Python-based Design and Implementation of Meteorological Data Acquisition

Ying Yuan

Yinchuan University of Energy, Ningxia, Yinchuan, 750105, China

Keywords: Meteorological Data, Python, Web Crawler.

Abstract: In the era of big data, people can gain all kinds of meteorological information through different channels of information. In the face of ever-increasing diversified meteorological data, higher requirements are put forward for improving precision and timeliness of meteorological data acquisition. Under such background, Python language is applied to develop a meteorological data acquisition program which can define the area to crawl the data from www.nmc.cn. The program can capture key meteorological data and agricultural production guidance information in real time in different cities. Besides, the crawled data are saved locally, and displayed through graphical user interface. In this way, the meteorological data acquired can be exhibited to the public in an efficient and vivid way, which carries positive significance for improving weather forecast.

1. Introduction

The rapid development of internet causes the internet data become massive and diversified. People cannot gain as much data as possible that users are interested in on one page. Thus, crawling more data that users are interested in on different webpages has become a trend. In daily life, people can gain all kinds of meteorological information through diverse channels of information, but the meteorological information that users are interested in only accounts for a small part. This study aims to crawl the desired meteorological information from the target website page to improve the efficiency of gaining useful meteorological information.

The content of this study is how to crawl the desired meteorological information from the target website page. At present, Python as a simple and efficient programming language is favored by more and more programmers. In this paper, Python is chosen as the programming language of data crawling, and complex data acquisition is achieved through utilizing all kinds of rich libraries and analyzing the key codes in the website. Since many websites own anti-crawler technology, this design bypasses the website protection through pretending as a browser. In the aspect of data acquisition, except capture of key meteorological data information in different cities, agricultural production guidance information provided by www.nmc.cn can be captured. The scientific suggestions on spring sowing, autumn harvest, irrigation and manuring can be provided fast to help agricultural workers acquire real-time meteorological data fast and adjust agricultural production and life style in time.

2. Introduction to Key Technology

2.1. Python language

Python is an object-oriented interpretive computer programming language. It is simple and easy to learn, with strong functions, free open source and cross multi-platform use. Compared with traditional programming languages, Python is executable just with the interpreter, with the advantages of fast computing speed and high memory management efficiency. Hence, it is widely applied in the fields of web development, game script, web crawler and data analysis, etc.

2.2. Web crawler

The data in internet are stored in dynamic webpages. The program will capture the specified data through the set crawling rules, just as a spider captures its preys along the spider's web. Thus, we call this program web crawler. The principle is that some sites are accessed by our crawler program

DOI: 10.25236/scmc.2019.016

which simulates the browser, and crawl HTML code/JSON data/ binary data (picture and video) returned by the sites to local place, thus extracting the desired data and storing them.

To be specific, four steps are involved:

① Initiate a request: Apply Http library to initiate a request to the target site. This request is a packed Request, including request header and request body, etc.

② Acquire site content and response: If the server can work and respond normally, we will acquire a packed Response which contains various pictures, videos, webpage codes, and Json data, etc.

③ Analyze contents: The methods to analyze Html data include several different ways. The first one is to customize the analysis format. The regular expression can be utilized to gain corresponding contents, and Re module is imported in the file. The second one is to utilize the third-party analysis libraries such as Beautifulsoup, Urllib and Xpath to analyze webpages.

④ Save data: Python provides the method to save data. The file form may be used to save data. If the data are massive, the data can be saved in the database (including MySQL and MongDB).

2.3. Requests library

Requests library is a very convenient Python HTTP library, and it will be basically used during writing crawler projects. Since it is based on Urllib3, it is unnecessary to add the query string manually for URL in the application process. Meanwhile, form coding is not required for Post data. The functions of Keep-alive and Http connection pool are of full automation. Besides, the built-in Json decoder exempts form Json data format conversion, which is convenient for developers to acquire the data on Web.

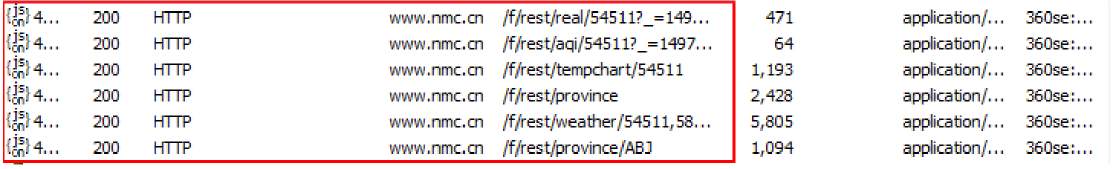
2.4. Remodule and regular expression

The major function of Re module is to match character strings in a dynamic and fuzzy way. The regular expression can match a series of character strings with similar features according to a mode. In the replacement and modification process of character strings, the regular expression also has the function of extracting character strings, except simply judging whether the character strings are matched.

3. Detailed Design and Implementation

3.1. URL construction strategy of target website

The current crawler can bring us a lot of development experience and matters needing attention. In this study, the target website is www.nmc.cn. In accordance with the characteristics of the website, we designed automatic URL construction for the website. Through the analysis, we find that the targets to be inquired are displayed in the last item of URL, and the backstage package capture is shown in Fig.1. When Json file is opened, messy codes are shown. Then, we need Json to decode. After the data saved are screened by the regular expression, the data can be saved locally.



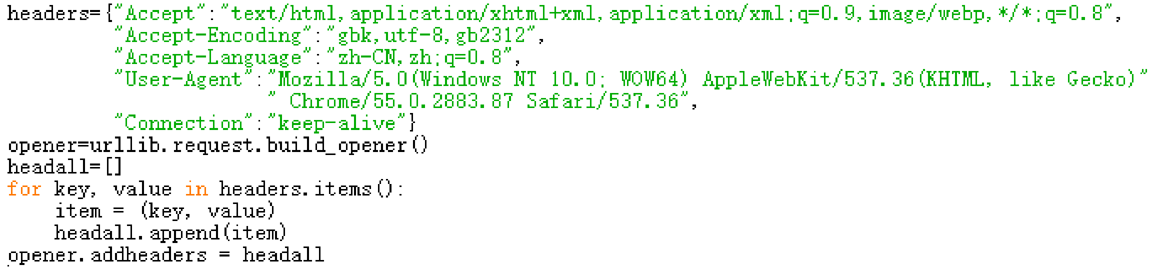
```
4...  200  HTTP  www.nmc.cn  /f/rest/real/54511?_=149...   471    application/...  360se:...
4...  200  HTTP  www.nmc.cn  /f/rest/aqi/54511?_=1497...   64     application/...  360se:...
4...  200  HTTP  www.nmc.cn  /f/rest/tempchart/54511       1,193  application/...  360se:...
4...  200  HTTP  www.nmc.cn  /f/rest/province              2,428  application/...  360se:...
4...  200  HTTP  www.nmc.cn  /f/rest/weather/54511,58...   5,805  application/...  360se:...
4...  200  HTTP  www.nmc.cn  /f/rest/province/ABJ          1,094  application/...  360se:...
```

Fig.1. Browser package capture

3.2. URL analysis and browser simulation

To prevent the target website from owing anti-pickpocket function, the single IP test shows that the target website does not add IP anti-pickpocket function. We can just use the simple agent information. We add agent information (USER_AGENT), and header information (DEFAULT_REQUEST_HEADERS). The two pieces of information will be included in the header file requested by Http and sent to the server of the target website. Thus, the webpage code of the page can be gained.

Http cache mechanism is opened. During sending Http request, it is no longer necessary to read configuration header information in the configuration file, and Http header information in cache can be directly read, thus reducing I/O overhead, and improving crawling speed. The codes are shown in Fig.2.

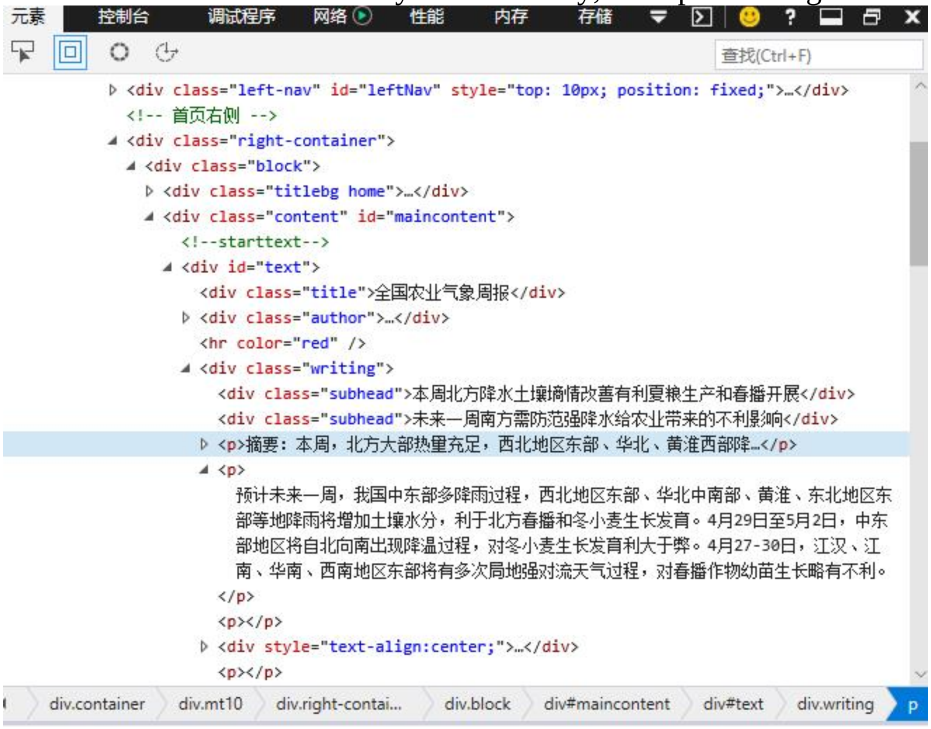


```
headers={"Accept":"text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
         "Accept-Encoding":"gbk,utf-8,gb2312",
         "Accept-Language":"zh-CN,zh;q=0.8",
         "User-Agent":"Mozilla/5.0(Windows NT 10.0; WOW64) AppleWebKit/537.36(KHTML, like Gecko)"
                     " Chrome/55.0.2883.87 Safari/537.36",
         "Connection":"keep-alive"}
opener=urllib.request.build_opener()
headall=[]
for key, value in headers.items():
    item = (key, value)
    headall.append(item)
opener.addheaders = headall
```

Fig.2. Codes of analog browser

3.3. Information matching and retrieval

After URL change rules of the target website are confirmed, the position of contents to be crawled on the website is analyzed in the next step. Open the target webpage, press F12, then analyze the source code in the page and find CSS code of the target information. It is supposed the target is the information of national weekly agrometeorological newspaper. The corresponding CSS codes of the information are shown in Fig.3. Then, the corresponding elements and Class names of the information to be extracted are locked. By the same way, code positioning is conducted.

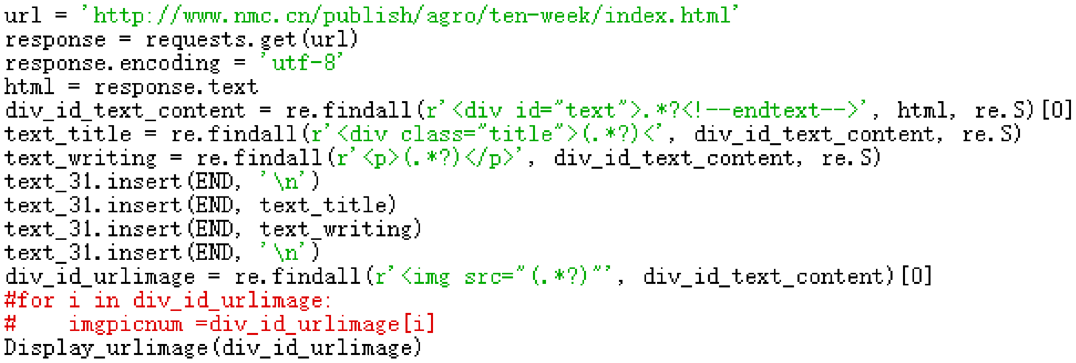


```
▷ <div class="left-nav" id="leftNav" style="top: 10px; position: fixed;">…</div>
  <!-- 首页右侧 -->
◢ <div class="right-container">
  ◢ <div class="block">
    ▷ <div class="titlebg home">…</div>
    ◢ <div class="content" id="maincontent">
        <!--starttext-->
      ◢ <div id="text">
          <div class="title">全国农业气象周报</div>
        ▷ <div class="author">…</div>
          <hr color="red" />
        ◢ <div class="writing">
            <div class="subhead">本周北方降水土壤墒情改善有利夏粮生产和春播开展</div>
            <div class="subhead">未来一周南方需防范强降水给农业带来的不利影响</div>
          ▷ <p>摘要：本周，北方大部热量充足，西北地区东部、华北、黄淮西部降…</p>
          ◢ <p>
              预计未来一周，我国中东部多降雨过程，西北地区东部、华北中南部、黄淮、东北地区东
              部等地降雨将增加土壤水分，利于北方春播和冬小麦生长发育。4月29日至5月2日，中东
              部地区将自北向南出现降温过程，对冬小麦生长发育利大于弊。4月27-30日，江汉、江
              南、华南、西南地区东部将有多次局地强对流天气过程，对春播作物幼苗生长略有不利。
            </p>
            <p></p>
          ▷ <div style="text-align:center;">…</div>
            <p></p>
```

Fig.3. Source code of webpage

3.4. Acquisition of webpage source code and extraction of information required from source code

Requests library offers a series of functions to operate URL so that we can read the data on www and ftp like reading local files. It is suitable for extracting the required element information which is directly saved in the source code of the page. requests.get(url).text can be directly used to extract page source code. Then, re module (regular expression) is introduced. The regular expression can be used to filter useless characters to prevent interference. For example, the codes to gain the information of weekly agrometeorological newspaper are shown in Fig.4.



```
url = 'http://www.nmc.cn/publish/agro/ten-week/index.html'
response = requests.get(url)
response.encoding = 'utf-8'
html = response.text
div_id_text_content = re.findall(r'<div id="text">.*?<!--endtext-->', html, re.S)[0]
text_title = re.findall(r'<div class="title">(.*?)<', div_id_text_content, re.S)
text_writing = re.findall(r'<p>(.*?)</p>', div_id_text_content, re.S)
text_31.insert(END, '\n')
text_31.insert(END, text_title)
text_31.insert(END, text_writing)
text_31.insert(END, '\n')
div_id_urlimage = re.findall(r'<img src="(.*?)"', div_id_text_content)[0]
#for i in div_id_urlimage:
#    imgpicnum =div_id_urlimage[i]
Display_urlimage(div_id_urlimage)
```

Fig.4. Codes to gain the information of weekly agrometeorological newspaper

3.5. Crawling result storage and display

The crawling results should be collated in time. After the data to be crawled are extracted from the webpage, the data need to be saved. In this design, the meteorological data crawled are saved in local Excel file.

To better display the crawling results, Python's tkinter module is utilized to design the operable graphical user interface (GUI). The display content contains real-time meteorological data, agricultural information guidance and system operation.

Users input different areas and acquire real-time meteorological data of different areas. The design supports the information types gained. The interface is shown in Fig.5.

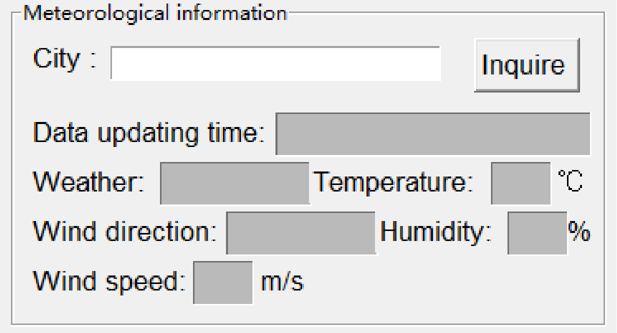


Fig.5. Query interface of city meteorological data

The design can dynamically capture various kinds of agrometeorological information, as shown in Fig.6. The information can bring scientific and reasonable guidance and suggestions for agricultural workers.

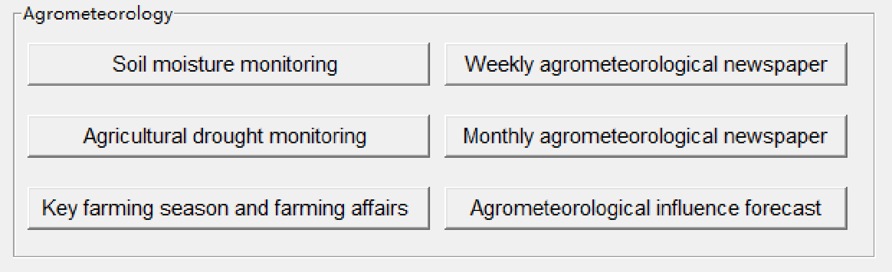
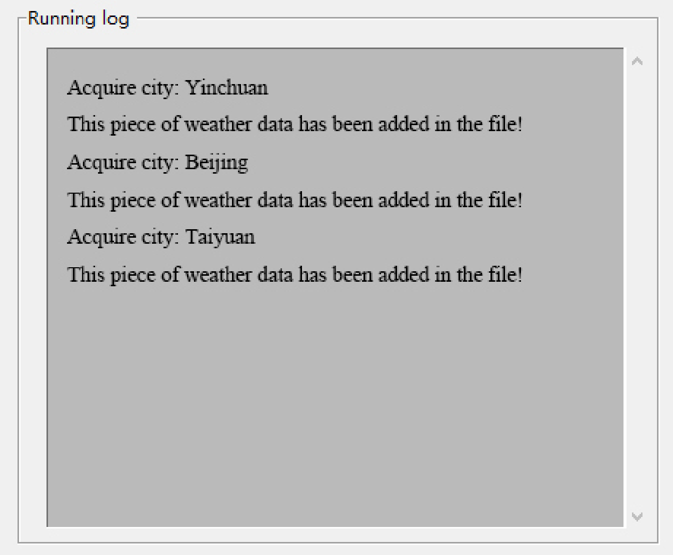

Agrometeorology
Soil moisture monitoring
Weekly agrometeorological newspaper
Agricultural drought monitoring
Monthly agrometeorological newspaper
Key farming season and farming affairs
Agrometeorological influence forecast

Fig.6. Query interface of Agrometeorology

Running log window can display user's operations in real time. In case of operation error, the prompt will be given. If the operation is correct, user's operations will be recorded. The running log interface is shown in Fig.7.



```
Acquire city: Yinchuan
This piece of weather data has been added in the file!
Acquire city: Beijing
This piece of weather data has been added in the file!
Acquire city: Taiyuan
This piece of weather data has been added in the file!
```

Fig.7. Running log interface

4. Summary

It is known from the study that the abundant standard libraries and fast development advantage of Python can be utilized to achieve crawling and display of meteorological data of www.nmc.cn. Not only is the vivid, real-time and effective meteorological information acquired, but also the data support is provided for follow-up data analysis.

Acknowledgement

The study was supported by Higher education and teaching reform project of the Ningxia Hui Autonomous Region (Grant No.NXJG2016092)

References

[1] Gong Weiwei, Qi Xiangchun, Pei Shikang. Research and Application of Python and R Language Hybrid Programming. Computer Applications and Software, 2018 (1) 28-31

[2] Yan Fei, Xiap Pu. Research and Implementation of Theme-based Data Crawling Technology with Python Framework. Computer Era, 2018 (11) 10-13.

[3] Yang Fan, Dong Jun, Tang Hongliang, Zhang Hao. Research of Taobao Comment Crawling Technology Based on Python. China Management Informationization, 2019 (4) 162-163.

[4] Guo Lirong. Python-based Web Crawler Programming. Electronic Technology & Software Engineering, 2017 (23) 248-249.

[5] Chen Lin, Ren Fang. Crawler Programming for Sina Microblog Data Based on Python. Information System Engineering, 2016 (09) 97-99.

[6] Du Xiaoxu, Jia Xiaoyun. Sina Microblog Crawler Analysis Based on Python. Software, 2019 40 (04) 182-185.