看雪精华帖爬虫

转载

weixin_30533797 于 2016-12-02 21:16:00 发布 71 收藏
文章标签： 爬虫 移动开发 php
原文链接：http://www.cnblogs.com/bingghost/p/6127384.html

看雪自带的搜索感觉不是太好用, 然后弄了个爬虫

目前支持4种功能

1. 爬取某个版块所有的链接, 并保持到文件

2. 自动把精华帖分类出来, 并保存到文件

3. 把含有指定关键字的链接单独保存为文件(针对所有链接)

4. 把含有指定关键字的链接单独保存为文件(针对所有精华帖链接)

github下载地址:

https://github.com/bingghost/pediy_spider

需要下载下面的依赖库

bs4
requests
html5lib

代码如下

```
#!/usr/bin/env python
# encoding: utf-8
"""
@author:     bingghost
@copyright:  2016 bingghost. All rights reserved.
@contact:
@date:       2016-12-1
@description: 看雪爬虫
"""
import re
import time
import requests
import argparse
from bs4 import BeautifulSoup


import sys
reload(sys)
sys.setdefaultencoding('utf8')
```

```
class PediySpider:
    def __init__(self, spider_url, specified_title):
        self._url = spider_url

        self.file_dict = {"all_title":"all_title.txt",
                          "good_title":"good_title.txt",
                          "filter_title":"filter_title.txt",
                          "filter_good_title":"filter_good_title.txt"}

        # good title
        self.filter_list = ['jhinfo.gif', 'good_3.gif', 'good_2.gif']

        # title specified
        self.specified_title = specified_title

        self.page_count = self.get_page_count()
        pass

    def get_page_content(self, page_num):
        rep_data = requests.get(self._url + str(page_num))
        soup = BeautifulSoup(rep_data.content, 'html5lib')
        return soup

    def get_page_count(self):
        value = int(self.get_page_content(1).select('div.pagenav td.vbmenu_control')[0].get_text().split('
        return value

    def save_file(self, content, filename):
        print content
        with open(filename, 'a+') as f:
            f.write(content.encode('utf-8') + '\n')

    def is_good_title(self, item):
        # The item in threads_box is a <td> tag. when we need to find
        # the img tag,we must find in its parent tag namely the <tr> tag.
        img_list = item.parent.find_all('img')
        for img in img_list:
            if img.get('src').split('/')[-1].lstrip() in self.filter_list:
                return True

        return False
        pass

    def is_specified_title(self, title_content):
        if self.specified_title is None:
            return False

        specified_title_encode = self.specified_title.encode('utf8')
        title_content_encode = title_content.encode('utf8')
        if specified_title_encode in title_content_encode:
            return True
            pass

        return False
        pass

    def is_good_specified_title(self, title_content):
        if self.specified_title is None:
```

```
            return False

        specified_title_encode = self.specified_title.encode('utf8')
        title_content_encode = title_content.encode('utf8')
        if specified_title_encode in title_content_encode:
            return True
            pass

        return False
        pass

    def check_content(self, threads_box):
        url_head = 'http://bbs.pediy.com/showthread.php?' + 't='

        for item in threads_box:
            title_box = item.find(id=re.compile('thread_title'))
            title = title_box.get_text()
            title_url = url_head + title_box.get('href').split('=')[-1]

            # now get the title and url
            self.save_file(title + '   ' + title_url, self.file_dict['all_title'])

            # excellent good and attention title
            is_good_title = self.is_good_title(item)
            if is_good_title:
                # print single_thread_box
                self.save_file(title + '   ' + title_url, self.file_dict['good_title'])
                pass

            # specified title content
            if self.is_specified_title(title):
                self.save_file(title + '   ' + title_url, self.file_dict['filter_title'])
                pass

            # specified good title content
            is_good_title_filter = self.is_good_specified_title(title)
            if is_good_title and is_good_title_filter:
                self.save_file(title + '   ' + title_url, self.file_dict['filter_good_title'])
                pass
        pass

    def worker(self):
        for i in range(1, 100000):
            if i > self.page_count:
                break

            # get all title info in current page
            threads_box = self.get_page_content(i).find_all(id=re.compile('td_threadtitle'))
            self.check_content(threads_box)

            time.sleep(3)

    def start_work(self):
        print "[-] start spider"

        self.worker()

        print "[-] spider okay"
        pass
```

```
    pass


def set_argument():
    # add description
    parser = argparse.ArgumentParser(
        description="A spider for the bbs of pediy's Android security forum,"
               "also you can modify the url to spider other forum.")

    # add argument
    group = parser.add_mutually_exclusive_group(required=True)
    group.add_argument(
        '-a', '--all',
        action='store_true',
        help='Get all titles')

    group.add_argument(
        '-f', '--filter',
        type=str,
        default=None,
        help='filter title')

    group.add_argument(
        '-gf', '--gfilter',
        type=str,
        default=None,
        help='filter good title')

    args = parser.parse_args()
    return args
    pass


def main():
    args = set_argument()

    spider_dict = {"android":"http://bbs.pediy.com/forumdisplay.php?f=161&order=desc&page=",
                   "ios":"http://bbs.pediy.com/forumdisplay.php?f=166&order=desc&page="}

    pediy_spider = None

    if args.all:
        pediy_spider = PediySpider(spider_dict['android'], None)
        pass

    if args.filter:
        pediy_spider = PediySpider(spider_dict['android'], args.filter)
        pass

    if args.gfilter:
        pediy_spider = PediySpider(spider_dict['android'], args.gfilter)
        pass

    pediy_spider.start_work()
    pass


if __name__ == '__main__':
    main()
```

效果:

```
/Library/Frameworks/Python.framework/Versions/2.7/bin/python2.7 /Users/bingghost/code/python/test/pediy_spider.py -gf 安卓
[-] start spider
    【招聘】看雪科技 2017 招聘开启!    http://bbs.pediy.com/showthread.php?t=212508
    【第15题进行中】看雪.博文视点CrackMe攻防大赛(踩楼送奖!)    http://bbs.pediy.com/showthread.php?t=213207
    "麦洛克菲"内核、移动安全培训(第11期2017年3月4日开课，支持远程和全日制零基础，可试听)    http://bbs.pediy.com/showthread.php?t=142155
    【推荐】『Android安全』版优秀和精华帖分类索引    http://bbs.pediy.com/showthread.php?t=179524
    【公告】看雪Android安全小组招募成员【注：20人左右】    http://bbs.pediy.com/showthread.php?t=212703
    【注意】企业账号发帖注意事项    http://bbs.pediy.com/showthread.php?t=197893
    Android开源项目收集[2013/0701更新]    http://bbs.pediy.com/showthread.php?t=158513
    【注意】Android安全版的规定和发帖建议（2014.12更新）    http://bbs.pediy.com/showthread.php?t=179475
    阿里聚安全攻防挑战赛报名开启!    http://bbs.pediy.com/showthread.php?t=214355
    闲聊阿里加固（一）    http://bbs.pediy.com/showthread.php?t=214116
```

转载于:https://www.cnblogs.com/bingghost/p/6127384.html