheet('上海富婆名单表', 2)
wb1.title = '红牛全国分公司信息表'
wb2.title = '红浪漫贵宾名单'  # 支持二次修改
wb2.sheet_properties.tabColor = '1072BA'

# 写入文件数据

"""
第一种写入方式，直接写入
"""
wb2['A1'] = 'jason'
wb2['B1'] = 'jason专属技师'

"""
第二种写入方式
"""

wb2.cell(row=3, column=2, value='张三')

"""
第三种写入方式
"""
wb1.append(['company_name_list', 'company_address_list', 'company_email_list', 'company_p
one_list'])
wb1.append(['红牛杭州分公司', '杭州市上城区庆春路29号远洋大厦11楼A座', '310009', '0571-8704
279/7792'])


# 3.保存文件
```

```
wb.save('爬取红牛信息.xlsx')
```

第三方模块的下载

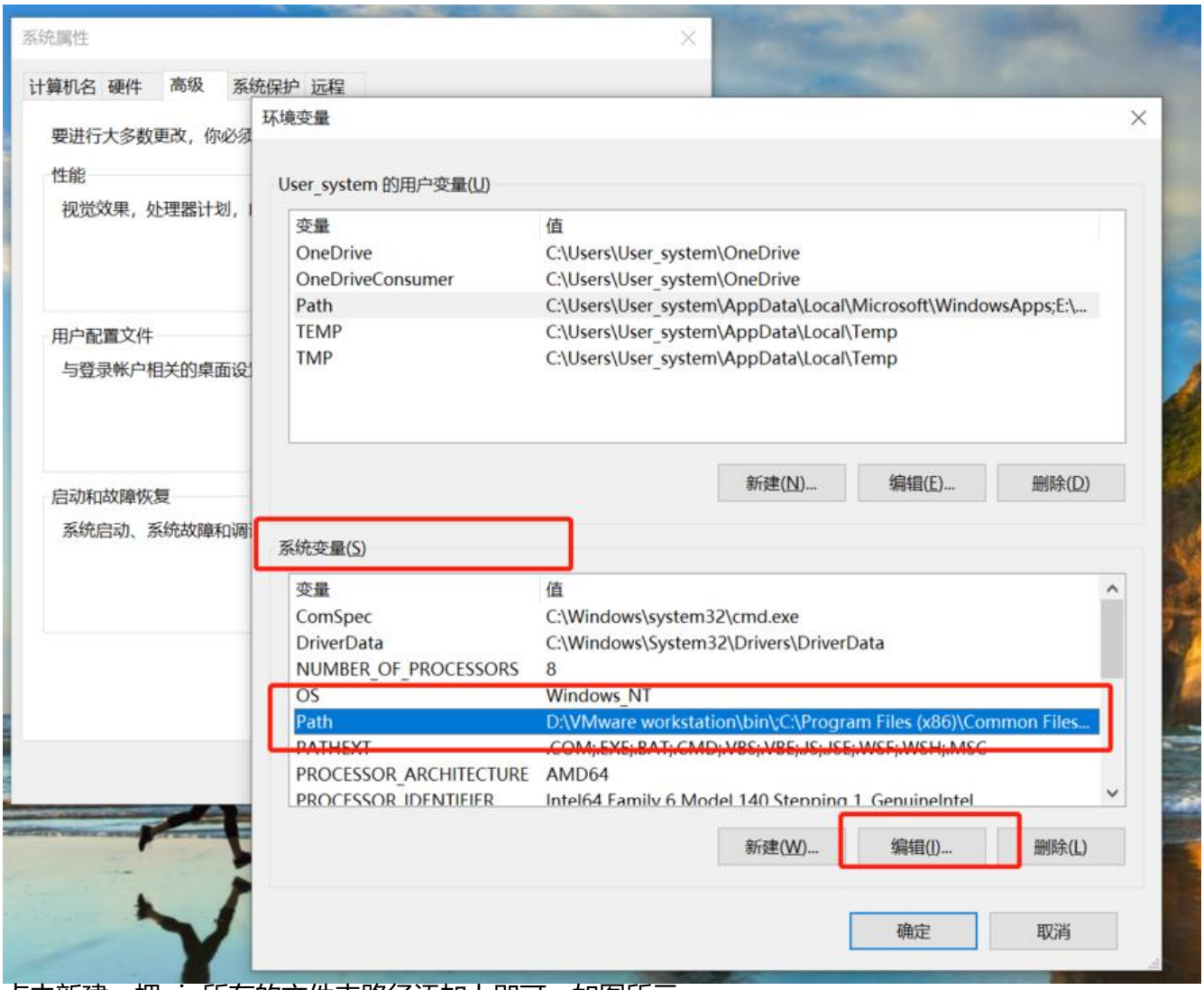
```
"""
1.下载速度很慢
    pip工具默认是从国外的仓库地址下载模块 速度很慢
    我们可以切换下载的地址(源地址)
        清华大学：https://pypi.tuna.tsinghua.edu.cn/simple/
        阿里云：http://mirrors.aliyun.com/pypi/simple/
        中国科学技术大学：http://pypi.mirrors.ustc.edu.cn/simple/
        华中科技大学：http://pypi.hustunique.com/
        豆瓣源：http://pypi.douban.com/simple/
        腾讯源：http://mirrors.cloud.tencent.com/pypi/simple
        华为镜像源：https://repo.huaweicloud.com/repository/pypi/simple/
    pip3.8 install 模块名 -i 源地址

    pycharm提供第三方模块下载快捷方式

    也可以直接修改python解释器源文件(课下自行查阅)
2.下载报错
    1.pip工具版本过低 直接拷贝提示信息里面的更新命令即可
        python38 -m pip install --upgrade pip
    2.网络波动 关键字是Read timed out
        只需要重新下载几次即可 或者切换一个网络稳定一点的
    3.有些模块在下载使用之前需要提前配置指定的环境
        结合具体情况 百度搜索

3.模块也有版本
    pip3.8 install 模块名==版本号
        pip3.8 install django==1.11.11
"""
```

- 第三方模块必须先进行下载才可以导入使用
- python下载第三方模块需要借助pip工具
- pip工具需要添加到环境变量才可以使用

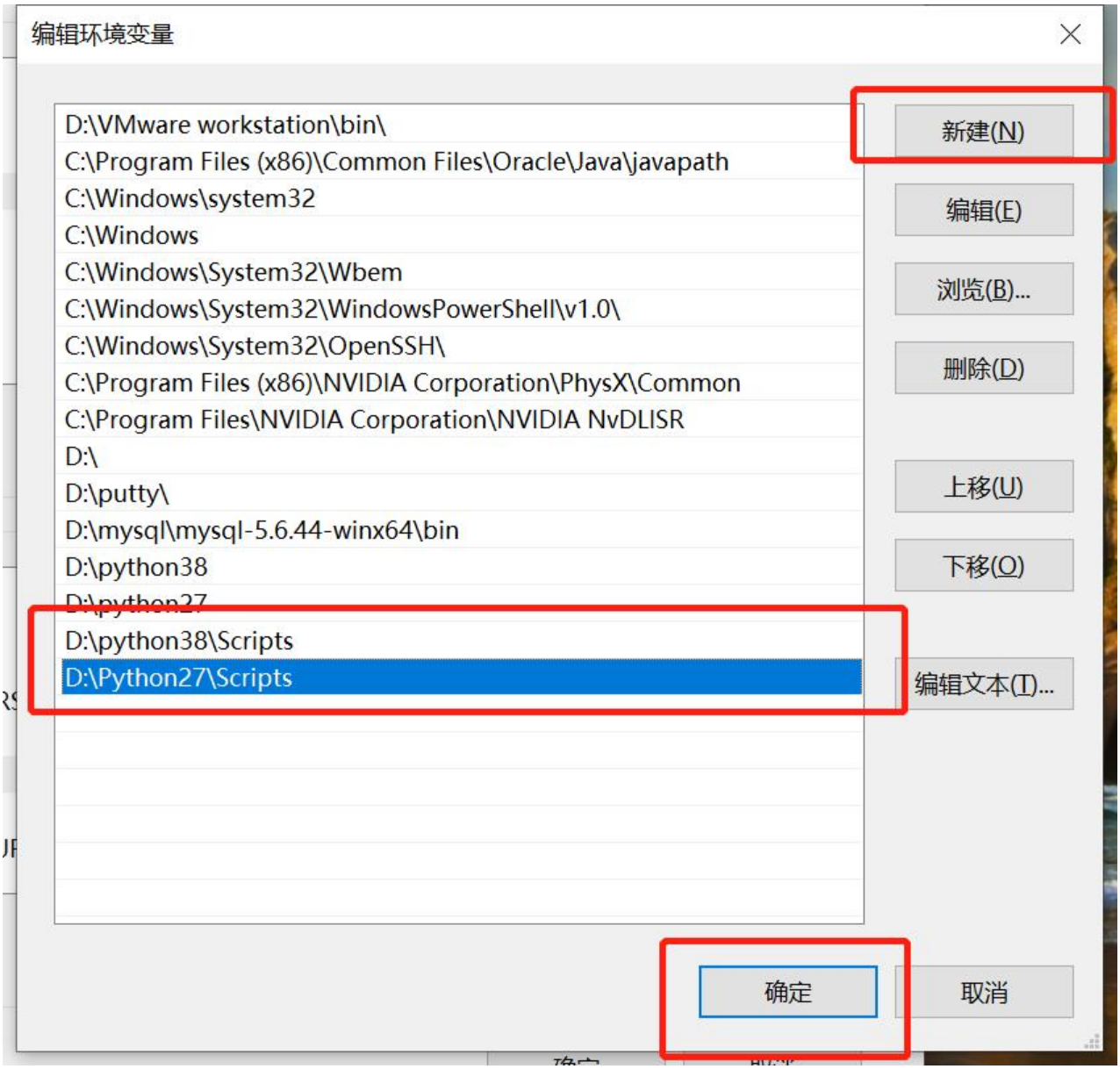
添加环境变量的步骤如下：在windows操作系统中可以通过鼠标右键依次点击我的电脑-〉系统属性-高级系统设置-〉环境变量，来设置系统的环境变量，如下图：

原文链接：python 爬虫简介


系统属性
计算机名
硬件
高级
系统保护
远程
要进行大多数更改，你必须作为管理员登录。
性能
视觉效果，处理器计划，内存使用，以及虚拟内存
设置(S)...
用户配置文件
与登录帐户相关的桌面设置
设置(E)...
启动和故障恢复
系统启动、系统故障和调试信息
设置(T)...
环境变量(N)...
确定
取消
应用(A)



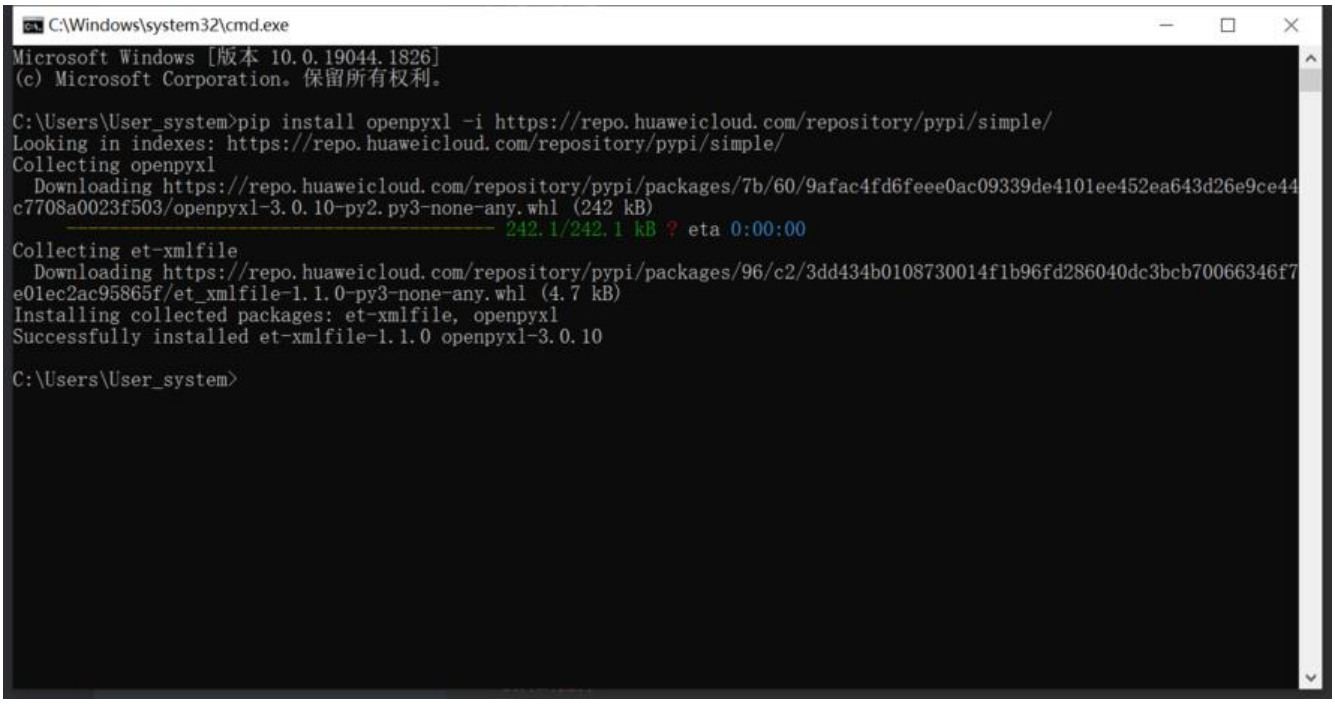
点击新建，把pip所在的文件夹路径添加上即可，如图所示：



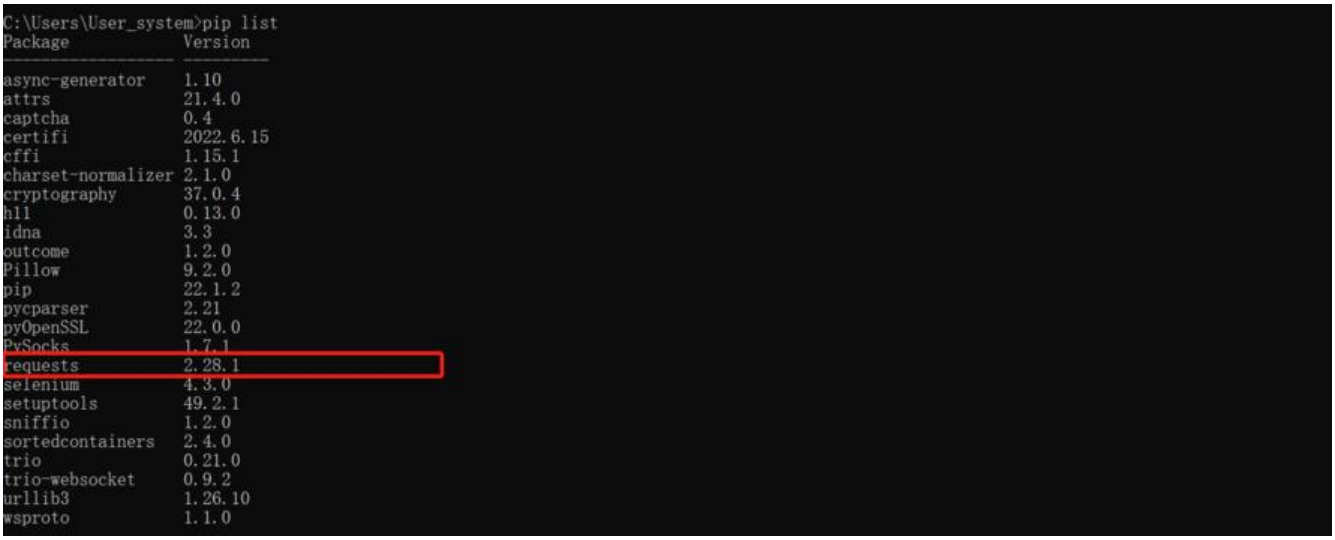
第三方模块下载的具体步骤：

使用win+R输入cmd，使用命令行安装，如下图：

```
C:\Users\User_system>pip install requests
Collecting requests
  Downloading requests-2.28.1-py3-none-any.whl (62 kB)
     ---------------------------------------- 62.8/62.8 kB 4.8 kB/s eta 0:00:00
Requirement already satisfied: idna<4,>=2.5 in d:\python38\lib\site-packages (from requests) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in d:\python38\lib\site-packages (from requests) (2022.6.15)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in d:\python38\lib\site-packages (from requests) (1.26.10)
Collecting charset-normalizer<3,>=2
  Downloading charset_normalizer-2.1.0-py3-none-any.whl (39 kB)
Installing collected packages: charset-normalizer, requests
Successfully installed charset-normalizer-2.1.0 requests-2.28.1
```

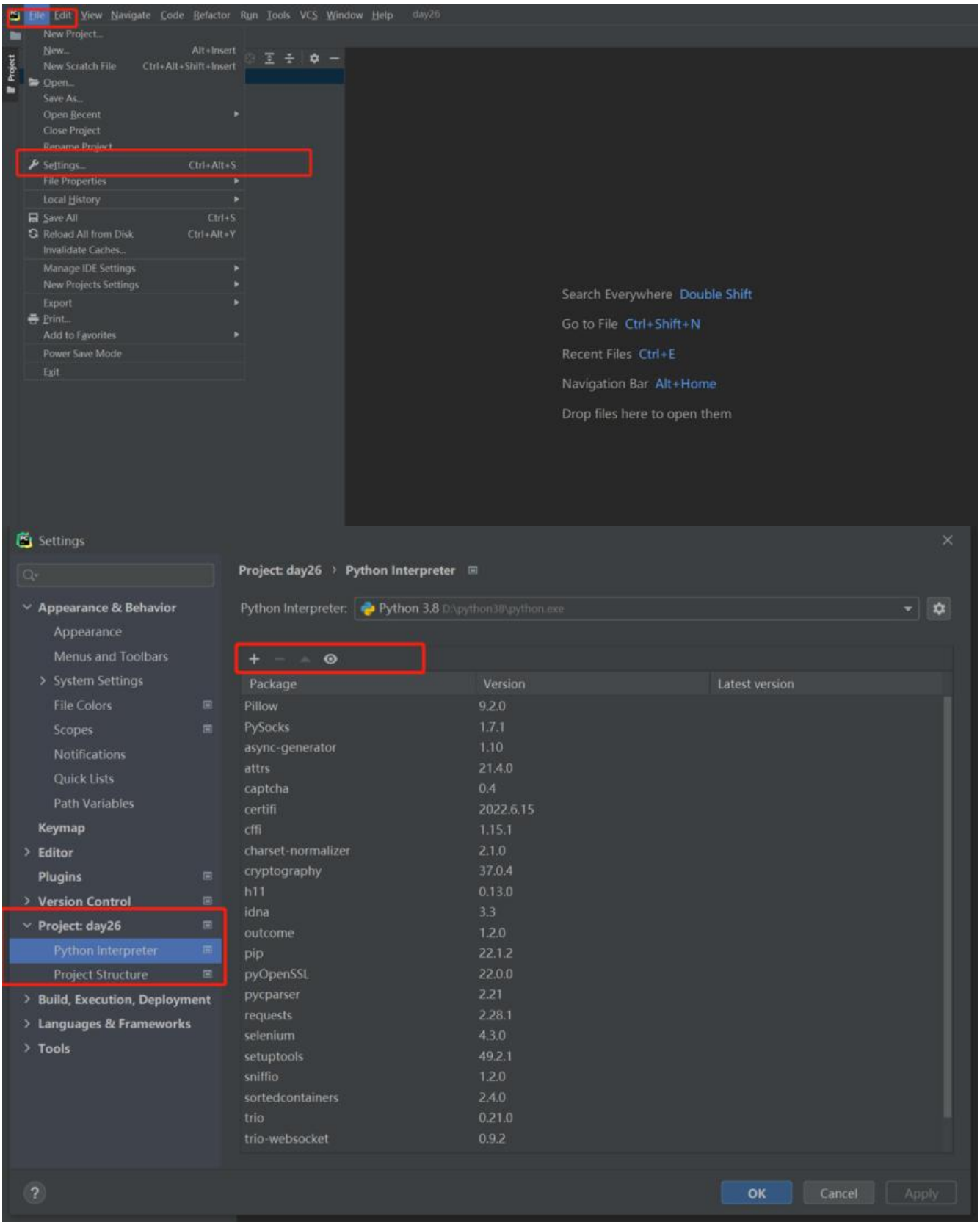


查看下载的版本信息

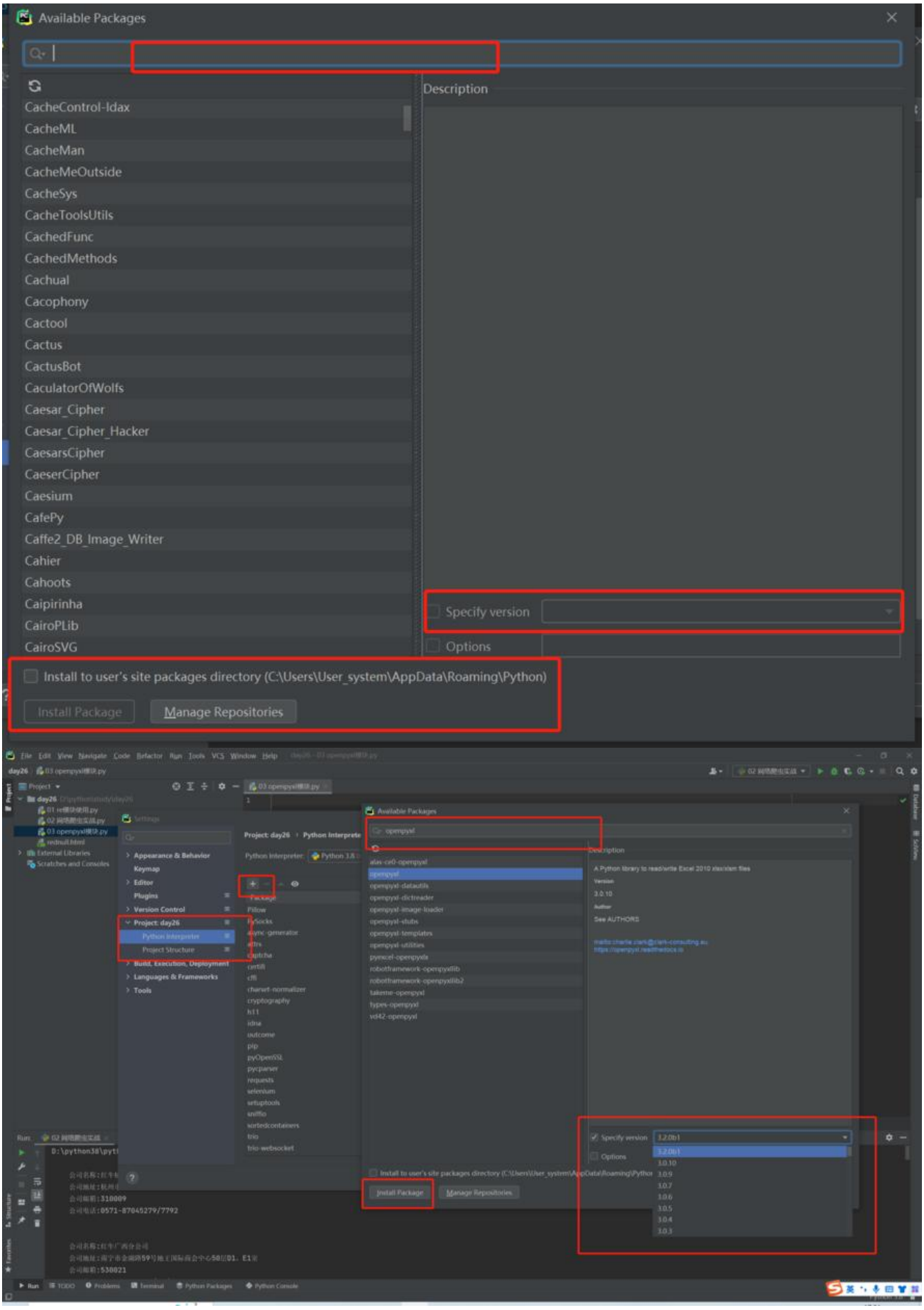


也可以使用pycharm下载，步骤如下：

选择File——setting——打开设置，依次进行如下步骤即可：



原文链接：python 爬虫简介



原文链接：python 爬虫简介

课题演练

```
import re
import openpyxl
import requests

# 获取红牛全国分公司的信息，eg:地址、电话、邮箱等

# 1.向目标地址发送网络请求获取相应的数据(相当于在浏览器地址栏中输入地址)

baseurl = 'http://www.redbull.com.cn/about/branch'

# 2.创建Excel表格并写入数据
from openpyxl import Workbook
wb = openpyxl.Workbook()
ws = wb.active
ws.append(['company_name_list', 'company_address_list', 'company_email_list', 'company_ph
ne_list'])
# 3.请求头
headers = {
    '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',
    'Connection': 'keep-alive',
    'User-Agent': 'Mozilla/5.0 (X11; linux x86_64; rv:60.0) Gecko/20100101 firefox/60.0',
    'upgrade-insecure-requests': '1'
}
content = requests.get(baseurl, headers=headers).content.decode('utf-8')
# 4.研究目标数据的特征，使用正则筛选数据
# 5.获取所有分公司的名称
company_name_list = re.findall('<h2>(.*?)</h2>', content)
# print(company_name_list)
# 6.获取所有分公司的地址
company_address_list = re.findall("<p class='mapIco'>(.*?)</p>", content)
# print(company_address_list)
# 7.获取所有分公司的邮箱地址
company_email_list = re.findall("<p class='mailIco'>(.*?)</p>", content)
# print(company_email_list)
# 8.获取所有分公司的电话
company_phone_list = re.findall("<p class='telIco'>(.*?)</p>", content)
# print(company_phone_list)
for i in range(len(company_name_list)):
    ws.append([company_name_list[i], company_address_list[i], company_email_list[i], company
phone_list[i]])
    wb.save('红牛全国分公司信息.xlsx')
```