数据爬取和数据分析

# 一、实验内容

Python一个重要的应用就是数据爬取，可以从互联网上轻松获取大量的数据用于数据分析。本案例使用Python爬取城市的历史天气，然后对天气进行分析。

# 二、实验步骤

1、检查实验中用到的库是否齐全，如未安装请先安装。

2、分析网站<http://www.tianqihoubao.com>中历史数据的网址规律，得到需要爬取的网址。

3、向服务器请求数据，将服务器返回的数据解析，并保存到文件中。

4、对数据进行分析（可考虑使用pandas库）。如某天的历史最高温和最低温是多少？该天晴天和下雨的概率是多少？哪天下雨的概率最高？哪天晴天的概率最高？最高温度和最低温度出现在哪天？

下面提供数据获取部分代码仅供参考

*# https://blog.csdn.net/weixin\_43327576/article/details/86514093*import requests  
from requests.exceptions import RequestException  
from bs4 import BeautifulSoup  
import os  
import csv  
import time  
  
def get\_one\_page(url):  
 *'''  
 获取网页  
 '''* print(**'正在加载'**+url)  
 headers={**'User-Agent'**:**'User-Agent:Mozilla/5.0'**}  
 try:  
 response = requests.get(url,headers=headers)  
 if response.status\_code == 200:  
 return response.content  
 return None  
 except RequestException:  
 return None  
  
def parse\_one\_page(html):  
 *'''  
 对网页内容进行解析  
 '''* soup = BeautifulSoup(html, **"lxml"**)  
 info = soup.find(**'div'**, class\_=**'wdetail'**)  
 rows=[]  
 tr\_list = info.find\_all(**'tr'**)[1:] *# 使用从第二个tr开始取* for index, tr in enumerate(tr\_list): *# enumerate可以返回元素的位置及内容* td\_list = tr.find\_all(**'td'**)  
 date = td\_list[0].text.strip().replace(**"**\n**"**, **""**) *# 取每个标签的text信息，并使用replace()函数将换行符删除* weather = td\_list[1].text.strip().replace(**"**\n**"**, **""**).split(**"/"**)[0].strip()  
 temperature\_high = td\_list[2].text.strip().replace(**"**\n**"**, **""**).split(**"/"**)[0].strip()  
 temperature\_low = td\_list[2].text.strip().replace(**"**\n**"**, **""**).split(**"/"**)[1].strip()  
  
 rows.append((date,weather,temperature\_high,temperature\_low))  
 return rows  
  
*#爬取的城市拼音名称，实验中可只选取一个城市*cities = [**'chengdu'**,**'aba'**,**'bazhong'**,**'dazhou'**,**'deyang'**,**'ganzi'**,**'guangan'**,  
 **'guangyuan'**,**'leshan'**,**'luzhou'**,**'meishan'**,**'mianyang'**,**'neijiang'**,**'nanchong'**,  
 **'panzhihua'**,**'scsuining'**,**'yaan'**,**'yibin'**,**'ziyang'**,**'zigong'**,**'liangshan'**]  
years = [**'2012'**,**'2013'**,**'2014'**,**'2015'**,**'2016'**,**'2017'**,**'2018'**]  
months = [**'01'**,**'02'**,**'03'**,**'04'**,**'05'**,**'06'**, **'07'**, **'08'**,**'09'**,**'10'**,**'11'**,**'12'**]  
  
if \_\_name\_\_ == **'\_\_main\_\_'**:  
 *# os.chdir() # 设置工作路径* for city in cities:  
 with open(city + **'\_weather.csv'**, **'a'**, newline=**''**) as f:  
 writer = csv.writer(f)  
 writer.writerow([**'date'**,**'weather'**,**'temperature\_high'**,**'temperature\_low'**])  
 for year in years:  
 for month in months:  
 url = **'http://www.tianqihoubao.com/lishi/'**+city+**'/month/'**+year+month+**'.html'** html = get\_one\_page(url)  
 content=parse\_one\_page(html)  
 writer.writerows(content)  
 print(city+year+month+**' is OK!'**)  
 time.sleep(2)