《网络爬虫》实训指导书

院系名称：经管信息学院

课程代码：31091430

总学时数：48

适用专业：大数据技术与应用专业

编 制 人：孙 检

编制日期：2020 年 7 月

审 核 人：

审 定 人：

《网络爬虫》实训指导书

一、实训目的与要求

《网络爬虫》实训的教学目的是学生通过学习该课程，掌握爬虫概述、Ullib 实现网站下载、使用正则表达式获取网页数据、使用 beautifulsoup 工具选择数据、使用 scrapy 编写网页爬虫程序、使用 item、pipeline 实现数据序列化与存储等基础知识。课程着眼于学生的长远发展，重点培养其网络爬虫、大数据及数据分析领域岗位基本工作技能、职业素养、社会适应能力、交流沟通能力、团队协作能力、创新能力和自主学习能力。

二、实训内容

（一）实例实训

以 Python 程序设计的实例指导学生如何独立完成 Python 程序设计与编写。让学生在机房实际操作，按照给定的要求完成相应任务。

（二）网络爬虫实训

让学生自己选择不同的网页爬取数据，根据项目要求，对数据分析转换，可以进行本地保存数据和数据可视化显示。

（三）总结

对学生的全部作品进行考核，并选择典型的案例对实训的结果进行考核。

三、参考课时

标题	实训内容	实训课时
实训一	Python 基本语法回顾	4
实训二	序列结构	4
实训三	文件处理	4
实训四	面向对象	4
实训五	Excel 数据	6
实训六	爬取社交网络数据	8
实训七	使用 Python 实现网络爬虫算	8
实训八	批量获取网络图片数据	6

实训九	使用 Python 处理图片尺寸和	6
实训十	使用 Python 统计分析社交数	6
实训十一	总结	4
总计		60

四、实训材料准备

（一）软件准备

Python3 以上版本 、PyCharm2019 中文版、RegexBuddy4.9（抓包工具，版本和工具不做要求）

（二）硬件准备

网络条件：与因特网连接的局域网。

教师用机：Windows 10。

学生用机：Windows 10。

五、综合实训考核办法：

系统文档　20 分

项目功能　40 分

程序调试　10 分

实训出勤　20 分

效果展示　10 分

目　　录

实训一 Python 基本语法回顾

一、实训目的和要求

1. 掌握 Python 语言的基本语法；
2. 掌握 Python 语言中创建模块的方法；
3. 了解 Python 语言中定义类及其使用方法；
4. 学习使用 Python 语言输出斐波那契数列的方法；
5. 学习使用 Python 语言实现删除一个 list 里面的重复元素的方法。

二、实训内容

1. 根据 Python 基本语法功能设计出实现输出斐波那契数列的方法，并比较不同实现方法的性能；
2. 根据 Python 语言中的排序和循环功能，实现删除一个 list 里面的重复元素。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计输出斐波那契数列的 Python 程序分析实验要求；
2. 逐个打印输出斐波那契数列的元素记录程序代码；
3. 记录并分析实验结果；
4. 设计程序删除一个 list 里面的重复元素分析实验要求；
5. 对 list 进行排序；
6. 从后向前查找并删除 list 中的重复元素记录程序代码；
7. 记录并分析实验结果。

设计输出斐波那契数列的 Python 程序：首先调用 raw_input 输入要打印的斐波那契数列的长度，然后把斐波那契数列存储于一个序列当中，并逐个打印序列的元素。

此实验部分实现代码如下

```
#通过输入斐波那契数列的长度打印斐波那契数列
FibonacciUptoNumer = int(raw_input('Please input a  Fibonacci
Series up to Number : '))
n = FibonacciUptoNumer
fibs = [0, 1]
for number in range(n): fibs.append(fibs[-2] + fibs[-1])
```

设计删除一个 list 里面的重复元素程序：首先调用 List.sort()对序列进行排序，然后调用 last = List[-1] 语句从后向前找出重复的元素，并逐个打印非重复的元素。

此实验部分实现代码如下

```
if List:
List.sort() last = List[-1]
for i in range(len(List)-2, -1, -1):
if last==List[i]:del List[i] else: last=List[i]
print List
```

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用上机练习的方法，考核以操作的正确性为评分标准。具体评分如下：

1. “ Goodmorning ，everyone ” 格式化输出。（20 分）
2. 水果价格表打印结果。（30 分）
3. 字典练习。（40 分）
4. 代码规范。（10 分）

七、思考和练习

1. Python3 中使用 print()函数默认是否换行？如何不换行？
2. Python3 数据类型有哪几种？
3. Python3 格式化输出的几种方式。

实训二 序列结构

一、实训目的和要求

1. 学习使用网络爬虫输出斐波那契数列的方法；
2. 学习使用网络爬虫实现删除一个 list 里面的重复元素的方法。

二、实训内容

1. 根据 Python 基本语法功能设计出实现输出斐波那契数列的方法，并比较不同实现方法的性能。
2. 根据网络爬虫中的排序和循环功能，实现删除一个 list 里面的重复元素。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计输出斐波那契数列的 Python 程序分析实验要求，逐个打印输出斐波那契数列的元素记录程序代码，记录并分析实验结果。

此实验部分实现代码如下

```
#通过输入斐波那契数列的长度打印斐波那契数列
FibonacciUptoNumer = int(raw_input('Please input a Fibonacci Series up to Number : '))
n = FibonacciUptoNumer
fibs = [0, 1]
for numberin range(n):
```

```
fibs.append(fibs[-2] + fibs[-1])
```

2. 设计程序删除一个 list 里面的重复元素分析实验要求，对 list 进行排序，从后向前查找并删除 list 中的重复元素记录程序代码，记录并分析实验结果。

此实验部分实现代码如下

```
if List:
List.sort()
last = List[-1]
for i in range(len(List)-2, -1, -1):
if last==List[i]: del List[i]
else:last=List[i]
print List
```

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. SchoolMem、Student、Teacher 类的创建及各自关系。（20 分）

2. SchoolMem、Student、Teacher 类的属性书写。（30 分）

3. Student、Teacher 类的 printInfo 方法实现。（40 分）

4. 代码书写规范。（10 分）

七、思考和练习

1. 面向对象的特征。
2. Python 的继承方式，Python 是否支持多继承？
3. Python 的重写、重载操作。

实训三 文件处理

一、实训目的和要求

1. 掌握文件读写等基本操作的实现；
2. 掌握异常处理的基本方法；
3. 掌握简单的正则表达式规则，能用正则表达式处理分析一些常见的网络数据。

二、实训内容

文件、异常处理和正则表达式。

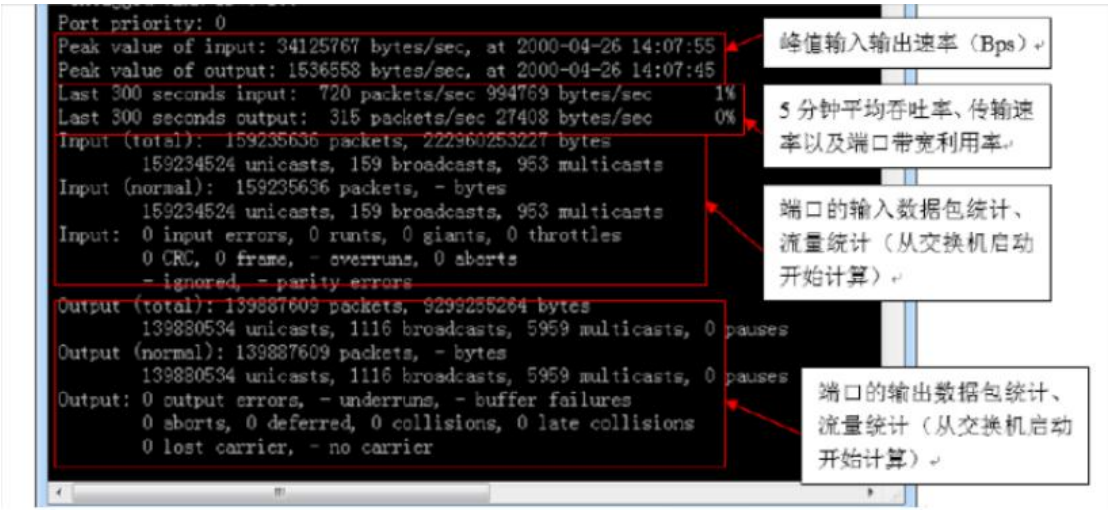
三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 创建文件 hello.txt,写入内容“hello，world”，向文件“hello.txt”中追加从 0 到 9 的随机整数,10 个数字一行，共 10 行整数。

2. 分析交换机中的数据，如下图所示，按照要求解析出数据，并保存到文本文件中。



输入峰值速率（ bytes/sec）：
输出峰值速率（ bytes/sec）：
5 分钟平均输入速率 (packets/sec,bytes/sec)：
5 分钟平均输出速率 (packets/sec,bytes/sec)：
5 分钟平均输入带宽利用率：
5 分钟平均输出带宽利用率 ：
输入总包数：（ packets）
输入总流量（ bytes）：
输出总包数：（ packets）
输出总流量（ bytes）：

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用上机练习的方法，考核以操作的熟练程度和正确性为评分标准，以 A（优秀）、B（良好）、C（及格）、D（不及格）为成绩标准。

七、思考和练习

1. file()函数的参数‘r’，‘w’，‘a’各自有什么不同?
2. Python 文件夹操作的方式。

实训四 面向对象

一、实训目的和要求

1. 掌握面向对象的基本概念，掌握 python 中面向对象的基本实现方法；
2. 能利用面向对象的基本思想解决实际问题。

二、实训内容

1. 掌握 python 基本程序设计流程和基本语法；
2. 掌握面向过程与面向对象编程思路。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 创建 SchoolMem 类，该类中包含三种属性：姓名、性别、年龄以及针对每个属性的 get 和 set 方法。
2. 创建 Student 类，继承自 SchoolMem 类，添加额外三个属性：班级、学号和数量统计。
3. 创建 Teacher 类，继承自 SchoolMem 类，添加额外三个属性：科室、工号和数量统计。
4. 要求在 Student 类和 Teacher 类中分别实现 printInfo 方法，该方法打印对象的多有属性信息。

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. SchoolMem、Student、Teacher 类的创建及各自关系。（20 分）

2. SchoolMem、Student、Teacher 类的属性书写。（30 分）

3. Student、Teacher 类的 printInfo 方法实现。（40 分）

4. 代码书写规范。（10 分）

七、思考和练习

1. 面向对象的特征。
2. Python 的继承方式，Python 是否支持多继承？
3. Python 的重写、重载操作。

实训五 Excel 数据处理

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 学习两种读取的 Excel 数据的方法；
3. 学习写入 Excel 数据的方法；
4. 掌握如何读写其他格式数据的方法；
5. 掌握如何比较不同读写方法的运算性能。

二、实训内容

1. 用 xlrd 模块中的 open_workbook 实现打开 Excel 数据表，并设计使用索引和名称两种方法读取 Excel 数据，最终写入 csv 文件中；

2. 用 datetime 模块中的 datetime.now 来计算两种不同的读取方法所用 CPU 时间，从而比较并分析不同算法的性能。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计按名称和按索引读取 Excel 数据的程序，分析实验要求，按行打印 Excel 表中的数据，记录程序代码，记录并分析实验结果。
2. 设计写入 Excel 数据的程序，分析实验要求，按行将数据写入 Excel 表中，记录程序代码记录并分析实验结果。
3. 设计计算程序运行时间的程序，分析实验要求，记录程序代码，比较并分析实验结果，总结、撰写实验报告。
4. Python 语句读取 Excel 表数据时，首先要调用 xlrd 模块，然后使用语句 data = xlrd.open_workbook('excelFile.xls')打开 Excel 表格。

此实验部分实现代码如下

```
from pyExceleratorimport *
w = Workbook()            #创建一个工作簿
ws = w.add_sheet('test') #创建一个工作表
ws.write(0,0,'uestc')    #在 1 行 1 列写入 uestc
ws.write(0,1,'Sofrware') #在 1 行 2 列写入 Software
ws.write(1,0,'cs')       #在 2 行 1 列写入
csw.save('mini.xls')     #保存至 mini.xls 文件中
```

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. Excel 表数据的读取。（30 分）

2. Excel 表数据分析。（40 分）

3. Excel 表数据写入。（30 分）

七、思考和练习

1. 还有哪些模块、库可以操作 Excel 数据表；
2. Excel 数据表写入格式转换。

实训六 爬取社交网络数据

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 学习社交网络 OAUTH 协议的原理；
3. 学习使用 Python 语言获取社交网络数据。

二、实训内容

1. 理解社交网络 OAUTH 协议的原理，并学习获取 CONSUMER_KEY；
2. CONSUMER_SECRET、USER_TOKEN、USER_SECRET 的方法。
3. 用 Python 语言中的 Json、OS、Linkedin 模块对 LinkedIn 网站中联系人名单进行搜集。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

社交网络 OAUTH 协议原理：

OAUTH 协议为用户资源的授权提供了一个安全的、开放而又简易的标准。与以往的授权方式不同之处是 OAUTH 的授权不会使第三方触及到用户的帐号信息（如用户名与密码），即第三方无需使用用户的用户名与密码就可以申请获得该用户资源的授权，因此 OAUTH 是安全的。

本实验中 Linkedin 网站的 OAUTH 协议是采用 HMAC-SHA1 加密的。开发者需要注册 LinkedIn 账户，获得 CONSUMER_KEY(即 APIKey)和 CONSUMER_SECRET。KEY 跟 SECRET 的使用方式跟其他一些协议中的公钥私钥的方案相类似，你可以使用你所熟悉的编程语言将 KEY 和 SECRET 结合，为你发出的每个请求添加签名，以此来向 Linkedin 开放平台表明自己身份的合法性。然后根据 CONSUMER_KEY 和 CONSUMER_SECRET 获取 USER_TOKEN 和 USER_SECRET。这个步骤主要有两个目的：第一，告诉 Linkedin 将要做什么；第二，告诉 Linkedin 在 callback 里要做什么。此外，USER_TOKEN 和 USER_SECRET 可以帮助提供 ACCESSTOKEN。

实现代码如下：

```
access_token_url='https://api.linkedin.com/uas/oauth/accessToken'token=oauth.Token(request_token['oauth_token'],request_token['oauth_token_secret'])token.set_verifier(oauth_verifier)
client=oauth.Client(consumer,token)
resp,content=client.request(access_token_url",POST")access_token=dict(urlparse.parse_qsl(content))
print"AccessToken:"
print"-oauth_token=%s"%access_token['oauth_token']
print"-oauth_token_secret=%s"%access_token['oauth_token_secret']print
```

print"Youmaynowaccessprotectedresourcesusingtheaccesstokensabove."

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. 网页数据的爬取。（30 分）
2. 网页数据分析。（40 分）
3. 网页数据保存。（30 分）

七、思考和练习

1. 爬虫程序相关模块；
2. 爬虫工具有哪些。

实训七 使用 Python 实现网络爬虫算法

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 学习网络爬虫算法的原理；
3. 学习使用 Python 语言实现网络爬虫算法。

二、实训内容

1. 理解网络爬虫算法的原理，并设计使用网络爬虫获取网页数据的程序；
2. 用网络爬虫中的 threading 和 GetUrl 模块对网站中 URL 进行搜集。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计某一个网页上获取数据的程序
分析实验要求
打印网页上获取的数据
记录程序代码
记录并分析实验结果。
2. 设计多线程的获取网站 URL 的程序
分析实验要求
打印网站上相关的 URL
比较不同线程数的算法性能
记录程序代码
记录并分析实验结果。

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. 网页数据的爬取。（30 分）

2. 网页数据分析。（40 分）

3. 网页数据保存。（30 分）

七、思考和练习

1. 爬虫程序相关模块；
2. 爬虫工具有哪些。

实训八 批量获取网络图片数据

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 了解大批量获取网络图片的原理；
3. 学习使用 Python 语言批量获取网络图片。

二、实训内容

1. 了解大批量获取网络图片的原理，并掌握批量获取网络图片的方法；
2. 用网络爬虫中的 urllib、urllib2 等模块对图虫网站中的图片进行批量下载，并存储在指定的文件夹中。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计批量获取图片的程序

分析实验要求

批量获取并存储网络图片

记录程序代码

记录并分析实验结果。

批量获取网络图片的方法：

3. 批量获取网络图片的方法是通过解析网页的 HTML 文件，利用正则表达式把源代码中的图片地址过滤出来，从而根据过滤出来的图片地址下载网络图片。具体来说，批量获取网络图片的方法分为三种，一是用微软提供的扩展库 win32com 来操作 IE，二是用 selenium 的 webdriver，三是用 python 自带的 HTMLParser 解析。

实现代码如下：

```
#获取二级页面 url
def findUrl2(html):
```

```
    re1=r'http://tuchong.com/\d+/\d+/|http://\w+(?<!photos).tuchong.com/\d+/'
    url2list = re.findall(re1,html)
    url2lstfltr = list(set(url2list))
    url2lstfltr.sort(key=url2list.index)
    #print url2lstfltr
    return url2lstfltr
#获取 html 文本
def getHtml(url):
    html = urllib2.urlopen(url).read().decode('utf-8')#解码为 utf-8
    return html
#下载图片到本地
def download(html_page, pageNo):
    #定义文件夹的名字
    x = time.localtime(time.time())
    foldername=str(x.__getattribute__("tm_year"))+"-"+str(x.__getattribute__("tm_mon"))+"-"+str(x.__getattribute__("tm_mday"))
    re2=r'http://photos.tuchong.com/.+/f/.+\.jpg'
    imglist=re.findall(re2,html_page)
    print imglist
    download_img=Nonefor imgurl in imglist:
    picpath='C:\\Users\\peterlindi\\Desktop\\lindi\\%s\\%s'%(foldername,str(pageNo))
    filename = str(uuid.uuid1())
    if not os.path.exists(picpath):
        os.makedirs(picpath)
        target= picpath+"\\%s.jpg" % filename
        print "The photoslocation is:"+target
        download_img = urllib.urlretrieve(imgurl, target)
        time.sleep(1)
        print(imgurl)
    return download_img
```

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. 批量获取图片。（30 分）
2. 图片批量保存到本地。（40 分）
3. 图片过滤与代码逻辑。（30 分）

七、思考和练习

1. 网络爬取的图片怎么保存到本地；
2. 图片的二进制格式转换。

实训九 使用 Python 处理图片尺寸和角度

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 学习图片的像素矩阵表示方法；
3. 学习使用 Python 语言调整图像尺寸和角度。

二、实训内容

1. 学习使用像素矩阵表示图片的方法；
2. 用 Python 语言中的 Image 等模块对图片尺寸、角度等进行处理的方法。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

1. 设计图片处理方法的程序

分析实验要求

实现指定图片的尺寸和角度进行调整

记录程序代码

记录并分析实验结果。

1）图片的像素矩阵表示：

数字图像数据可以用矩阵来表示，因此可以采用矩阵理论和矩阵算法对数字图像进行分析和处理。最典型的例子是灰度图像。灰度图像的像素数据就是一个矩阵，矩阵的行对应图像的高（单位为像素），矩阵的列对应图像的宽（单位为像素），矩阵的元素对应图像的像素，矩阵元素的值就是像素的灰度值。在计算机数字图像处理程序中，通常用二维数组来存放图像数据。二维数组的行对应图像的高，二维数组的列对应图像的宽，二维数组的元素对应图像的像素，二维数组元素的值就是像素的灰度值。采用二维数组来存储数字图像，符合二维图像的行列特性，同时也便于程序的寻址操作，使得计算机图像编程十分方便。图像的位图数据是一个二维数组（矩阵），矩阵的每一个元素对应了图像的一个像素，当保存一幅图像时，不但要保存图像的位图数据矩阵，还要将每个像素的颜色保存下来，颜色的记录是利用颜色表来完成的。颜色表，也叫颜色查找表，是图像像素数据的颜色索引表。以一个 4 色位图为例，则其颜色表有 4 个表项，表中每一行记录一种颜色的 R、G、B 值，这样，当表示一个像素的颜色时，只需要指出该颜色在第几行，即该颜色在表中的索引值即可。

2） Python 语句调整图片的尺寸和角度时，首先要调用 Image 模块中的 im=Image.open("xxx.jpg") 语句打开指定的预处理图片，并调用 im.size 和 im.resize 记录并调整图片的尺寸，调用 im.rotate 语句调整图片的角度，最后调用 im.convert 实现图片格式的转换。

此实验部分实现代码如下

```
import Image
```

```
im = Image.open("messi.jpg" )
print im.size
width = 200
ratio = float(width)/im.size[0]
height = int(im.size[1]*ratio)
nim1 = im.resize( (width, height), Image.BILINEAR )
print nim1.size
nim1.save("resize.jpg" )
nim2 = im.rotate( 45, Image.BILINEAR )
nim2.save("rotated45.jpg" )
nim3 = im.rotate( 90, Image.BILINEAR )
nim3.save("rotated90.jpg" )
gray_img = im.convert("L")gray_img2=
gray_img.resize((128,128),Image.BILINEAR)
print gray_img2.histogram()
```

在该实验中，学生需用前述的图片处理方法对指定图片的尺寸和角度进行调整，并在此基础上，思考如何实现调整其他图片参数的方法，记录 Python 代码，
并分析实验结果。

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

1. 网络图片的爬取。（30 分）
2. 图片的尺寸和角度处理。（50 分）
3. 代码逻辑与规范。（20 分）

七、思考和练习

1. 图片处理的方式；
2. 网络爬取的图片能否修改格式。

实训十 使用 Python 统计分析社交数据

一、实训目的和要求

1. 强化 Python 程序的设计和编程能力；
2. 学习社交网络数据清洗和数据统计分析的方法；
3. 学习使用 Python 语言统计分析社交网络数据。

二、实训内容

1. 学习社交网络中联系人职位、公司、年龄等信息的数据清洗和统计分析方法；

2. 用 Python 语言中的 Counter、itemgetter 等模块对 LinkedIn 网站中联系人名单信息进行初步的统计分析。

三、实训准备

Python3、PyCharm2019 中文版。

四、实训步骤

数据清洗：

数据清洗是指发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等。由于数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把脏数据清除，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。

本实验中使用的数据来源是 Linkedin 网站中联系人信息，需要清洗的数据主要是由于数据名称不统一造成的。例如，联系人公司中很多都带有后缀 Inc.，Co.等，联系人职位中很多带有 Prof.,Dr.等，这些信息在统计时会有干扰作用。例如 IBMInc.和 IBM 代表的都是 IBM 公司，但程序在进行统计分析时会误认为是两个不同的公司。

部分实现代码如下：

```
fortransformintransforms:
companies[i]=companies[i].replace(*transform)
```

在该实验中，学生需用前述的数据清洗方法实现对 Linkedin 社交网络联系人信息进行数据清洗，并在此基础上，思考如何实现清洗其他社交网络(如新浪微博)联系人信息的方法，记录 Python 代码，并分析实验结果。

Python 语句分析清洗后的社交网站联系人信息时，首先要调用 Counter 模块语句将联系人相关信息进行统计，并调用 PrettyTable 模块将联系人信息存储在表中，最后调用 print 语句按照降序打印用户信息表。

此实验部分实现代码如下

```
pt=PrettyTable(field_names=['Company','Freq'])pt.align='c'
c=Counter(companies)
[pt.add_row([company,freq])
for(company,freq)insorted(c.items(),key=itemgetter(1),reverse=Tru
e)iffreq>0]
printpt
titles=[c['JobTitle'].strip()forcincontactsifc['JobTitle'].strip()
!='']
```

在该实验中，学生需用前述的统计分析方法分析 LinkedIn 联系人信息，并在此基础上，思考如何实现分析其他社交网络(微博)联系人信息的方法，记录 Python 代码，并分析实验结果。

五、实训方法

使用投影进行讲解演示，并上机进行练习。

六、考核办法

此部分实训内容采用全体考察的方法，以百分制为满分，具体评分标准如下：

设计社交网络数据清晰的程序分析实验要求（10 分）

清洗网络中的脏数据（20 分）

记录程序代码（30 分）

设计统计分析社交网络联系人信息的程序分析实验要求（20 分）

记录并分析实验结果（20 分）

七、思考和练习

1. 爬虫程序相关模块；
2. 爬虫工具有哪些。

实训十一　总结

一、 实训目的和要求

将学生制作的作品进行综合的考核，并进行总结。

二、 实训内容

1. 对学生作品进行考核。
2. 选择典型的（优秀的和劣质的）作品分别进行总结。

三、 实训准备

Python3、PyCharm2019 中文版。

四、 实训步骤

1. 对学生的作品依次进行综合考核。
2. 抽取典型（优秀和劣质）的作品进行全面的解析。

五、 实训方法

在机房利用 Python 编辑器完成。

六、 考核办法

此部分实训以考核为主，对学生组品进行整体的考核。采用全体考察的方法，以百分制为满分，具体评分标准如下：

系统文档　20 分

项目功能　40 分

程序调试　10 分

实训出勤　20 分

效果展示　10 分

七、思考与练习

无。

附录一　Python 开发环境搭建

本课程实验使用的 Python 开发环境为 Python IDLE，其用户界面图见图 1 所示。IDLE 是开发 python 程序的基本集成开发环境，具备基本的 IDE 的功能，是 Python 教学的不错的选择。当安装好 python 以后，IDLE 就自动安装好了，不需 要另外去找。同时，使用 Eclipse 这个强大的框架时 IDLE 也可以非常方便的调 试 Python 程序。其基本功能包括语法加亮、段落缩进、基本文本编辑、TABLE 键控制、调试程序。

打开 Idle 后出现一个增强的交互命令行解释器窗口（具有比基本的交互命令提示符更好的剪切、 粘贴、回行等功能）。除此之外，还有一个针对 Python 的 编辑器（无代码合并，但有语法标签高亮和代码自动完成功能）、类浏览器和调试器。菜单为 TK “剥离”式，也就是点击顶部任意下拉菜单的虚线将会将该 菜单提升到它自己的永久窗口中去。特别是 "Edit" 菜单，将其“靠”在桌面一角非常实用。Idle 的调试器提供断点、步进和变量监视功能。

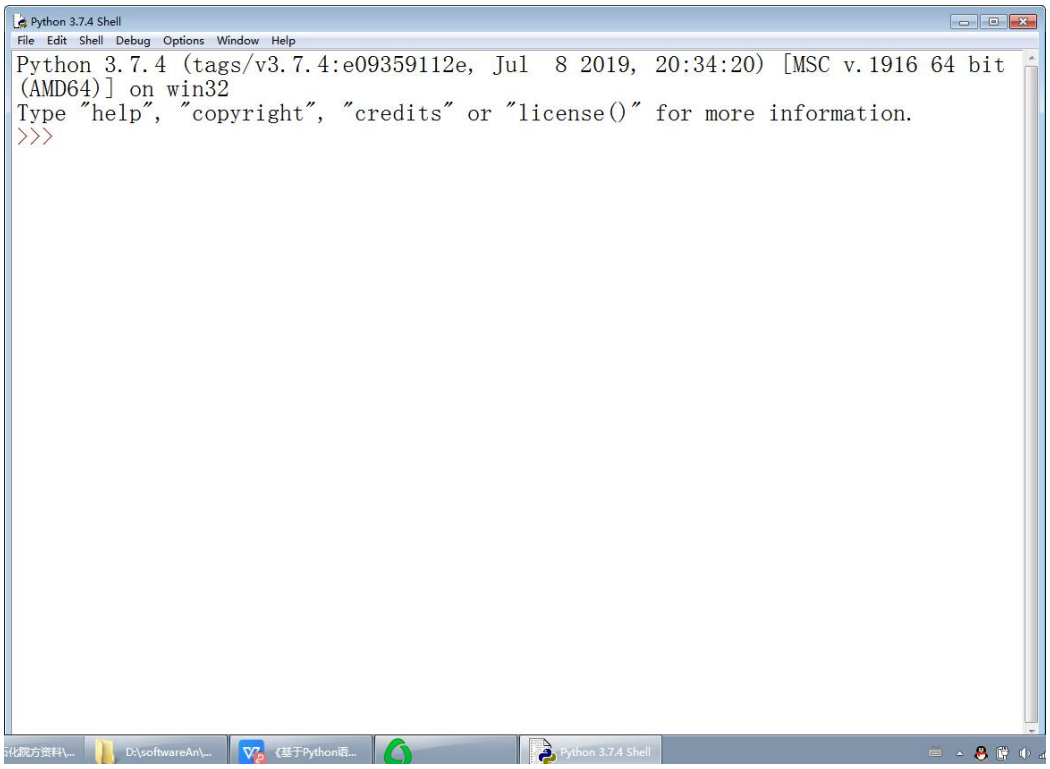


图 1 Python IDLE 界面图