

计算机网络基础及数据获取

【实验目的】

(1) 了解 Internet 协议的基本含义,掌握计算机的 IP 地址设置方法。

(2)掌握页面免费邮箱的申请过程,掌握如何修改电子邮箱的个人信息,掌握电子邮件的 各项基本功能。

(3) 了解邮件的基本功能,掌握如何使用邮件完成多邮件的收发功能。

(4) 掌握 Python 网络爬虫 requests 库的使用。

(5) 掌握正则表达式的数据抽取方法。

(6) 掌握 Python 写入 Excel 的方法。

【实验环境】

中文 Windows, Outlook 2016, Python 3。

【实验内容】

(1)分别设置家中三台计算机的 IP 地址,使得这三台计算机能正常连接互联网。

- (2) 免费电子邮箱的注册。
- (3) 注册后个人信息的修改。
- (4) 电子邮箱的基本功能使用。

(5)邮件收发软件的使用。注册了好几个页面邮箱,但是每次都登录邮箱页面收发邮件 很麻烦,可以通过设置不用打开网站就接收所需要的邮件。

- (6) requests 库的安装。
- (7) 用 requests 库获取数据。
- (8) 用正则表达式抽取数据。

<u>______ 实验 3.1 IP 地址设置</u>

IP 地址是指 Internet 协议使用的地址。每个 Internet 服务提供商(ISP)须向有关组织申请一组 IP 地址,一般是将该组 IP 地址动态分配给其用户,当然用户也可以向 ISP 申请一个 IP 地址,这就是为什么在配置 Windows 10 的"网络"时,一般让系统自动分配 IP 地址,也可以手工设置。下面讲解设置 IP 地址的过程。



【实验要求】

- (1) Internet 协议基础知识。
- (2) 自动获取 IP 地址。
- (3) 手工设置 IP 地址。

【实验步骤】

(1)打开资源管理器,右击"网络",在弹出的快捷菜单中选择"属性"命令,如图 3-1 所示, 进入"网络和共享中心"窗口,如图 3-2 所示。

💣 🖸 📗 🖬 🖬	– 🗆 X	展开(A)
文件 网络 查看 ← → ↑ ● →	5 v	在新窗口中打开(E) 固定到快速访问
★ 快速访问	此文件夹为空。	固定到"开始"屏幕(P)
		映射网络驱动器(N) 断开网络驱动器的连接(C)
● 此电脑		删除(D)
0 个项目		属性(R)

图 3-1 右击"网络"再选择"属性"命令

暨 网络和共享中心		– 🗆 ×	<
← → · · ↑ 壁 « 网络和 Inte	* > 网络和共享中心 > ひ	搜索控制面板 。	•
文件(F) 编辑(E) 查看(V) 工具(D		
控制面板主页	查看基本网络信息并设置连	接	^
	查看活动网络		I
更改适配器设置 更改高级共享设置 媒体流式处理选项	HUAWEI-amy1 专用网络	访问类型: Internet 连接: 創 WLAN (HUAWEI-amy1)	
另请参阅	更改网络设置		I
Internet 选项	💼 设置新的连接或网络		1
Windows Defender 防火墙	设置宽带、拨号或 VPN 过	崔接; 或设置路由器或接入点。	,

图 3-2 "网络和共享中心"窗口

(2) 在图 3-2 所示窗口的左侧单击"更改适配器设置",进入"网络连接"窗口,如图 3-3 所示,在该窗口中右击需要配置的网卡,在弹出的菜单中选择"属性"命令,弹出本地连接属性对话框,如图 3-4 所示,勾选"Internet 协议版本 4(TCP/IPv4)"复选框,单击"属性"按钮,进入 IP 地址的设置界面,如图 3-5 所示。



图 3-3 "网络连接"窗口

WLAN METE	×	Internet	t 协议版本 4 (TCP/IPv4) 屬	ŧ	
络 共享		常规	备用配置		
连接时使用:		1000			
🖤 Intel(R) Wireless-N 7265		如果) 緒系的	网络支持此功能,则可以获用 克管理员处获得适当的 IP 设	权自动指派的 IP 设置。否则,《 置。	弥震要从网
	配置(C)				
此连接使用下列项目(0):		۲	自动获得 IP 地址(O)		
☑ ■Microsoft 网络赛户端	^	0	使用下面的 IP 地址(S):		
☑ 聖 Microsoft 网络的文件和打印机共享 ☑ 哩 Oos 数据和计划程序		IP	地址①:	• • •	
☑ 및 Intel(R) Technology Access Filter Driver		子	网播码(U):	a. a. a	
☑ ▲ Internet 协议版本 4 (TCP/IPv4) □ ▲ Microsoft 网络适配器多路传送器协议		10	认网关(D):	· · · ·	
□ Microsoft LLDP 协议驱动程序 ☑ * Internet 协议版本 6 (TCP/IPv6)	~	۲	自动获得 DNS 服务器地址(B)	
<	>	-0	使用下面的 DNS 服务器地均	£(E):	
安装(<u>N</u>) 卸 <u>载(U</u>)	届性(图)	20	违 DNS 服务器(的):	a (a) a	
描述			用 DNS 影響器(A)+		
传输控制协议/Internet 协议。该协议是默认的广 于在不同的相互连接的网络上通信。	域网络协议,用				
			退出时验证设置(L)		高级(⊻)
80	RISK	5		18:00	Pota

图 3-4 本地连接属性对话框

图 3-5 进入 IP 地址的设置界面

IP 地址的获取有两种方式:自动获取方式和手工设置方式。

如果想自动获取 IP 地址,也就是由系统自动分配 IP 地址,可以选择图 3-5 中的"自动获得 IP 地址"单选按钮。这种方式是系统根据当前连接网络段,自动分配对应的 IP 地址号码。 采用这种自动获取的方式,IP 地址的分配带有一定的随机性。

IP 地址的随机自动获得方式是目前最主要的 IP 地址分配方式。这种方式为用户上网提供了极大的便利,使得网络配置不再困扰用户。但从管理的角度来看,手工设置固定 IP 地址的方式更有利于管理。选择图 3-5 中的"使用下面的 IP 地址"单选按钮,可以自行手工设置 IP 地址以及网关和 DNS 服务器。如图 3-6 所示,分别设置 IP 地址、子网掩码、默认网关、首选 DNS 服务器和备用 DNS 服务器等内容。

	臣
规	
如果网络支持此功能,则可以获明 络系统管理员处获得适当的 IP 设	取自动指派的 IP 设置。否则,你需要从网 2置。
○自动获得 IP 地址(Q)	
④使用下面的 IP 地址(S):	
IP 地址(I):	192 . 168 . 202 . 12
子网掩码(U):	255.255.255.0
默认网关(D):	192.168.202.1
○ 自动获得 DNS 服务器地址	(B)
●使用下面的 DNS 服务器地	址(E):
首选 DNS 服务器(P):	202 . 114 . 232 . 1
备用 DNS 服务器(A):	· · ·

图 3-6 手工设置 IP 地址以及网关和 DNS 服务器

大数据分析导论实验指导与习题集

DNS 服务器用于网络域名地址(也就是网址)与 IP 地址之间的转换,用户可以选择自动获取 DNS 服务器。

シンシ 实验 3.2 注册免费邮箱

电子邮件是最早的网络应用之一,也是使用最为广泛的。电子邮件最早作为法律电子证据。它突破了时空的限制,大大地提高了办公的效率,为办公自动化、商业活动、公务活动等提供了极大的便利。

【实验要求】

掌握注册免费的 Web 页面邮箱。

【实验步骤】

(1) 打开 IE 浏览器,在地址栏中输入 http://www.163.com,进入网易首页,如图 3-7 所示。



图 3-7 网易首页

(2) 单击"注册免费邮箱"按钮,进入网易首页的注册免费邮箱界面,如图 3-8 所示。

邮件地址	建议用手机号码;	主册	@	163.com
	6~18个字符,可使用	用字母、数字、下引	则线,	雷以字母开头
*密码				
	6~16个字符,区分;	大小写		
确认密码				
	请再次填写密码			
千切里雨				
丁加与的	-+80-			
-1-01-2-0-)	忘记密码时,可以通)	过该手机号码快速	找回题	鋼
* 16:07:03	■ * +80- 忘记密码时,可以通う	过该手机号码快速		Non
+105円 *捡证码	************************************	过该手机号码快速		NOR
*验证码	 本30- 忘记密码时,可以通 请填写图片中的字符 	过该手机号码快速 ,不区分大小写	找回: 有不	
*验证码	() () () () () () () () () () () () () (过该手机号码快速 ,不区分大小写 码	找回語 着不	四 如 有是? 排张图片
+ 验证码 至信验证码	1 +86- 忘记密码时,可以通 请填写图片中的字符 免费获取验证	过该手机号码快速 ,不区分大小写 码	北回語	码 1990年 清楚? 操乐图片
 ナロバラ時 * 验证码 直信验证码 	 (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)	动族手机号码快速: ,不区分大小写 码 真写短信中的验证	找回话 看不	四 9 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

图 3-8 网易首页的注册免费邮箱界面

第3章 计算机网络基础及数据获取

29

(3) 在图 3-8 中,有三种邮箱可供选择,其中"注册 VIP 邮箱"是需要付费的。这里选择 的是"注册字母邮箱"。在图 3-8 中,带"*"的都是必填项目,输入要注册的邮件地址(也即 邮箱地址)、密码、确认密码、手机号码(手机号码是为了收取短信验证码)、验证码以及短信 验证码,单击"立即注册"按钮,即可完成注册流程。注册成功即可弹出注册成功的提示 窗口。

(4) 如果该邮箱与其他邮箱关联(将现有的邮箱与其他邮箱建立关联,则无须重复登录, 即可查看相应的收件信息)。

(5) 注册完成后,输入要登录的邮箱地址,如图 3-9 所示,输入密码,单击"登录"按钮,进 入我的邮箱,如图 3-10 所示。网易邮箱主界面如图 3-11 所示。单击"设置"按钮,进入邮箱设 置界面。"设置"菜单下的子菜单如图 3-12 所示,包括"常规设置""邮箱密码修改""账号与邮 箱中心""邮箱安全设置""手机服务""POP3/SMTP/IMAP"和"换肤"。

登录	注册免费邮箱、	- 网易考拉
dxj	sjyyjch@163.com	ø
睛	输入密码	é
	登 录	
	十天内免登录	忘记密码?

darj	sjyyj	ch91		1
进)	、用户	中心		
进)	大我的	的邮箱	ĩ	
进)	、我的	LOFT	ER	
进)	我的	相册		

图 3-9 输入要登录的邮箱地址和密码 图 3-10 进入我的邮箱



图 3-11 网易邮箱主界面

受置	~	報助~	退出
Ф	常規	设置	
	邮箱	喀码修改	
	账号	与邮箱中	сiv
	邮箱	安全设置	
	手机	服务	
	POP	P3/SMTP/	IMAP
¥	接肤	ŧ	

图 3-12 "设置"菜单下的子菜单

"常规设置"包含的设置功能包括基本设置、自动回复/转发、发送邮件后设置、邮件撤回、 写信设置、读信设置和其他设置。这些设置可以提高邮箱的管理。

"邮箱密码修改"可以修改邮箱的密码和密保问题(密保问题是在密码丢失的情况下通过 回答问题找回密码的一种方式)。

"账号与邮箱中心"的功能如图 3-13 所示。"多账户关联"是指在拥有多个网易邮箱的情况下,设置了多账户关联以后,只需要登录任一账号,即可登录关联的多个网易邮箱账户,方便用户使用。"邮箱中心"更是提供了对其他非网易邮箱的邮件的管理和迁移。"发件人管理"可以设置多个发件人,用户可根据需要确定发件人,如在发件时未指定,则自动填写的是默认发件人的信息。

多账户关联 无须登录,一键切换	其他网易邮箱账号	
账号	未该部件	
您还没有关联其他邮箱账户? 立	即关联	
邮箱中心 通过POP3自动收取,管	管理其他邮箱的邮件,20秒实现QQ/G	mail邮件迁移
账号 收取文件夹	未读/总邮件数	通讯录
您还没有其他邮箱? 立即创建 发件人管理 他用其他邮箱账号发	送邮件	
账号	代发方式	状态
dxjsjyyjch <dxjsjyyjch@163.co< td=""><td>网易代发</td><td>默认发件人</td></dxjsjyyjch@163.co<>	网易代发	默认发件人
+ 运加发性人		

图 3-13 "账号与邮箱中心"的功能

"邮箱安全设置"如图 3-14 所示。

"手机服务"如图 3-15 所示。这是基于现在社会智能手机应用的需求而提供的功能。

POP3/SMTP/IMAP 设置如图 3-16 所示。POP3/SMTP/IMAP 是邮件传输的常用协议,当我们使用第三方的邮件管理客户端,登录对应的邮箱账号时,需要开启 POP3/SMTP/IMAP 协议才可以登录,否则就会显示权限不足。

"换肤"菜单是对邮件背景颜色的管理。用户可根据个人喜好,选择设置不同风格的邮箱 皮肤。"换肤"菜单提供"动态皮肤""主题皮肤"和"基础皮肤"三种风格供用户选择。

31

登录安全	
登录二次验证	为提高安全性,除了要输入用户名和密码,还需短信动态验证码才能 登录使用邮箱。
人脸识别	人脸认证登录,加强您的邮箱账号安全。
隐私安全	
安全锁	给重要的资料加安全锁,让您的邮件信息资料更加安全。
安全提醒	
客户端删信提醒	当邮件客户课删除邮件时,系统会通过邮件和短信发送提醒信息。
自动转发提醒	在设置自动转发时会进行短信提醒(需绑定手机号码),以提高邮箱 安全性。

图 3-14 邮箱安全设置

6	援作短信通知 设置自动转发/POP3删信后有短信通知	0	已开通 验证手机后自动获得	默认开启
您	还可以开通以下手机服务			
	手机号码邮箱 以手机号码作为用户名的邮箱账号	•	未激活 开通后可使用免费通知短信	与上开通
6	邮箱登录二次验证 整录邮箱时需要通过短信验证,账号221	•	未开通 开通后可使用邮箱二次验证	马上开通
۵	安全锁 阿盘、文件夹等双重密码保护	0	未开通 加訫范围:无	马上开通
1	陸身邮(收费) 收发邮件都有短信通知	•	未开通 剩余条数:0条	马上开通

图 3-15 手机服务

POP3/SMTP/IMAP	
设置POP3/SMTP/IMAP:	 POP3/SMTP服务 IMAP/SMTP服务 水取最近30天邮件 > 温馨爆示:请使用授权码登录第三方邮件客户编
设置POP3/SMTP/IMAP:	☑ 开启客户读删除邮件提醒 当邮件客户读删除邮件时,系统会通过邮件发送提醒信息
	健存 取満
提示	
服务器地址:	POP3服务器: pop.163.com SMTP服务器: smtp.163.com IMAP服务器: imap.163.com
网易官方部件客户端:	网易邮箱大师
其他邮件客户读:	PC满设置帮助
	移动端设置帮助(iOS、Android、Symbian)

图 3-16 POP3/SMTP/IMAP 设置

______实验 3.3 使用 Windows 10 的"邮件"功能收发电子邮件

【实验要求】

掌握基于 Windows 10 操作系统自带的一个客户端软件,用于处理邮件收发。

【实验步骤】

(1)单击"开始"菜单,在弹出的程序列表中单击"邮件"按钮,如图 3-17 所示。进入"邮件"界面,如图 3-18 所示。



图 3-17 单击"邮件"按钮



图 3-18 "邮件"界面

在界面上单击"账户",即可打开"管理账户"功能,如图 3-19 所示。

收件	箱 - Outlook			- 🗆 X
			搜索	管理账户
+	新邮件		重点其他	选择要编辑设置的账户。
R	账户		其他:3 条新消息 微软官方商城, Windows, 微软 Of	Outlook liuqiangel@hotmail.com
	Outlook liuqiangel@hotmail.com	18	2019年7月18日	
Ē	文件夹		Microsoft 使用条款更新 您好,您之所以会收到」	 链接收件箱 一 还加账户
	收件箱	18	2018年3月22日	May111,2021

图 3-19 管理账户

(2) 在图 3-19 所示的界面中单击"+添加账户",弹出"添加账户"对话框,如图 3-20 所示。

添加账户	×
将账户 联系。	^白 添加到 邮件、日历 和 人员,以访问电子邮件、日历事件和 人。
0	Outlook.com Outlook.com, Live.com, Hotmail, MSN
0	Office 365 Office 365, Exchange
G	Google
	Yahoo!
\square	iCloud
Δ	其他账户 POP, IMAP
	× 关闭

图 3-20 "添加账户"对话框

(3)如果有 Google、Yahoo 和 iCloud 的邮箱,直接单击相应名称即可进入设置;如果是其他邮箱,则选择"其他账户 POP,IMAP"。这里选择"其他账户 POP,IMAP"。单击"其他账户 POP,IMAP"进入下一步设置,如图 3-21 所示。

因为安全保护的问题,绝大部分的 IMAP/POP3 邮箱的电子邮件客户端的 IMAP/POP3 的功能都是关闭的,需要首先登录邮箱页面,开启这些功能以后,才能在"邮件"系统中完成该邮箱的收发功能。这里以网易邮箱为例展示如何设置。其他邮箱的设置方法大同小异。

下面依次解释 POP3、SMTP 和 IMAP。

POP3 是 Post Office Protocol 3 的简称,即邮局协议的第 3 个版本,它规定怎样将个人计算机连接到 Internet 的邮件服务器和下载电子邮件的电子协议。它是因特网电子邮件的第一个离线协议标准,POP3 允许用户从服务器上把邮件存储到本地主机(即自己的计算机)上,同

大数据分析导论实验指导与习题集

时删除保存在邮件服务器上的邮件,而 POP3 服务器则是遵循 POP3 协议的接收邮件服务器, 用来接收电子邮件。

和账户		×
其他账户		
某些账户需要其他步骤才可登录。		
了解详细信息		
电子邮件地址		
someone@example.com		
使用此名称发送你的邮件 liuqi		
密码		
我们将保存此信息,以便你无须每次都进行	登录。	

图 3-21 账户信息设置

SMTP 的全称是 Simple Mail Transfer Protocol,即简单邮件传输协议。它是一组用于从源地址到目的地址传输邮件的规范,通过它来控制邮件的中转方式。SMTP 属于 TCP/IP 协议簇,它帮助每台计算机在发送或中转信件时找到下一个目的地。SMTP 服务器就是遵循 SMTP 的发送邮件服务器。SMTP 认证简单地说就是要求必须在提供了账户名和密码之后 才可以登录 SMTP 服务器,这就使得那些垃圾邮件的散播者无可乘之机。

增加 SMTP 认证的目的是使用户避免受到垃圾邮件的侵扰。

IMAP 的全称是 Internet Mail Access Protocol,即交互式邮件存取协议,它是跟 POP3 类 似的邮件访问标准协议之一。不同的是,开启 IMAP 后,用户在电子邮件客户端收取的邮件 仍然保留在服务器上,同时在客户端上的操作都会反馈到服务器上,如删除邮件、标记已读等,服务器上的邮件也会做相应的动作。所以无论从浏览器登录邮箱或者从客户端软件登录邮箱,看到的邮件以及状态都是一致的。

网易 163 邮箱相关服务器信息如图 3-22 所示。

服务器名称	服务器地址	SSL 协议端口号	非 SSL 协议端口号
IMAP	imap. 163. com	993	143
SMTP	smtp. 163. com	465/994	25
POP3	pop. 163. com	995	110

图 3-22 网易 163 邮箱相关服务器信息

IMAP和 POP3 的区别如下。

POP3 协议允许电子邮件客户端下载服务器上的邮件,但是在客户端的操作(如移动邮件、标记已读等)不会反馈到服务器上,比如通过客户端收取了邮箱中的3封邮件并移动到其

他文件夹,邮箱服务器上的这些邮件是没有同时被移动的。

而 IMAP 提供 Web 邮箱与电子邮件客户端之间的双向通信,客户端的操作都会反馈到服务器上,对邮件进行的操作,服务器上的邮件也会做相应的动作。

同时,IMAP像POP3那样提供了方便的邮件下载服务,让用户能进行离线阅读。IMAP 提供的摘要浏览功能可以让用户在阅读完所有的邮件到达时间、主题、发件人、大小等信息后 才做出是否下载的决定。此外,IMAP更好地支持了从多个不同设备中随时访问新邮件。

IMAP 和 POP3 的区别如图 3-23 所示。

操作位置	操作内容	IMAP	POPS
收件箱	阅读、标记、移动、删除邮件等	客户端与邮箱更新同步	仅客户端内
发件箱	保存到已发送	客户端与邮箱更新同步	仅客户端内
创建文件夹	新建自定义的文件夹	客户端与邮箱更新同步	仅客户端内
草稿	保存草稿	客户端与邮箱更新同步	仅客户端内
垃圾文件夹 接收误移入垃圾文件夹的邮件		支持	不支持
广告邮件	接收被移入广告邮件夹的邮件	支持	不支持

图 3-23 IMAP 和 POP3 的区别

总之,IMAP整体上为用户带来更为便捷和可靠的体验。POP3更易丢失邮件或多次下载相同的邮件,但IMAP通过邮件客户端与Web邮箱之间的双向同步功能很好地避免了这些问题。

注: 若在 Web 邮箱中设置了"保存到已发送",使用客户端 POP 服务发信时,已发邮件也 会自动同步到网页端"已发送"文件夹内。

为了邮箱安全,网易邮箱默认关闭 IMAP 服务,需要用户开启客户端协议以及授权码方 可使用,如果需要开启可以通过以下方式开启。

登录网页邮箱,单击邮箱页面上方的"设置"按钮,选择 POP3/SMTP/IMAP,根据实际需求开启 IMAP/SMTP 服务或者 POP3/SMTP 服务,如图 3-24 所示。开通后即可使用 Foxmail、Outlook 等第三方客户端进行收发邮件(注:单击"关闭"按钮,即可关闭成功。)

規设置 POP3/SMTP/IMA 箱密码修改 名 开启服	IAP
洛 开启服	路: IMAP/SMTP服务 已关闭 开启 手机服务
に 信分类	POP3/SMTP服务 已关闭 开启 POP3/SMTP/IMAP 登 损肤
长号与邮箱中心 ◎箱安全设置 ◎箱手机服务 反垃圾/黑白名单	POP3/SMTP/IMAP服务能让你在本地客户端上收发邮/ 温馨提示:在第三方登录网易邮箱,可能存在邮件泄 全,建议使用邮箱大师登录,扫描右侧二堆码,下數 PLUS会员专星皮肤

图 3-24 根据实际需求开启 IMAP/SMTP 服务或者 POP3/SMAP 服务

设置完成后,返回图 3-21 所示的账户信息设置界面,输入相应的邮箱信息,就可以使用 "邮件"客户端软件来收发邮件了。

实验 3.4 基于正则表达式获取中南财经政法大学教务

部通知公告

【实验要求】

36

(1) 安装第三方库 requests。

(2)使用浏览器查看中南财经政法大学教务部的通知公告网站网页源代码。

(3) 通过正则表达式获取所有教务部通知的标题和发表日期。

(4) 将数据写入 Excel 文件。

【实验步骤】

(1) 安装 Python 第三方库 requests。

在 Windows 环境下,单击左下角的"开始"菜单,选择"附件"→"命令提示符"命令,如图 3-25 所示。



图 3-25 选择"附件"→"命令提示符"命令

输入命令 pip install requests,如图 3-26 所示。该状态需要联网,若出现 requests successfully installed 则表示安装成功;如果安装失败则有可能是 Python 环境变量问题,需要重新安装 Python。

(2) 用浏览器打开中南财经政法大学教务部通知公告网页(http://jwc.zuel.edu.cn/ 5768/),如图 3-27 所示。

在网页空白处右击,在弹出的快捷菜单中选择"查看网页源代码"命令,如图 3-28 所示。





thoras	b财佳敬法大	学教务音	ß			游输入关键学	
网站首页	机构设置	规章制度	办事流程	常用下载	人才培养	支部建设	联系我们
通知公告.学生	j@;	即公告 学生				当前位置:	百页 通知公古,学生
	· 30	迎关注中南财经政法	大学教务部假信公众	5			2019-12-03
	· 关	于2019年下半年全国	大学英语四六级考试	战绩和2020年上半年	全国		2020-02-27
	· 20	20届同学关于毕业传	◎准备好了吗?			2020-02-17	
	* 全校本科教务管理咨询方式集锦					2020-02-16	
	·×	于调整2019-2020学	年第二学期本科教育	教学活动安排的通知((=	2020-02-04	
	· ¥	于调整2019-2020学	年第二学期本科教育			2020-01-29	
	· ¥	于2019-2020-2学期	辅修双学位上课安排			2020-01-15	
	· 关	于取消2019-2020学	年第二学期部分本科	课堂的通知			2020-01-07
	· ¥	于2017、2018级辅作	8双学位补缴学费的通			2020-01-02	
	· ¥	于2019-2020学年第	一学期期末考试安排			2019-12-17	
	· ¥	于2019-2020学年第	二学期本科生选课安			2019-12-13	
	· ¥	于2019-2020学年第	一学期期末考试拟安			2019-12-11	
	· ¥	于开展2019-2020学	年第一学期学生网上	*********	******************	2019-12-11	
	· 关	于2018级学生校内辅	修双学位报名工作的			2019-12-10	
			毎页 14	记录 总共 180 记录	第一页 <<上一页 下	一页>> 尾页 页码 1	/13 跳转到

图 3-27 用浏览器打开中南财经政法大学教务部通知公告网页

(3) 查找需要获取的关键词,在网页源代码中搜索标题"欢迎关注中南财经政法大学教务 部微信公众号",找到该标题对应的源代码位置,如图 3-29 所示。

分析网页源代码,构建抽取模式。通过观察可以发现,标题"欢迎关注中南财经政法大学教务部微信公众号"的模式如图 3-30 所示。

大数据分析导论实验指导与习题集

网站首页	机构设置	规章制度	办事流程	常用下载	人才培养	支部建设	联系我们	
				eî-îl			St.dy	
通知公告.学生	-	通知公告.学生				当前位置:	首页 通知公告,学生	
		· 欢迎关注中南财经政法;	大学教务部微信公众	号			2019-12-03	
		·关于2019年下半年全国	大学英语四六级考试	战成绩和2020年上半年	全国		2020-02-27	
		·2020届同学关于毕业你	你准备好了吗?				2020-02-17	
	返回(日) 前进(F) 建新加歇(R)	Alt+向左箭头 Alt+向右箭头 CriteR	式集視		2020-02-16			
			第二学期本科教育	1数学活动安排的通知(=		2020-02-04	
	另存为(A)	CUITR	第二学期本科教育	軟学活动安排的通知			2020-01-29	
		Ctrl+S	修双学位上课安排	的通知			2020-01-15	
	打印(P) Ctrl4 配成中文(简体中文)(T) 查看何页源代码(V) Ctrl4		第二学期部分本和	4课堂的通知			2020-01-07	
			又学位补缴学费的通知			2020-01-02		
	查看妈页信息(1)		学期期末考试安排	的通知			2019-12-17	
	审查元素(N)	Ctrl+Shift+I	学期本科生选课会	对非的通知			2019-12-13	
		· 关于2019-2020学年第一		胡时间的通知			2019-12-11	
	· 关于开爆2019-2020学 · 关于2018限学生校内辅			年篇一学期学生网上评败活动的通知			2019-12-11	
				的紧急通知			2019-12-10	
			每页 1	4 记录 总共 180 记录	第一页 <<上一页]	下一页>> 尾页 页码:	1/13	





图 3-29 找到该标题对应的源代码位置

源代码:target = '_blank' title = '欢迎关注中南财经政法大学教务部微信公众号'> 模式:target = '_blank' title = '(. * ?)'> 八 代码:re.findall("target = '_blank' title = '(. * ?)'>",webpage)

图 3-30 标题"欢迎关注中南财经政法大学教务部微信公众号"的模式

38

图 3-30 中的粗体字部分"欢迎关注中南财经政法大学教务部微信公众号"是需要从网页 源代码中抽取的内容,只需要把粗体字部分替换为(.*?)就变成所需要的模式,最后在 Python 代码中使用 re. findall("target='_blank' title='(.*?)>",webpage),会把 webpage (HTML 源代码)中所有满足模式条件的字符串以列表的形式返回。

同理,在网页源代码中搜索日期"2019-12-03",如图 3-31 所示。

energy and parts 7 Transfer (Laurie where and where " part (many " transfer (Laurie wash transfer)) (The state (Laurie van (Laurie where and where (Laurie (Lauree Laurie van (Laurie van transfer)))	2819-12-03	#15.819 A W #
(2) (UM "AND" NOT")		
(date)		
vite of an of the second state's		
(Table 11 Sele antice area in the rest of the area antice in the second se		
(Tree)	a contraction of the second second second	
A set of the set of	in the state of th	total share the term shareful
hereine under Beget trilstage "Baget "Baget "Baget "Baget "Baget "Baget "Bill (1965) in dieser Benefind 1966 Same Benefind 1	of (TTM/Rist.Mts' targets' 1907	ANAMIA (a) the class? senor
#110年19月1日,11月1日(11月1日)),11月1日)),11月1日(11月1日)(11月1日)(11月1日)))(11月1日))(11月1日))(11月1日)))(11月1日)))(11月1日)))(11月1日) 11月1日)(11月1日)(11月1日))(11月1日))(11月1日))(11月1日))(11月日))(11月日))(11月1日))(11月日))(11月日))(11月日))(11月日))(11月日))(11月日))(11日)	Taugetter _ end C (D) m A (F (G) / L)	1/12/02 class" sale-stee 22-3" ha
rappy, and ABERVS is also important way of the family advant death of the set of the family of the f	(bd' Ha thein "ed-bash" herb?	/SHOULAL. hts" targets", self") #
有文章のシーンは1991年19月1日は、dater "entries 14" An Section 14" Netric 14" 18年2月19日 - An Alexandro 2014 - Alexandro 201	a state" with link" heads (1989)	int.hts, targets, self (##6#8
and "Marine by insert" all SETENCE on dust more statement out of the set dust of the factor is for the dust of the bet (Willish by insert all SEEE) of	Whi share advice 15-7 Na (last'rd-las'
net/Williamin' uper/adi/#E/#ERAV#_UINI dae/ad-test#F/s dae/ad-bai/ tet/Williamin' tape/,ad//#BBAV#_UIND/DING dae/ad-bai/ad-bai/ad-bai/	of million bes' targets' pall")	Little Cal the Carrisson
的过去式和过去分词 化化合金 化合金 化合金 化合金 化合金 化合金 化合金 化合金 化合金 化合	harder and many here to	and share an or other
instante in the same and a best within the team, all BYSERS and in the show at 7 % due to the 'Shikin Ma' team' all BEREBYSE Since Solution	" "lair" and the if'd a later	"mearlish" hod?/1941/list.bts"
territ, wir (1998) in der verwalterrit (201) is darf dense Garte Als Gart der Geren (201) is der verster in (201) Statister (201) is der versterrit (201)	P.T. Ha (Laro) "sid-Luss" horth"	78740/3141.Html fielgettr", pdII 7/3
Construction of the second residence of the second seco	CLORENCE AND THE REPORT OF THE	altriate target and the start the
and mid 10 Australia Contact and and and and and and an an an	"His class" spisses" portlata	uder similelalumattes' frege #10
1271年後の日本の予知時代である「大阪市市」「大阪市市においた」「おいた」はなど、「おいた」のおいてのか、「おいた」のなどの「おいた」のなどのなどのなどのなどの「おいた」ではないで、「おいた」である「おいた」である「おいた」である「おいた」である「おいた」である」である「おいた」である「おいた」である「おいた」である」である「おいた」である「おいた」である「おいた」である「おいた」である」である「おいた」」「おいた」「おいた」」「おいた」「おいた」」「おいた」「おいた」「お	rises clearly rike class w	Produce terms in the set of the set of the
anticipation application in a TDF sector and a later any conference of the data and characterized in the data in t	picthetaider"stapleColumnities"	frage #128' bld chase where
+ris*: 麦利亚美学家/Norther states* educer#12 /right=Line*/addresser/addresse	白鼻、芋茸 ジョ・ジネル ジネル ジネル・	the charry column sent cost
11月、11日、11日、11日、11日、11日、11日、11日、11日、11日、	(株社大学教会研究部はならちいかい)	tentions dans'rdenastriate
see date half. The set of Table 2 and the set of the se	课和2021年上半年全国大学员道顶;	18年18年1月1日日1月1日日1月1日日1日日1日日1日日1日日1日日1日日1日日1日日1
가지의 가수로 들려갔지 않는 것은	HER Y SHOULD HAVE A HER AND A HER AN	an its targets hims totler 200
	/0001/0014//0110142301002/2424.14	a tarpets' black titler Kyuda
21日本2017年年第二年第三月第三年前の学校の連載(二)「大子通常に体の空空発展二年第三月第三日に登録を通知で通知です」。ジャンクルン「Part」では、Parting State Table (2017年)	an itee 6 elevefur blagen clara	- collour rest title 20
Ame 7 Ganetia 「America Come Come Come Come Come Come Come Come	Game million more data poser tar	whate' (2020-00-15//spect-0/16/01a
[1994] 《Jan new Institute deadle" (1994) [1995] [1905] [1995] [1905] [1	「本料理算的書料」というからになか	classy columneer dets neer date-
And	and an electron by the Provide Street	(1) 第一次目前のありたり目が通信につい
(appropriate and the second seco	于2019-2029年後第二学校支持支持	唐於明於萬知 1年于2013年2025年年展
二字形式 日本 第二字 第二字形式 W 小の (1999) (1999) (1999) (1999) (1999) (1999) (1997	hts targets that totas Rt	「おけていいする書へ手規規未有に設計
Terpite Line Televic STREED 201948-988910-201948-988910-201948-978910-201948-989101-201948-000100-001000-00000000000000000000000	him neer the ther it clearftr	"Hope class" mise new trils")
这个Proce_CORPUTITIONAL的内容的因为 ¹⁰⁰ Proce_Corputational 中国的中国的全国的有关的主要事件。这上DRAM在全部由学校由学校由学校由学校的生活中,Proce_Corputational elementation and elementation and elementational elementation.	H2-MO gentOTD-0425-Utp-F	

图 3-31 在网页源代码中搜索日期"2019-12-03"

通过观察可以发现,日期"2019-12-03"的模式如图 3-32 所示。

图 3-32 日期"2019-12-03"的模式

图 3-32 中的粗体字部分"2019-12-03"是需要从网页源代码中抽取的内容,只需要把粗体 字部分替换为(.*?)就变成所需要的模式,最后在 Python 代码中使用 re. findall('< span class="column-news-date news-date-hide">(.*?)',webpage),会把 webpage (HTML 源代码)中所有满足模式条件的字符串以列表的形式返回。

获取中南财经政法大学教务部首页的标题和日期的完整代码如图 3-33 所示。

import re	#引入正则库
import requests	#引入 requests 库
url = 'http://jwc.zuel.edu.cn/5768/'	#设置需要访问的 URL 网址
response = requests.get(url)	#访问中南财经政法大学教务部通知公告网
response.encoding = 'utf - 8'	#设置编码格式为 utf-8,防止乱码
webpage = response.text	#webpage 中获取网页源代码
<pre>titles = re.findall("target = '_blank' title =</pre>	'(. * ?)'>",webpage) #获取所有标题
dates = re.findall('< span class = "column - ne	ews - date news - date - hide">(. * ?)',webpage)
	# 获取日期
for title in titles:	#依次打印所有标题
print(title)	
for date in dates:	#依次打印所有日期
print(date)	

图 3-33 获取中南财经政法大学教务部首页的标题和日期的完整代码

为了能够获取所有的通知,实现自动翻页,观察每次翻页后网址的变化规律,如图 3-34 所示。

- 第1页:http://jwc.zuel.edu.cn/5768/list1.htm
- 第2页:http://jwc.zuel.edu.cn/5768/list2.htm

第 13 页:http://jwc.zuel.edu.cn/5768/list13.htm

图 3-34 观察每次翻页后网址的变化规律

通过分析我们知道这些网址的前面都是 http://jwc.zuel.edu.cn/5768/list,后面都是 htm,只是中间的数字代表了页码,所以使用 for 循环连续生成翻页网址,如图 3-35 所示。

```
lefturl = "http://jwc.zuel.edu.cn/5768/list"
righturl = ".htm"
for i in range(1,13 + 1):
    url = lefturl + str(i) + righturl
    print(url)
```

图 3-35 使用 for 循环连续生成翻页网址

这里通过 for i in range(1,13+1)可以生成 1~13 的序列,通过 url=lefturl+str(i)+ righturl 将 URL 通过加号拼接起来,形成完整的 URL。通过翻页功能形成新的代码获得所 有标题和日期,如图 3-36 所示。

```
import re
                                            #引入正则库
import requests
                                            #引入 requests 库
lefturl = "http://jwc.zuel.edu.cn/5768/list"
                                            #每页 URL 的左边部分
righturl = ".htm"
                                            #每页 URL 的右边部分
for i in range(1,13 + 1):
                                            #让i形成1~13的序列
   url = lefturl + str(i) + righturl
                                            #形成每页的 URL
   response = requests.get(url)
                                            #访问中南财经政法大学教务部通知公告网
                                            #设置编码格式为 utf-8, 防止乱码
   response. encoding = 'utf - 8'
   webpage = response.text
                                            #webpage 中获取网页源代码
   titles = re.findall("target = '_blank' title = '(. * ?)'>",webpage) #获取所有标题
   dates = re.findall('< span class = "column - news - date news - date - hide">(. *?)</span >', webpage)
                                            #获取日期
   for title in titles:
                                            #依次打印所有标题
       print(title)
                                            #依次打印所有日期
   for date in dates:
       print(date)
```

图 3-36 通过翻页功能形成新的代码获得所有标题和日期

(4) 将数据写入 Excel 表:为了将数据写入 Excel,需要引入 csv 库。在操作 Excel 表格时,需要将 Excel 数据转换为一个二维列表,如图 3-37 所示。

标题1	日期1	rows=_ ['标题 1','日期 1'],
标题 2	日期 2	['标题 2','日期 2'],
标题 3	日期3	['标题 3','日期 3'] 」

图 3-37 将 Excel 数据转换为一个二维列表

在图 3-37 中的代码里,直接打印出来标题和日期,需要使用 append()方法将其添加到列 表中,于是定义两个新的空白列表 titlelist 和 datelist,将循环中的 title 和 date 添加到列表中, 如图 3-38 所示。

for title in titles:		for title in titles:
print(title)	\Rightarrow	<pre>titlelist.append(title)</pre>
for date in dates:		for date in dates:
print(date)		datelist.append(date)

图 3-38 将循环中的 title 和 date 添加到列表中

需要将 titlelist 的每一个元素和 datelist 的每一个元素组队,形成图 3-39 的结构,需要通过 zip()方法进行列表合并,如图 3-39 所示。

titlelist = [标题 1,标题 2,标题 3…标题 n] datelist = [日期 1,日期 2,日期 3…日期 n] ↓ newlist = zip(titlelist,datelist) newlist = [(标题 1,日期 1),(标题 2,日期 2),….(标题 n,日期 n)]

图 3-39 通过 zip()方法进行列表合并

新的列表形成之后,将其列表写入 Excel,如图 3-40 所示。

<pre>file = open('download.csv', 'w', no</pre>	ewline = '')
<pre>f_csv = csv.writer(file)</pre>	#准备写入
f_csv.writerows(newlist)	#写人数据
file.close()	#关闭文件

图 3-40 将其列表写入 Excel

在写入数据时,需要用 open()函数建立一个 csv 文件并打开它,即 file=open(' download. csv', 'w',newline=''),open()函数的第一个参数 download. csv 是需要打开的文件名字,第二个参数 w 代 表以写入的方式打开文件,第三个参数 newline="表示在行与行之间不需要空行。 获取标题和日期并写入 Excel 文件的完整代码如图 3-41 所示。

```
#引入正则库
import re
import requests
                                               #引入 requests 库
                                               #引入 csv 库
import csv
lefturl = "http://jwc.zuel.edu.cn/5768/list"
                                               #每页 URL 的左边部分
righturl = ".htm"
                                               #每页 URL 的右边部分
datelist = []
                                               #用于存储所有日期的列表
titlelist = []
                                               #用于存储所有标题的列表
for i in range(1, 13 + 1):
                                               #让 i 形成 1~13 的序列
   url = lefturl + str(i) + righturl
                                               #形成每页的 URL
   response = requests.get(url)
                                               #访问中南财经政法大学教务部通知公告网
   response. encoding = 'utf - 8'
                                               #设置编码格式为 utf-8, 防止乱码
   webpage = response.text
                                               #webpage 中获取网页源代码
   titles = re.findall("target = '_blank' title = '(. * ?)'>", webpage) #获取所有标题
   dates = re.findall('< span class = "column - news - date news - date - hide">(. *?)</span>', webpage)
                                                #获取日期
   for title in titles:
                                                #将 title 加入列表
```

图 3-41 获取标题和日期并写入 Excel 文件的完整代码

<pre>titlelist.append(title)</pre>	
for date in dates:	#将 date 添加到列表中
datelist.append(date)	
<pre>newlist = zip(titlelist,datelist)</pre>	#为了写入 Excel 文件将列表合并
<pre>file = open('download.csv', 'w', newline = '')</pre>	#打开 Excel 文件
<pre>f_csv = csv.writer(file)</pre>	#准备写入
f_csv.writerows(newlist)	#写入数据
<pre>file.close()</pre>	#关闭文件
print(len(titlelist),"条数据写入 download.c	sv 文件中")

图 3-41 (续)

该程序运行后会在当前目录下形成一个 download 的 Excel 文件,下载的 download. csv 的内容如图 3-42 所示。

X	a 17 - C1	* [*								_				_			¢	download
文作	+ 开始	插入	页	面布局	公式	数据	审阅	视图	A	crobat								
[BC				12			alla	.M.	-			1	1	1.0	1	turn.		0
4		100	BE.	UP	40	100 +		WX.	-	-	Press (1 m	In	1.00°		60
数据	著 表格	图片 1	乾贴画	形状	SmartArt	屏幕截图	柱形图	折线图	饼图	条形图	面积图	散点图	其他图表	折线置	柱形图	盈亏	切片器	超链接
透视	£ •										•	*						
	表格			插图	100					图表			6		迷你图		筛选器	链接
	Å1		• (**	f,	次迎;	<注中南!	财经政济	去大学	教务部	B微信公	公号							
(A					A							В			С	D		E
1	欢迎关注	中南财经	政法	大学教	务部微信	公众号							2019/12	2/3		-	_	
2	关于2019-	-2020学	年第:	二学期	体育保健	班教学多	7排的通	的	_				2020/3/	12				
3	天于2019	-2020字	年第.	二字期	本科课程	↑退选」	_作安排	的通知] 山(左)				2020/3/	11			_	
4	大于2019	キトキキ	F 全国	大子兵	と宿四六気	这专认财务	喷和202	20年上	羊牛 至	리그			2020/2/	27		-		
0	2020)油回	子大丁斗 新久等項	三亚小	/住留幻	r] µ≕] ? :čé								2020/2/	17				
0 7	王仪平村	秋芳昌玛 2010-20	20学	力式県	计师	教育教育	的活动学	2111-653	5 ± Π (-			2020/2/	10		-		
0	大」 明堂(2019-20	20子	十-77-2	子期本相	教育教子	か活动学	241103.00	11 11 11 11 11 11 11 11	-			2020/2	20				
q	关于2019-	-2020-2	学期	捕修羽	学位上课	安排的证	百年(1)54 百年(1)54	011110.00	1/1				2020/1/	15				
10	关于取消	2019-20	20学:	年第一	学期部分	本科课堂	201通知						2020/1	/7		-		
11	关于2017	· 2018	及辅修	双学位	沐嶽学	書的通知							2020/1	/2				
12	关于2019-	-2020学	年第	一学期	期末考试	安排的通	重矢口						2019/12/	17				
13	关于2019-	-2020学	年第:	二学期	本科生选	课安排的	的通知						2019/12/	13				
14	关于2019-	-2020学	年第·	一学期	期末考试	拟安排时	时间的通	重失口					2019/12/	11				
15	关于开展2	2019 - 2	020学	年第-	一学期学生	主网上评	教活动	的通知				1	2019/12/	/11				
16	关于2018	级学生相	交内辅	修双字	6位报名]	C作的紧;	急通知					1	2019/12/	10				
17	关于2017	· 2018\$	及辅修	双学位	立缴纳学到	贵的通知							2019/12	2/9			-	
18	第十五届	数学竞赛	获奖	名单与	领奖通知	1				_			2019/12	2/5				
19	关于开展	栈校201	9年校	内辅修	双学位打	8名工作1	的通知			_			2019/12	2/4		-	_	
20	关于2019	年下半年	国全4	大字英	语四六朝	5考试准:	考证打印	7的通9	ŧΩ	_			2019/12	2/3		-	_	
21	2019-202	0字年第	一字!	明明中	考试违纪	舞弊情力	1.通报					-	2019/11/	27			_	
22	甲闸财经1	以法大字	第十	立 盾 釣	子克费者	10次排		** 6528	40				2019/11/	19				
23	大于2019	-2020子	牛弗	一子期	坐村味在 1回太利義	(第9周7	「味り返	达的通	χη.				2019/11	/6		-		
29	大于2019-	4秋字口 -2020学	1111运	列本界	加中科学	和耳運台	い周生の					- 1	2019/10/	29				
26	关于延长:	2020子	十分	于197	时间的道	- */4/1 い不口 単午日	176 21			-			2019/10/	20				
27	关于领取	2019年音	(二)方	普通词	6水平测试	北等级证:	书的通知	60				1	2019/10/	14				
28	关于做好	2018级者	福本	科学生	专业分别	在工作的i	通知						2019/10	0/9				
29	关于开展	2019年第	三次	计算机	l辅助普i	1话水平)	则试报	名工作的	的通知	1			2019/10	0/8				
30	关于2019	级本科新	ff生体	育课送	课安排的	句通知							2019/9/	20				
31	关于2019	级新生本	5科教	材发放	安排的证	重矢口							2019/9/	18				
32	关于我校的	第十五届	数学	竞赛报	名的通知	1							2019/9/	12				
33	关于2019	年下半年	F全国	大学英	语四六组	吸考试报:	名工作的	的通知					2019/9/	10				
34	大学数学	教研中心	2019	-2020	学年第一	学期答判	安排		1. 10	_			2019/9/	10				
35	关于开展2	2019年丁	、半年	中国小	>数民族;	又语水平	手级考证	a, (MHH	()报:	名.			2019/9	9/9				
36	天于2019	-2020字	年第	一字期	本科课程	补选工作	安排的	通知	hr 4-6-17	. he			2019/9	9/5		-	-	
37	大十升展	推存202	い届优	穷本科	*罕业生9	日113.以1实行	则工研》	七王二1	F的通	1×L			2019/9	/1			-	_
38	大于2018-	-2019字	牛弗.	_子期	级考安排 期士委员	町建知	ST/E#	1) and the					2019/8/	31			-	
39	大丁2018-202	-2019子	牛弟.	_子明 約材少	前承知	四次反复重	TIFR	기표지나		-			2019/8/	20		-	-	
40	2019-202 关于南湖:	61子所	出口	我们仅自习室	安排的译	ñ4n							2019/8/	29			-	
42	2018-201	9学年筆	一学	日ク里 相相主	老试住纪	■ 毎幣情S	7通报 ((-)					2019/1	1/2		-	_	
43	我校本科	通识法的	课程	一皆表	- J 44.0250	247118//	overant /			-			2019/6	12		-	-	
44	关于2018-	-2019学	年第	二学期	期末考试	安排的证	前午口						2019/6/	12				
	MT	Dear J		11 11/ 10	Hart Art St.	34 200 65 12	X Lo							1.0				

图 3-42 下载的 download. csv 的内容