美国联邦储备银行爬虫开发手册

Snail XHR

2021.7.14

目录

1 关于本爬虫

本爬虫用于获取美国联邦储备银行网站上的内容。(`https://www.federalreserve.gov/newsevents.htm`)。

爬虫获取的是新闻以及演讲稿。

备注：该网站有三个栏目，分别是新闻、演讲、证词。该爬虫没有爬取证词。

2 实现原理

爬虫首先爬取新闻与活动页面中新闻稿、演讲稿的地址。对于每篇文章，分别获取其标题及内文，并写入文件。

据观察，新闻、演讲的链接都在一对 class 属性为 news news___title 的 p 标签中。我们初步的解决方案是，先以空格为分割符，将该 p 标签的源码进行切片，在提取其中的链接。此时，需要保证该标签中没有除所需内容以外的链接。

接下来，对于每篇新闻，其标题被包含在一对 class 属性为 title 的 h3 标签中。所有的正文则都被包含在了一个 class 属性为 col-xs-12 col-sm-8 col-md-8 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)

该函数用于获取所有新闻与演讲稿的 URL。

参数：你应当提供新闻与活动页面的 URL ，目前为 `https://www.federalreserve.gov/newsevents.htm`。

该函数将会返回一个列表，包含所有新闻、演讲稿的 URL。

4.2 get_news(news_url)

该函数用于获取一篇文章的标题和正文。

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：

$$<(\backslash\ S*?)*[\hat{}>]*>.*?|<.*?\ />$$

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。