

# 半月谈 · 今日谈爬虫开发手册

Snail XHR

2021.7.14

# 目录

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>关于本爬虫</b>                                       | <b>3</b> |
| <b>2</b> | <b>实现原理</b>  | <b>3</b> |
| <b>3</b> | <b>实现方法</b>  | <b>3</b> |
| <b>4</b> | <b>函数说明</b>  | <b>3</b> |
| 4.1      | get_news_list(base_url) . . . . .                  | 3        |
| 4.2      | get_news(news_url) . . . . .                       | 4        |
| 4.3      | delete_html_tag(string) . . . . .                  | 4        |
| 4.4      | write_into_file(file_name,title,article) . . . . . | 4        |

## 1 关于本爬虫

本爬虫用于获取半月谈网上今日谈的内容 (<http://www.banyuetan.org/byt/jinritan/index.html>)。

爬虫获取的是今日谈中的 Top10 要闻。

## 2 实现原理

爬虫首先爬取今日谈页面中 Top10 要闻的地址。对于每篇文章，分别获取其标题及内文，并写入文件。

据观察，Top10 要闻一栏在一对 class 属性为 title2 的 div 标签中。进一步的，我们发现，所有在这一栏中的 a 标签（指向文章的链接）的 class 属性都为 special1。因此，在提取新闻地质这一阶段，我们只需要在整个网页中寻找 class 为 special1 的 a 标签。

接下来，对于每篇新闻，我们发现整个页面中有且只有一个 h1 标签，该标签中即是新闻标题。所有的正文则都被包含在了一个 class 属性为 detail\_content 的 div 标签中。我们只需提取这两个标签中的内容即可。

## 3 实现方法

我们选择了用 Python 实现爬虫，依赖的库包括 Beautiful Soup4 , requests , re(正则表达式) , lxml(XML 解析)。我们一般使用 requests 库下载页面，并使用 Beautiful Soup4 过滤结果。最后，用正则表达式剔除 HTML 标签。

## 4 函数说明

### 4.1 get\_news\_list(base\_url)

该函数用于获取“Top10 要闻”的 URL。

参数：你应当提供“Top10 要闻”所处页面的 URL，目前为 <http://www.banyuetan.org/byt/jinritan/index.html>。

该函数将会返回一个列表，包含“Top10 要闻”的 URL。

## 4.2 get\_news(news\_url)

该函数用于获取一篇新闻的标题和正文。

参数：你应当提供一个 URL，即是你需要获取的新闻的网址。

该函数将会返回一个字典，其中包含两个键：title 和 news。title 键对应的值是该新闻的标题的 HTML，news 键则对应正文的 HTML。值得注意的是，此时返回的数据是 HTML 代码，而非纯文本。

## 4.3 delete\_html\_tag(string)

该函数用于删除一个字符串中的 HTML 标签，注意，它只是寻找尖括号，并将其中的内容删除而以。因 HTML 标签形如：<tag> text </tag>，所以我们可以很轻易地删除 HTML 标签。该函数的原理是正则替换。使用的正则表达式：

```
<(\ S*?)*[>]*>.*?|<.*? />
```

## 4.4 write\_into\_file(file\_name,title,article)

该函数用于将一篇新闻写入文件。

参数：包含三个参数，file\_name 用于是目标文件名。title 则是新闻的标题，将在文件的最上方打印，并在其下留出空行。article 是正文，将被原样输出至文件中。

请注意，该函数纯粹只将参数写入文件，不会对其内容作改变。请保证你的参数与你的期望格式相同。