太湖创意职业技术学院

**实 验（实 训）报 告**

**实验（实训）名称 python大数据**

**所属课程名称 python大数据**

**班 级**  计算机应用202

**学 号**  12042034

**姓 名**  马云骥

**指导教师**  李成渊

**项目名称：**数据获取与存储  **日期：**2022-04-28

|  |
| --- |
| **一、实验（实训）概述：** |
| **【目的及要求】**  1.掌握Python的基本编程语法。  2.利用Python编写一段小程序。  **【实施环境】（使用的材料、设备、软件）**  **设备：**计算机 、Pycharm |
| **二、实验（实训）内容：** |
| **【实验内容】**  指定一个网站（例如中国天气网站http://www.weather.com.cn/weather/101280601.shtml），利用多线程控制爬取这个网站中的所有图像文件，同时把这些文件保存到程序所在文件夹的子文件夹中，分别对深圳、无锡、上海三个城市完成图片爬取任务，并分别存储在image-sz、image-wx、image-sh文件夹中。  **【实验（实训）过程】（步骤、记录、数据、程序等）**  **程序代码**  from bs4 import BeautifulSoup from bs4 import UnicodeDammit import urllib.request import threading ​ ​ def imageSpider(start\_url, city):     global count     global threads     try:         urls = []         req = urllib.request.Request(start\_url, headers=headers)         data = urllib.request.urlopen(req)         data = data.read()         dammit = UnicodeDammit(data, ["utf-8", "gbk"])         data = dammit.unicode\_markup         soup = BeautifulSoup(data, "lxml")         images = soup.select("img")         for image in images:             try:                 src = image["src"]                 url = urllib.request.urljoin(start\_url, src)                 if url not in urls:                     print(url)                     count = count + 1                     T = threading.Thread(target=download, args=(url, count, city))                     T.setDaemon(False)                     T.start()                     threads.append(T)             except Exception as err:                 print(err)     except Exception as err:         print(err) ​ ​ def download(url, count, city):     try:         if (url[len(url) - 4] == "."):             ext = url[len(url) - 4:]         else:             ext = ""         req = urllib.request.Request(url, headers=headers)         data = urllib.request.urlopen(req, timeout=100)         data = data.read()         fobj = open("image-" + str(city) + "\\" + str(count) + ext, "wb")         fobj.write(data)         fobj.close()         print("downloaded " + str(count) + ext)     except Exception as err:         print(err) ​ ​ def process(city):     global headers     global count     global threads     cityCode = {"sz": "101280601", "wx": "101190201", "sh": "101020100"}     start\_url = "http://www.weather.com.cn/weather/" + cityCode[city] + ".shtml"     headers = {         "User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/15.4 Safari/605.1.15"}     count = 0     threads = []     imageSpider(start\_url, city)     for t in threads:         t.join()     print("The process of " + str(city) + " has END") ​ ​ process("sz") process("wx") process("sh") print("The Final END")  我们需要定义三个函数   * process()   + 通过输入城市缩写，和函数内存放城市代码的数组 cityCode 一起定义了需要爬取图片的网站url   + 定义用户代理 User-Agent   + 调用函数 imageSpider() 来爬取图片 * imageSpider()   + 负责爬取目标网站中的图片   + 调用函数 download() 对爬取到的图片进行保存 * download()   + 将爬取到的图片下载下来并存放在相应文件夹中   **【结论】（结果、分析）**  **运行程序**  Windows下成功运行程序  可以看到我们爬取到的图片都已存放在了提前新建好的对应文件夹中  Windows下程序运行的效果（标注）  需要注意的是，本程序仅且仅能在 Windows 环境下正常运行。  因为 Windows 使用 \ 作为路径分隔符，而 Linux 和 macOS 使用的是符号 / ，如果不对程序源代码进行修改的话，爬取到的图片会错误地以 image-xx\xx.png 的命名方式直接存放在程序根目录下。  ubuntu下程序运行的效果  ubuntu 下程序错误运行的效果  macOS下程序运行的效果  macOS 下程序错误运行的效果  解决办法也很简单，只要对第45行代码稍作修改即可。  ... # fobj = open("image-" + str(city) + "\\" + str(count) + ext, "wb") fobj = open("image-" + str(city) + "//" + str(count) + ext, "wb") ... |
| **三、指导教师评语及成绩：** |
| **评语：**  该同学基本掌握Python的基本编程语法，能利用Python编写一段小程序，完成爬取天气预报数据的功能，达到了实训的要求。  **成绩：** A    **批阅日期：**2022-5-2 |