

Python专题教程：抓取网站，模拟登陆，抓取动态网页

版本：**v1.0**

Crifan Li

摘要

本文是针对Python的中级开发人员，介绍如何用Python语言去实现抓取网站，模拟登陆，抓取动态网页。其中主要涉及到，网络处理方面的模块（urllib，urllib2等），以及HTML解析相关的模块（BeautifulSoup，json等）。



本文提供多种格式供：

在线阅读	HTML ¹	HTMLs ²	PDF ³	CHM ⁴	TXT ⁵	RTF ⁶	WEBHELP ⁷
下载（7zip压缩包）	HTML ⁸	HTMLs ⁹	PDF ¹⁰	CHM ¹¹	TXT ¹²	RTF ¹³	WEBHELP ¹⁴

HTML版本的在线地址为：

http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/html/python_topic_web_scrape.html

有任何意见，建议，提交bug等，都欢迎去讨论组发帖讨论：

http://www.crifan.com/bbs/categories/python_topic_web_scrape/

修订历史

修订 1.0	2013-02-06	crl
1. 把之前教程的地址整理过来		

¹ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/html/python_topic_web_scrape.html

² http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/htmls/index.html

³ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/pdf/python_topic_web_scrape.pdf

⁴ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/chm/python_topic_web_scrape.chm

⁵ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/txt/python_topic_web_scrape.txt

⁶ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/rtf/python_topic_web_scrape.rtf

⁷ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/webhelp/index.html

⁸ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/html/python_topic_web_scrape.html.7z

⁹ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/htmls/index.html.7z

¹⁰ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/pdf/python_topic_web_scrape.pdf.7z

¹¹ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/chm/python_topic_web_scrape.chm.7z

¹² http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/txt/python_topic_web_scrape.txt.7z

¹³ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/rtf/python_topic_web_scrape.rtf.7z

¹⁴ http://www.crifan.com/files/doc/docbook/python_topic_web_scrape/release/webhelp/python_topic_web_scrape.webhelp.7z

Python专题教程：抓取网站，模拟登陆，抓取动态网页:

Crifan Li

版本：v1.0

出版日期 2013-02-06

版权 © 2013 Crifan, <http://crifan.com>

本文章遵从：[署名-非商业性使用 2.5 中国大陆\(CC BY-NC 2.5\)](#)¹⁵

¹⁵ http://www.crifan.com/files/doc/docbook/soft_dev_basic/release/html/soft_dev_basic.html#cc_by_nc

目录

前言	iv
1. 本文目的	iv
2. 前提	iv
1. 如何用Python实现网站抓取，模拟登陆，抓取动态网页	1
2. Python中的网络处理	2
3. Python中的HTML解析	3
参考书目	4

前言

1. 本文目的

本文目的在于，在已经了解了抓取网站，模拟登陆，抓取动态网页方面的逻辑后，如何用Python语言去实现这部分的逻辑。

2. 前提

讨论如何用Python去实现，网站抓取，模拟登陆，抓取动态网页的话，前提是你需要对这部分的逻辑已经比较清楚了。

如果不清楚，请先去参考：

[详解抓取网站，模拟登陆，抓取动态网页的原理和实现（Python，C#等）](http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/html/web_scrape_emulate_login.html)¹

¹ http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/html/web_scrape_emulate_login.html

第 1 章 如何用Python实现网站抓取，模拟登陆，抓取动态网页



相关旧帖

[【教程】抓取网并提取网页中所需要的信息之 Python版](#) ¹

[【教程】模拟登陆网站之 Python版（内含两种版本的完整的可运行的代码）](#) ²

其实，对于urllib等库，已经做得够好了，尤其是易用性上，已经很方便使用了。

比如，直接可以通过如下代码，即可获得从网页的地址，而得到其网页的源代码了

TODO：add code

但是呢，由于实际上，和网页抓取，网页模拟登陆等方面，需要用到cookie，以及其他header参数，导致想要获得一个，功能强大且好用的，用于网络抓取方面的函数，则还是需要额外花很多功夫的

而我后来就是在折腾网络抓取方面，前前后后，经过实际使用而积累出来很多这方面的经验，最终，写了个相关的，功能更加强大一些，更加方便使用的函数的。主要是2个函数：

getUrlResponse和getUrlRespHtml

TODO：添加两个函数来自crifanLib的解释

TODO：再添加这两个函数的几种用法

TODO：再添加另外几个相关的函数的解释，包括downloadFile等函数

其实主要分两大方面：

一方面是把网站的内容抓取下来，涉及到和网络处理方面的模块

另外一方面是如何解析抓取下来的内容，即涉及到HTML解析等方面的模块

下面就来解释这两大方面相关的逻辑，以及如何用Python实现对应的这部分的功能。

¹ http://www.crifan.com/crawl_website_html_and_extract_info_using_python/

² http://www.crifan.com/emulate_login_website_using_python/

第 2 章 Python中的网络处理

主要涉及的一些，和网络处理方面有关的模块是，urllib，urllib2等



相关旧帖

[【整理】Python中用于解析Http数据包的模块/库](#)¹

[【已解决】Python中使用cookielib的FileCookieJar去save\(\)，结果出错：NotImplementedError](#)²

[【整理】Python中Cookie的处理：自动处理Cookie，保存为Cookie文件，从文件载入Cookie](#)³

TODO：整理对应的，带urllib和urllib2方面的帖子进来。

¹ http://www.crifan.com/python_http_package_parser_lib_module

² http://www.crifan.com/python_cookiejar_filecookiejar_save_error_notimplementederror

³ http://www.crifan.com/python_auto_handle_cookie_and_save_to_from_cookie_file

第 3 章 Python中的HTML解析



相关旧帖

[BeautifulSoup模块简介](#)¹

[【教程】Python中第三方的用于解析HTML的库：BeautifulSoup](#)²

[【总结】Python的第三方库BeautifulSoup的使用心得](#)³

[【整理】关于Python中的html处理库函数BeautifulSoup使用注意事项](#)⁴

[【已解决】用BeautifulSoup解析Html格式的Json字符串](#)⁵

[【经验记录】Python中json.loads的时候出错->要注意要解码的Json字符的编码](#)⁶

[【已解决】Python中json.loads解析包含\n的字符串会出错](#)⁷

[【已解决】Python中使用json.loads解码字符串时出错：ValueError: Expecting property name: line 1 column 1 \(char 1\)](#)⁸

[【已解决】Python中用json.loads解码字符串出错：ValueError: No JSON object could be decoded](#)⁹

Python中和解析抓取的网站内容，即解析HTML，JSON等方面，相关的模块有，BeautifulSoup，json等

¹ http://www.crifan.com/files/doc/docbook/python_summary/release/html/python_summary.html#python_lib_beautifulsoup

² http://www.crifan.com/python_third_party_lib_html_parser_beautifulsoup

³ http://www.crifan.com/summary_usage_of_beautifulsoup_in_python

⁴ http://www.crifan.com/some_notation_about_python_beautifulsoup_parse_html

⁵ http://www.crifan.com/use_beautifulsoup_parse_the_backslash_style_html_json_string

⁶ http://www.crifan.com/notation_about_use_python_json_loads

⁷ http://www.crifan.com/use_python_json_loads_parse_string_contain_newline_will_fail_error

⁸ http://www.crifan.com/python_json_loads_valueerror_expecting_property_name_line_1_column_1_char_1

⁹ http://www.crifan.com/python_json_loads_valueerror_no_json_object_could_be_decoded

参考书目

[1] [【教程】抓取网并提取网页中所需要的信息 之 Python版](http://www.crifan.com/crawl_website_html_and_extract_info_using_python/)¹

¹ http://www.crifan.com/crawl_website_html_and_extract_info_using_python/