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）

版本：v1.1

Crifan Li

摘要

本文主要介绍了抓取网站，模拟登陆，抓取动态网页相关的逻辑，原理和如何实现。

主要包括：

- 抓取网页，模拟登陆等背后的通用的逻辑和原理
- 以提取songtaste网页中标题为例，详解如何抓取网站并提取网页内容
- 以模拟登陆百度为例，详解如何模拟登陆网站
- 以抓取网易博客帖子中的最近读者信息为例，详解如何抓取动态网页中的内容
- 详解了在模拟登陆和抓取动态网页过程中，如何用对应的网页分析工具，如IE9的F12，Chrome的Ctrl+Shift+J，Firefox的Firebug，去分析出对应的逻辑
- 针对抓取网站，模拟登陆，抓取动态网页，全部给出了完整的可用的，多种语言的示例代码：Python，C#，Java，Go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/web_scrape_emulate_login/release/html/web_scrape_emulate_login.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/web_scrape_emulate_login/

---

[1] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/html/web_scrape_emulate_login.html
[2] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/htmls/index.html
[3] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/pdf/web_scrape_emulate_login.pdf
[4] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/chm/web_scrape_emulate_login.chm
[5] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/txt/web_scrape_emulate_login.txt
[6] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/rtf/web_scrape_emulate_login.rtf
[7] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/webhelp/index.html
[8] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/html/web_scrape_emulate_login.html.7z
[9] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/htmls/index.html.7z
[10] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/pdf/web_scrape_emulate_login.pdf.7z
[11] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/chm/web_scrape_emulate_login.chm.7z
[12] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/txt/web_scrape_emulate_login.txt.7z
[13] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/rtf/web_scrape_emulate_login.rtf.7z
[14] http://www.crifan.com/files/doc/docbook/web_scrape_emulate_login/release/webhelp/web_scrape_emulate_login.webhelp.7z

修订历史		
修订 1.1	2013-09-22	crl
1. 把之前教程的地址整理过来 2. 添加新帖子的链接：模拟登陆百度的java版，go语言版		

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）：

Crifan Li

版本：v1.1

出版日期 2013-09-22

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

目录

前言

1. 本文目的

本文目的在于，如何从无到有的，了解抓取网站，模拟登陆，抓取动态网页方面的逻辑和具体实现。

第 1 章 网站抓取，模拟登陆，抓取动态网页的通用逻辑

相关旧帖

如何用Python，C#等语言去实现抓取静态网页+抓取动态网页+模拟登陆网站 [1]

【整理】各种浏览器中的开发人员工具Developer Tools：IE9的F12，Chrome的Ctrl+Shift+J，Firefox的Firebug [2]

【总结】浏览器中的开发人员工具（IE9的F12和Chrome的Ctrl+Shift+I）-网页分析的利器 [3]

【整理】关于抓取网页，分析网页内容，模拟登陆网站的逻辑/流程和注意事项 [4]

【教程】如何利用IE9的F12去分析网站登陆过程中的复杂的（参数，cookie等）值（的来源） [5]

【整理】关于http(GET或POST)请求中的url地址的编码(encode)和解码(decode) [6]

【整理】关于HTML网页源码的字符编码（charset）格式（GB2312，GBK，UTF-8，ISO8859-1等）的解释 [7]

【整理】网页抓取，模拟登陆，抓取动态网页内容等过程中，所涉及的Headers信息，Cookie信息，POST数据的处理逻辑 [8]

【整理】关于用正则表达式处理html代码方面的建议 [9]

[1] http://www.crifan.com/how_to_use_some_language_python_csharp_to_implement_crawl_website_extract_dynamic_webpage_content_emulate_login_website
[2] http://www.crifan.com/summary_webbrowser_developer_tool_ie9_f12_chrome_ctrl_shift_j_firefox_firebug
[3] http://www.crifan.com/browser_developer_tool_chrome_vs_ie9
[4] http://www.crifan.com/summary_about_flow_process_of_fetch_webpage_simulate_login_website_and_some_notice
[5] http://www.crifan.com/use_ie9_f12_to_analysis_the_root_source_of_values_of_parameter_cookie
[6] http://www.crifan.com/summary_url_encode_and_decode_during_http_get_post_request
[7] http://www.crifan.com/summary_explain_what_is_html_charset_and_common_value_of_gb2312_gbk_utf_8_iso8859_1
[8] http://www.crifan.com/website_crawl_process_related_headers_cookies_post_data_handle_logic
[9] http://www.crifan.com/for_process_html_with_many_tag_recommend_use_third_lib_while_simple_html_use_regular_expression

第 2 章 如何抓取静态网页并提取特定内容

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】抓取网并提取网页中所需要的信息 之 C#版 [2]

[1] http://www.crifan.com/crawl_website_html_and_extract_info_using_python/
[2] http://www.crifan.com/crawl_website_html_and_extract_info_using_csharp

第 3 章 如何模拟登陆网站

下面，给出足够多的例子：

3.1.（多种语言实现）模拟登陆百度

先去用工具分析逻辑：

【教程】手把手教你如何利用工具(IE9的F12)去分析模拟登陆网站(百度首页)的内部逻辑过程 [1]

再去用代码实现，此处，目前已经实现了：

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

- Python版
  【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）[3]

- Java版
  【教程】模拟登陆百度之Java代码版[4]

- Go语言版
  【记录】用go语言实现模拟登陆百度[5]

3.2.（多种语言实现）模拟登陆gogole

另外，也弄了个，模拟登陆google：

【记录】模拟登陆google [6]

[1] http://www.crifan.com/use_ie9_f12_to_analysis_the_internal_logical_process_of_login_baidu_main_page_website
[2] http://www.crifan.com/emulate_login_website_using_csharp/
[3] http://www.crifan.com/emulate_login_website_using_python/
[4] http://www.crifan.com/emulate_login_baidu_use_java_code/
[5] http://www.crifan.com/emulate_login_baidu_using_go_language/
[6] http://www.crifan.com/analysis_process_of_emulate_login_google/

第 4 章 如何抓取动态网页并提取特定内容

先去看看：

【教程】如何抓取动态网页内容 [1]

搞懂，抓取动态网页的逻辑。

再去看下面的例子：

4.1. 抓取动态网页示例：网易163博客的心情随笔FeelingCard

【记录】给BlogsToWordPress添加支持导出网易的心情随笔 [2]

【教程】以抓取网易博客帖子中的最近读者信息为例，手把手教你如何抓取动态网页中的内容 [3]

【记录】用Python解析网易163博客的心情随笔FeelingCard返回的DWR-REPLY数据 [4]

[1] http://www.crifan.com/how_to_crawl_dynamic_webpage_content
[2] http://www.crifan.com/blogstowordpress_add_feeling_card_for_163_netease_blog/
[3] http://www.crifan.com/example_to_crawl_dynamic_webpage_content_of_recent_reader_info_for_netease_blog_post
[4] http://www.crifan.com/parse_netease_163_post_emotion_feelingcard_dwr_reply_data/

第 5 章 抓取静态或动态网页和模拟登陆的注意事项和总结

相关旧帖

【总结】静态网页抓取，动态网页抓取，模拟登陆的注意事项和心得 [1]

[1] http://www.crifan.com/note_about_website_crawl_and_emulate_login/

参考书目

[1] 如何用Python，C#等语言去实现抓取静态网页+抓取动态网页+模拟登陆网站[1]

[1] http://www.crifan.com/how_to_use_some_language_python_csharp_to_implement_crawl_website_extract_dynamic_webpage_content_emulate_login_website