爬取宜昌公交车信息

**1．前言**

**1.1 课题描述**

爬取宜昌公交车信息（<https://yichang.8684.cn>）：公交车名称（lineName），运行时间（time），票价信息（price），所属公司（campony），往返线路（upline和downline），并将其保存在bus\_messages.csv文件中。

**1.2 课题背景及意义**

练习爬虫能力，获取宜昌公交车线路的详细信息。

**1.3 相关技术介绍**

**Jupyter notebook：**Jupyter Notebook是基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。

**xPath helper：**一款Chrome浏览器的开发者插件，安装了xPath helper后就能轻松获取HTML元素的xPath。

本次项目主要使用xPath helper来进行xPath内容查看。

**request**：一个Python第三方库，处理URL资源特别方便。Requests 继承了urllib2的所有特性。Requests支持HTTP连接保持和连接池，支持使用cookie保持会话，支持文件上传，支持自动确定响应内容的编码，支持国际化的 URL 和 POST 数据自动编码。

**lxml**：一个Python第三方库，使用它可以轻松处理XML和HTML文件，还可以用于web爬取。库的主要优点是易于使用，在解析大型文档时速度非常快，归档的也非常好，并且提供了简单的转换方法来将数据转换为Python数据类型，从而使文件操作更容易。

本次项目主要使用lxml.etree来进行XML信息处理。

**time**：Python库，可以用于格式化日期和时间。

本次项目主要使用time.sleep来设置时间间隔爬取。

**re**：Python库，用于正则匹配。

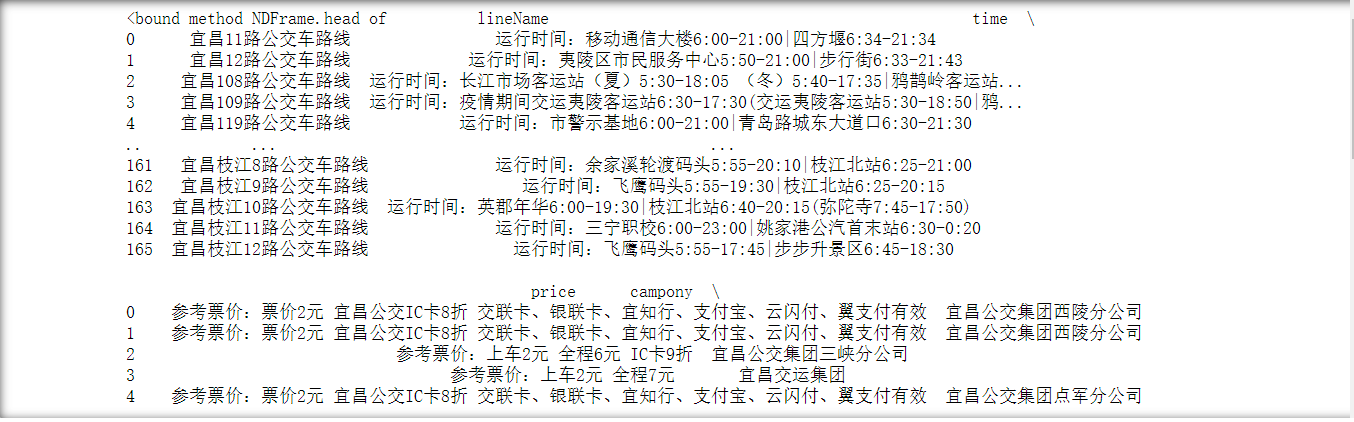
**pandas**：Python第三方库，它使用强大的数据结构提供高性能的数据操作。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas提供了大量能使我们快速便捷地处理数据的函数和方法。

本次项目主要使用pandas.DataFrame来完成结果的保存。

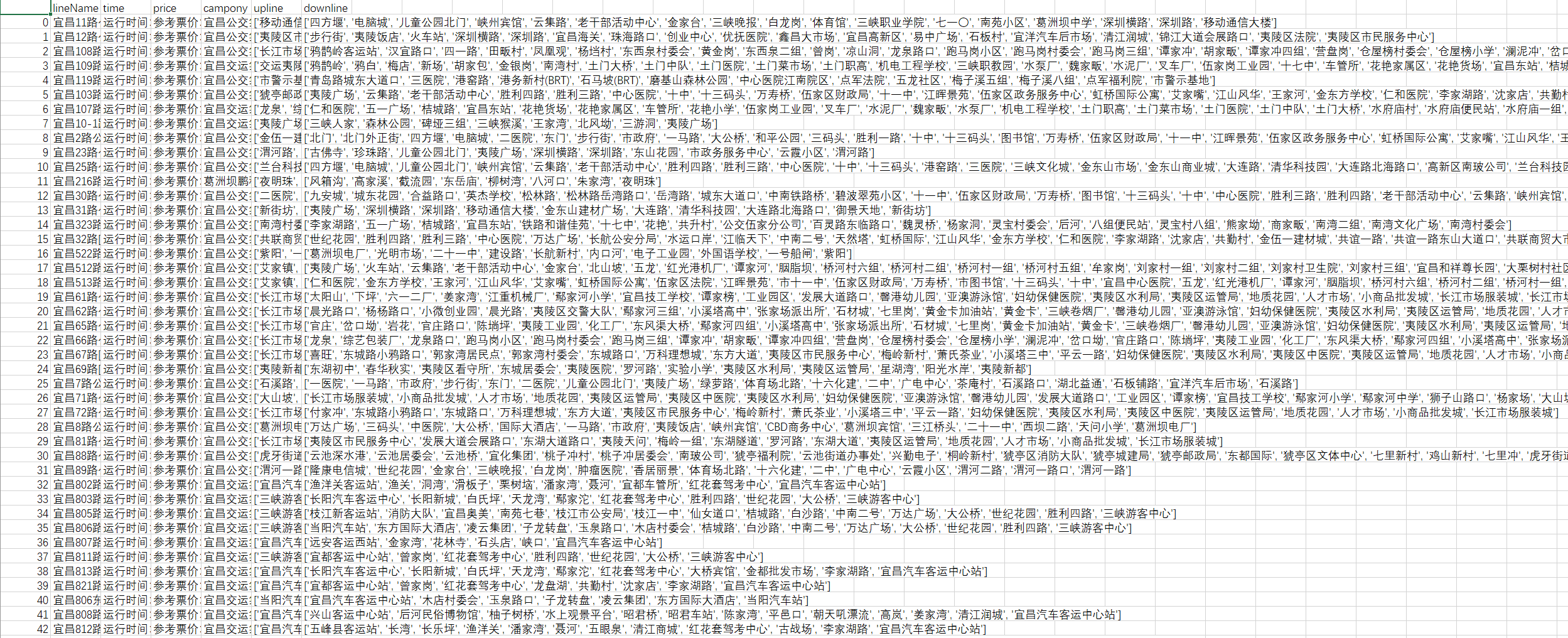
**2.系统分析**

可以做到爬取宜昌公交车信息（<https://yichang.8684.cn>）：公交车名称（lineName），运行时间（time），票价信息（price），所属公司（campony），往返线路（upline和downline），并将其保存在bus\_messages.csv文件中。

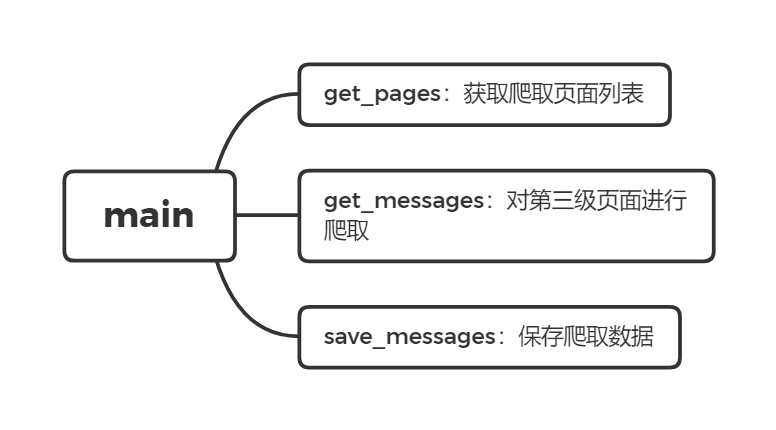
**输出结果的图片（部分）**：



**保存的csv的图片（部分）**：

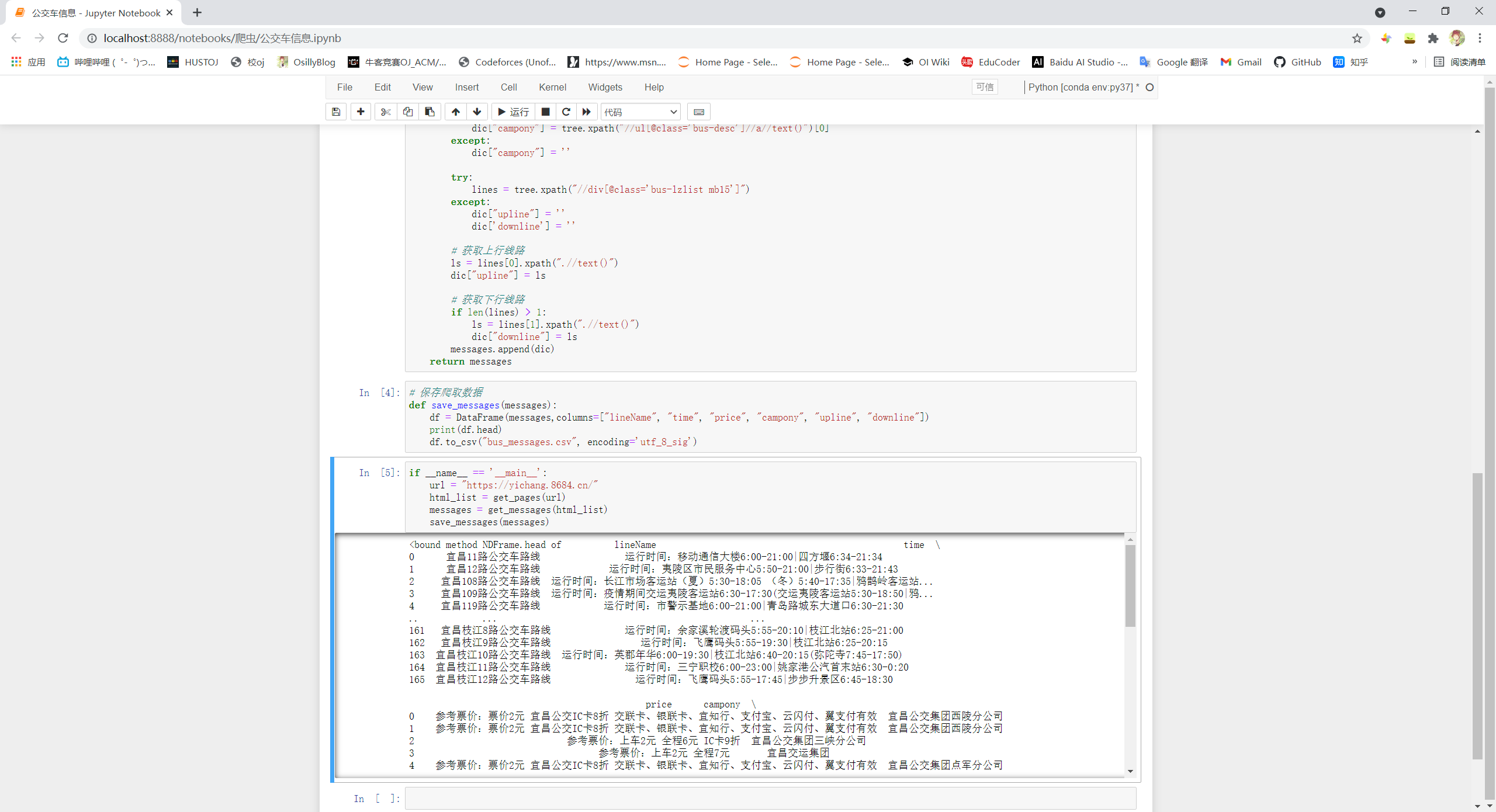


**3.系统设计**



**4.系统实现**

**4.1 运行界面**



**4.2 代码实现**

*#!/usr/bin/env python*   
*# coding: utf-8*   
   
*# In[1]:*   
   
   
import requests   
from lxml import etree   
from time import sleep   
import re   
import pandas as pd   
from pandas import DataFrame, Series   
   
headers = {   
 'Accept': '\*/\*',   
 'Accept-Language': 'en-US,en;q=0.8',   
 'Cache-Control': 'max-age=0',   
 'User-Agent': 'Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.116 Safari/537.36',   
 'Connection': 'keep-alive',   
 'Referer': 'http://www.baidu.com/'   
}   
   
   
*# In[2]:*   
   
   
*# 获取爬取页面列表*   
**def** get\_pages(url):   
 second\_pages = []   
 third\_pages = []   
 html\_list = []   
 res = requests.get(url=url,headers=headers)   
 pat = re.compile(r'/list[0-9A-Z]')   
 ls = pat.findall(res.text)   
 **for** next\_url **in** ls:   
 second\_pages.append(url + next\_url)   
*# print(second\_pages)*   
   
 **for** page **in** second\_pages:   
 res = requests.get(url=page,headers=headers)   
 tree = etree.HTML(res.text)   
 line\_list = tree.xpath("//div[@class='list clearfix']/a/@href")   
 **for** line **in** line\_list:   
 third\_pages.append(url+line)   
*# print(third\_pages)*   
   
 **for** page **in** third\_pages:   
 html = requests.get(url=page,headers=headers)   
 sleep(0.5)   
 html\_list.append(html.text)   
*# print(html\_list)*   
 **return** html\_list   
   
   
*# In[3]:*   
   
   
*# 对第三级页面进行爬取*   
**def** get\_messages(html\_list):   
 messages = []   
 **for** html **in** html\_list:   
 tree = etree.HTML(html)   
 dic = {}   
 **try**:   
 dic["lineName"] = tree.xpath("//h1//text()")[0]   
 **except**:   
 dic["lineName"] = ''   
 **try**:   
 dic["time"] = tree.xpath("//ul[@class='bus-desc']//text()")[0]   
 **except**:   
 dic["time"] = ''   
 **try**:   
 dic["price"] = tree.xpath("//ul[@class='bus-desc']//text()")[1]   
 **except**:   
 dic["price"] = ''   
 **try**:   
 dic["campony"] = tree.xpath("//ul[@class='bus-desc']//a//text()")[0]   
 **except**:   
 dic["campony"] = ''   
   
 **try**:   
 lines = tree.xpath("//div[@class='bus-lzlist mb15']")   
 **except**:   
 dic["upline"] = ''   
 dic['downline'] = ''   
 *# 获取上行线路*  
 ls = lines[0].xpath(".//text()")   
 dic["upline"] = ls   
 *# 获取下行线路*   
 **if** len(lines) > 1:   
 ls = lines[1].xpath(".//text()")   
 dic["downline"] = ls   
 messages.append(dic)   
 **return** messages   
   
   
*# In[4]:*   
   
   
*# 保存爬取数据*   
**def** save\_messages(messages):   
 df = DataFrame(messages,columns=["lineName", "time", "price", "campony", "upline", "downline"])   
 print(df.head)   
 df.to\_csv("bus\_messages.csv", encoding='utf\_8\_sig')   
   
   
*# In[5]:*   
   
   
**if** \_\_name\_\_ == '\_\_main\_\_':   
 url = "https://yichang.8684.cn/"   
 html\_list = get\_pages(url)   
 messages = get\_messages(html\_list)   
 save\_messages(messages)

**5.结束语**

对爬虫有了新的了解，学会了一些基本爬虫操作