实验报告

课程名称	内容安全实验			成　　绩		教师签名	
实验名称	Python 网络爬虫			实验序号		实验日期	2021.04.07
姓　　名		学　　号		专　　业		年级-班	

一、实验目的及实验内容
（本次实验所涉及并要求掌握的知识；实验内容；必要的原理分析）　　小题分

实验目的：

使用 requests-BeautifulSoup-re 技术路线，编写程序爬取网页。

实验内容：

1、参考实例 4，爬取百度搜索风云榜 http://top.baidu.com/ 任一榜单，搜索结果按顺序逐行输出（含编号），榜单自选。

本次实验选取的目标榜单为“百度搜索风云榜-娱乐-电影榜”（http://top.baidu.com/category?c=1），结果将输出并保存该页面的六个榜单：全部电影榜单、爱情榜单、喜剧榜单、惊悚榜单、科幻榜单、剧情榜单这六个板块的搜索指数排名前 50 的电影名称及其搜索指数。

结果将额外被保存在 data 目录下的 txt 文本文档中。

2、爬取当当图书排行榜（榜单自选），格式：爬取结果包含但不限于[排名 书名 作者]，注意输出格式对齐。

本次实验选取的目标榜单为“当当网-图书榜-好评榜（top 500）-哲学/宗教”（http://bang.dangdang.com/books/fivestars/01.28.00.00.00.00-all-0-0-2-1），结果将输出并保存宗教/哲学系列的累计好评榜排行前 500 本书的排名、书名、作者及出品方、出版社、出版年份、现价、原价、折扣信息。

结果将额外被保存在 data 目录下的 csv 文件中。

原理分析：

1、使用 python 的 request 库的 get 方法可以很方便地完成对网页的访问请求并获取网页的 html 源码；

2、使用 python 的 BeaustifulSoup 方法可以很方便、灵活地选择对 html 的解析方式（如 find 方法、select 方法等），进而获取每个节点的属性、内容，为爬虫爬取爬取者关注的、存储在网页上的数据创造条件；

3、使用 python 的 lxml 库的 etree 方法也可以对 html 源码进行解析，其原理与 BeaustifulSoup 方法的 find 方法、select 方法原理差不多，但更为灵活，我个人更喜欢用这种方法。鉴于实验要求使用 requests-BeautifulSoup-re 技术路线，因此 etree 方法在本实验中仅作为辅助方法被使用一次；

4、使用 python 的 re 库可以将正则表达式应用于对结果的过滤，从而从 html 节点中过滤并提取到自己想要的结构化数据，进而进行存储。

二、实验环境
（本次实验所使用的器件、仪器设备等的情况）　　小题分：

(1)处理器：Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz　2.40 GHz
(2)操作系统环境：Windows 10 家庭中文版 x64 19042.867
(3)编程语言：Python 3.8
(4)其他环境：16 GB 运行内存
(5)IDE 及包管理器：JetBrains PyCharm 2020.1 x64， anaconda 3 for Windows（conda 4.9.0）
(6)借助的第三方库及使用目的：

BeautifulSoup：解析 html 网页结构并从中提取指定数据；
csv：用于结构化保存结果；
lxml：解析 html 网页结构并从中提取指定数据；
os：用于判断文件是否存在、创建文件路径；
random：创建随机选择；
re：正则，用于网址过滤；
requests：模拟浏览器行为，发送 GET 请求以获取目标网站的数据；
time：用于停止等待，避免因为访问过于频繁而被目标网页所在服务器限制访问。

三、实验步骤及实验过程分析
（详细记录实验过程中发生的故障和问题，进行故障分析，说明故障排除的过程及方法。根据具体实验，记录、整理相应的数据表格、绘制曲线、波形等）

小题分：

说明：

由于网页时刻在更新，本篇实验报告所记录的内容仅为写报告时（2021/04/07）的情况，可能与实际实验时（2021/04/04）结果有出入。

一切以运行时所得到的结果为准。

百度搜索风云榜爬取过程：

“百度搜索风云榜-娱乐-电影榜”（http://top.baidu.com/category?c=1）的页面布局如下图所示：

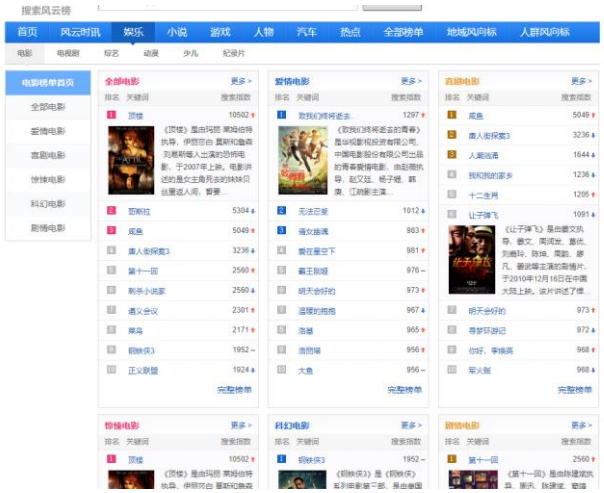

图 1 页面布局

可以看到，榜单分为六个部分：全部电影榜单、爱情榜单、喜剧榜单、惊悚榜单、科幻榜单、剧情榜单。由于每个榜单在这个页面所展现的仅仅是排名前十的结果，要想获取对应的榜

单下的排名前 50 的结果，需要在该页面中提取每个榜单的链接。

首先通过 request 库的 get 方法获取该页面的 html 源码，并用 BeaustifulSoup 创建一个 BeautifulSoup 对象。

在该页面对元素“全部电影”进行“右键-检查”的操作，可以看到该部分元素对应的 html 代码为 全部电影 。

该 html 代码包含了类别名字以及其链接，只要能获取这六个板块的链接，就能访问这六个板块所在的页面，进而获取排名前 50 的电影榜单。

因此接下来的工作就是获取这六个板块的链接。

右键该元素对应的 html 代码，在“复制”标签下可以看到多个选项，本次实验中会用到的有“复制 selector”选项和“复制 Xpath”选项。其中“复制 selector”选项将配合 BeaustifulSoup 的 select 方法使用，“复制 Xpath”选项将配合 lxml 的 etree 方法使用。这里选择复制 selector，会得到一个路径，该路径即为该节点在 html 的 DOM 树上的路径，通过该路径即可在 html 的 DOM 树上访问该节点的内容。

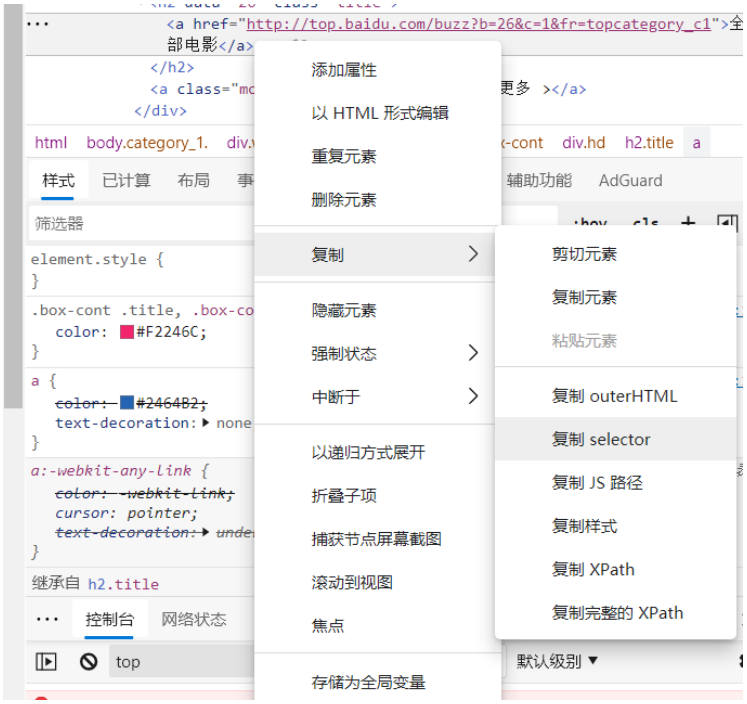

图 2 检查元素并获取其路径

下图 3 中的 path_1 即为上述元素的路径。其中“div: nth-child(1)”说明它在该页面的布局中占据了第一个板块的位置（事实上也是如此）。实际上，“nth-child(1)”是该元素所对应板块的属性，对于这六个板块而言，他们仅仅在这个属性上有区别，而在其他的路径、属性上毫无区别。事实上也是如此，如果检查元素“喜剧电影”，就会发现喜剧电影的路径如下图 3 的 path_2 所示，它的 div 属性为“nth-child(4)”。

因此，如果想获取全部的 6 个榜单及其对应的链接，只需要将这些能唯一标识其身份的“属性”去掉即可。因此，实际上要查找的路径为“#main > div > div.hd > h2 > a”，即下图 3 中 cla_path 所示。

将 cla_path 传入 BeaustifulSoup 的 select 方法，即可获取结果如下图 3 所示：

```
In [1]:
from crawl import *
```

```
In [2]:
url = "http://top.baidu.com/category?c=1"
html = get_html(url)
soup = BS(html, 'html.parser')
```

```
url: http://top.baidu.com/category?c=1
get response successfully!  2021-04-07 19:44:53
```

```
In [3]:
path_1 = '#main > div:nth-child(1) > div.hd > h2 > a'
path_2 = '#main > div:nth-child(4) > div.hd > h2 > a'
cla_path = "#main > div > div.hd > h2 > a"
classes = soup.select(cla_path)
print(classes)
```

```
[<a href="./buzz?b=26&amp;c=1">全部电影</a>, <a href="./buzz?b=338&amp;c=1">爱情电影</a>, <a href="./buzz?b=340&amp;c=1">喜剧电影</a>, <a href="./buzz?b=339&amp;c=1">惊悚电影</a>, <a href="./buzz?b=437&amp;c=1">科幻电影</a>, <a href="./buzz?b=337&amp;c=1">剧情电影</a>]
```

图 3 获取单个榜单页面链接

可以看到，结果列表中确实返回了这六个榜单的名字及其链接，然而其链接并不是直接指向其所在页面的链接，而是相对于当前页面的相对路径。当前页面的 url 为 http://top.baidu.com/category?c=1，其中“?c=1”为尾缀参数，真正的 url 应为 http://top.baidu.com/category，而当前页面所在的路径应为 http://top.baidu.com；再加上图 3 中获取到的结果显示“./”为当前路径，因此单个榜单对应的真正的 url 应为 http://top.baidu.com/buzz?b=*&id*，其中*&id* 为每个页面对应的 id，如“全部电影”对应的 id 即为 26、爱情电影对应的 id 为 338，以此类推。

通过构造合适的正则表达式可获取单个榜单的相对路径，然后再使用字符串拼接即可获取 6 个榜单对应的真正的 url。实际实验时发现，参数“c=1”对结果不产生实质性影响，因此不将其纳入拼接范围。构造正则表达式如图 4 中 cla_rule 所示，该表达式会返回字串“buzz?b=\d*”，其中“\d*”表示任意长度的数字。

```
In [4]:
cla_rule = r"buzz\?b=\d*"
head = "http://top.baidu.com/"
classes_ = {}
# tail = "&c=1"  # useless
rule = r"buzz\?b=\d*"
for i in classes:
    link = re.compile(rule).findall(str(i))
    classes_[i.text] = head + link[0]
for key in list(classes_.keys()):
    print(key, ':', classes_[key])
```

```
全部电影 : http://top.baidu.com/buzz?b=26
爱情电影 : http://top.baidu.com/buzz?b=338
喜剧电影 : http://top.baidu.com/buzz?b=340
惊悚电影 : http://top.baidu.com/buzz?b=339
科幻电影 : http://top.baidu.com/buzz?b=437
剧情电影 : http://top.baidu.com/buzz?b=337
```

图 4　获取单个榜单的真正 url

随后分别对每个 url 进行访问、数据获取。考虑到这六个网页的布局是完全一样的，因此仅以今日喜剧电影排行榜（http://top.baidu.com/buzz?b=340）为例进行说明。

先看页面布局如图 4 所示：

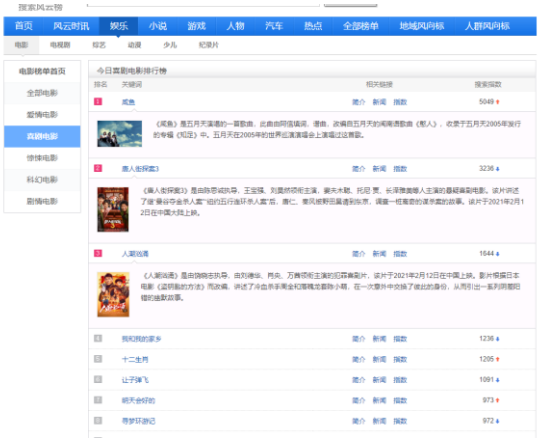

图 5　今日喜剧电影排行榜页面布局

首先检查“今日喜剧电影排行榜”获取标题的元素路径并使用 select 方法获取标题内容如图 6 所示；

```
In [5]:
url = "http://top.baidu.com/buzz?b=340"
html = get_html(url)
soup = BS(html, 'html.parser')
title_path = "#main > div.mainBody > div > div > h2"
title = soup.select(title_path)[0].text
print(title)

url: http://top.baidu.com/buzz?b=340
get response successfully!  2021-04-07 20:40:29
今日喜剧电影排行榜
```

图 6　获取榜单标题

由于在榜单页面上所有内容一字排开、平铺直叙，因此 find_all 方法能很方便地获取该页面上所有的电影的排名、名称、搜索指数信息。

获取电影名称的信息过程如图 7 所示，任选一部电影的名称进行检查，可以发现电影名称被保存在标签为 a、属性为“list-title”的节点。

```
▼<td class="keyword">
    <a class="list-title" target="_blank" href="http://www.baidu.com/baidu?cl=3&tn=SE_baiduhomet8_jmjb7mjw&rsv_dl=fyb_top&fr=top1000&wd=%C8%C3%D7%D3%B5%AF%B7%C9" href_top="./detail?b=340&c=1&w=%C8%C3%D7%D3%B5%AF%B7%C9">让子弹飞</a> == $0
```

图 7　电影标题信息

使用 find_all 对其进行查找，可获取全部的前 50 的信息。其中前三分别为《咸鱼》、《唐人街探案 3》、《人潮汹涌》。如图 8 所示。

```
In [6]:
movie = []
m_list = soup.find_all('a', 'list-title')
for i in m_list:
    movie.append(i.text.strip())
print(len(movie))
print(movie[:3])

50
['咸鱼', '唐人街探案3', '人潮汹涌']
```

图 8　获取电影名称

获取电影搜索指数的过程类似，搜索指数的标签为 td、属性为“last”，其结果如图 9 所示。

```
In [7]:
hot = []
hot_list = soup.find_all('td', 'last')
for i in hot_list:
    hot.append(i.text.strip())
print(len(hot))
for i in hot:
    print(type(i), i)

50
<class 'str'> 5049
<class 'str'> 3236
<class 'str'> 1644
<class 'str'> 1236
<class 'str'> 1205
<class 'str'> 1091
<class 'str'> 973
<class 'str'> 972
```

图 9　获取搜索指数

最终的百度搜索风云榜-娱乐-电影榜爬取结果（节选）（2021年4月4日，00:30:14）如图 10 所示：

今日电影排行榜

排名	电影名称	搜索指数
1	顶楼	8317
2	哥斯拉	8118
3	咸鱼	4153
4	唐人街探案3	4072
5	第十一回	3901
6	送你一朵小红花	3822
7	正义联盟	3033
8	刺杀小说家	2954
9	金刚	2641
10	阿凡达	2161
11	人潮汹涌	1939
12	钢铁侠3	1910
13	菜鸟	1881
14	谁都有秘密	1860
15	银魂	1721
16	环太平洋：雷霆再起	1655

图 10　格式化输出爬取结果

当当网图书榜爬取过程：

“当当网-图书榜-好评榜（top 500）-哲学/宗教”（http://bang.dangdang.com/books/fivestars/01.28.00.00.00.00-all-0-0-2-1）的页面布局如下图 11 所示：

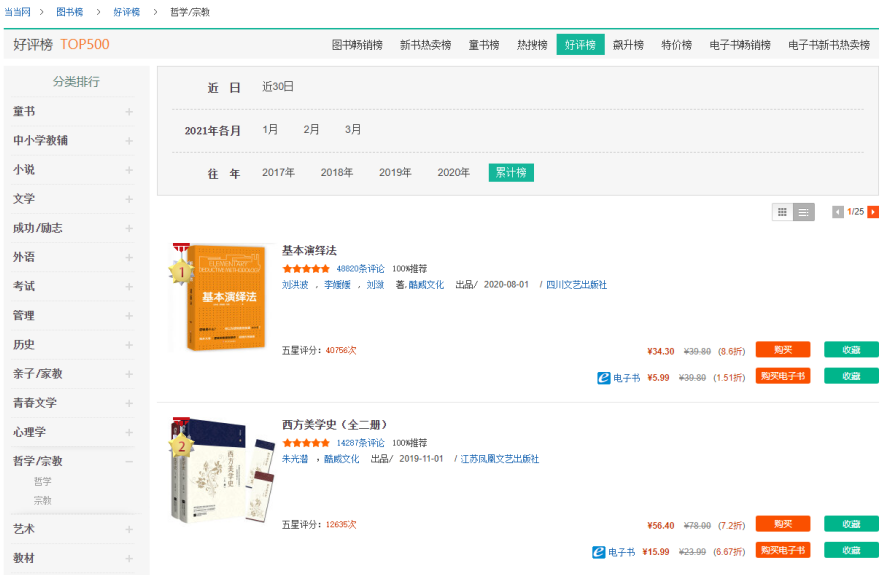

图 11　页面布局（列表视图）

通过多翻几页可以看到，该网页网址的最后一个“-”后的数字代表了页数，爬虫工作时只需更换这个数字就可以遍历 25 页*20 条共计 500 条内容。

事实上，当当网的图书榜单有两种视图，一种是大图平铺视图，另一种是列表视图。选择后者而非前者的原因是，列表视图的信息及页面结构更为清晰，方便爬取。（实际上，大图平铺视图和列表视图的参数由网址倒数第二个“-”及其后的数字控制，若倒数第二个数字为 1 则为大图平铺视图；若为 2 则为列表视图）

检查任意一本书的名字，得书名的通用路径为“body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.name > a”，由于当当网在某些图书的书名后会用括弧的形式做简介，因此选择用正则去掉简介。构造正则式如图 12 中 split_所示。split_是一个正先行断言，将保留“（”前的内容，即对于有简介的书而言将保留其简介前的部分；对于无简介的书而言将什么都匹配不到。因此还需加入一次判断，即若正则结果为空则用原始字串作为结果，若正则结果不为空则以正则结果为准。获取图书标题的结果如图 12 所示。

```
In [2]: url = 'http://bang.dangdang.com/books/fivestars/01.28.00.00.00.00-all-0-0-2-1'
        # 当当网> 图书榜 > 好评榜 > 哲学/宗教 > 累计榜 （列表视图）
        html = get_html(url)
        bs = BS(html, 'html.parser')

        url: http://bang.dangdang.com/books/fivestars/01.28.00.00.00.00-all-0-0-2-1
        get response successfully!  2021-04-08 09:11:55

In [3]: book = []
        book_path = 'body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.name > a'
        book_pre = bs.select(book_path)
        split_ = r'.*(?= (）'
        for item in book_pre:
            pre = item.text
            done = re.compile(split_).findall(pre)
            if len(done) == 0:
                done = pre
            else:
                done = done[0]
            book.append(done)
        print(book)
        print(len(book))

        ['基本演绎法', '西方美学史', '论语译注', '简单逻辑学 改变思维方式第一书 ', '李泽林谈人生', '资本论', '单独中的洞见', '易经杂说', '陈果：好的孤独', '老子他说', '美的历程（精装）', '每时每刻皆为逍遥时光', '懂你', '南怀瑾选集（套装全十册）', '尼采的心灵咒语2', '菜根谭', '给快节奏时代的简单哲学', '做才是得到——拿达吉堪布给你点滴加持', '天堂的证据', '次第花开 修订版']
        20
```

图 12　获取图书标题

检查任意一本书的作者名字可以看到出版社信息的通用路径为“body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.publisher_info”，通过简单分割可以发现该标签下包含了一本书的作者、出版时间、出版社信息。

```
In [4]: author_org_path = 'body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.publisher_info'
        unit_ = bs.select(author_org_path)
        ls = 0
        for i in unit_:
            undone = i.text.strip().split('/')
            ls += 1
            print(ls, undone)

        1 ['刘洪波, 李媛媛, 刘渤  著,酷威文化 出品', '\r\n                2020-08-01\n', '四川文艺出版社']
        2 ['朱光潜', '酷威文化 出品', '\r\n                2019-11-01\n', '江苏凤凰文艺出版社']
        3 ['杨伯峻 译注', '\r\n                2017-08-01\n', '中华书局']
        4 ['吴昱荣', '\r\n                2013-07-01\n', '中国华侨出版社']
        5 ['李泽林 著，李泽林研究所 编', '\r\n                2009-03-01\n', '当代中国出版社']
```

图 13　获取出版社信息

由于字串中多了一些转义字符及空格，因此需要通过清洗字串将其去掉；考虑到有些书的作者、出版时间、出版社信息可能会有空缺，因此需要对清洗结果中相关缺失项用字串“None”补位缺省值。清洗及补充结果如图 14 所示（’\u300’为）：

```
In [5]: author = []
        year = []
        org = []
        for item in unit_:
            undone = item.text.strip().split('/')
            if len(undone[0]) == 0:
                author_ = 'None'
            else:
                author_ = undone[0].strip('\u3000')
            author.append(author_)
            if len(undone[1]) == 0:
                year_ = 'None'
            else:
                year_ = undone[1].strip().strip('\r').strip('\n')
            year.append(year_)
            if len(undone[2]) == 0:
                org_ = 'None'
            else:
                org_ = undone[2].strip('\u3000')
            org.append(org_)
        print(len(author), len(year), len(org))
        print(author[0])
        print(year[0])
        print(org[0])

        20 20 20
        刘洪波, 李媛媛, 刘渤  著,酷威文化 出品
        2020-08-01
        四川文艺出版社
```

图 14　对出版信息的清洗

在获取书本的价格时，通过观察发现，所有书都有纸质版的价格，但仅有部分书有电子版的价格。为方便起见，实际爬取时将仅爬取纸质版价格的结果。通过观察网页的 html 源码发

现，价格所在的路径为“body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.price > p > span.price_n”，而电子书价格的路径为“body > div.bang_wrapper > div.bang_content > div.bang_list_box > ul > li > div.price > p.ebook_line > span.price_n”，不论是 BeautifulSoup 的 find_all 方法还是 select 方法都没办法简单地区分一个价格是纸质书的价格还是电子书的价格，因此考虑用 lxml 的 etree，通过 xpath 方法的 not 命令就可以获取属性值不是“ebook_line”的 p 节点的子节点数据，这样就可以方便地获取一本书的纸质版价格。相关结果如图 15 所示：

```
In [6]: from lxml import etree
        ht = etree.HTML(html)

        html_data = ht.xpath("/html/body/div/div/div/ul/li/div/p[not(@class='ebook_line')]/span")
        no = 0
        ls = 0

        while no < len(html_data):

            tmp = []
            bias = 0
            while bias < 3:
                tmp.append(html_data[no].text)
                bias += 1
                no += 1
            ls += 1
            print(ls, tmp)
```

```
1 ['¥34.30', '¥39.80', '8.6折']
2 ['¥56.40', '¥78.00', '7.2折']
3 ['¥16.90', '¥26.00', '6.5折']
4 ['¥26.70', '¥29.80', '9.0折']
5 ['¥13.60', '¥19.00', '7.2折']
6 ['¥130.60', '¥158.00', '8.3折']
7 ['¥33.10', '¥38.00', '8.7折']
8 ['¥16.50', '¥21.00', '7.9折']
9 ['¥36.00', '¥36.00', '10.0折']
```

图 15 获取价格信息

最终的当当网-图书榜-好评榜（top 500）-哲学/宗教爬取结果（节选）（2021年4月4日，11:38:24）如图 16 所示：

```
哲学/宗教好评榜TOP500



--------------------------------------------------
1：《基本演绎法》
作者及出品方：  刘洪波，李媛媛，刘潋  著,酷威文化 出品
出版社：    四川文艺出版社
现价：¥17.10，  原价：¥39.80，  折扣：4.3折
--------------------------------------------------
--------------------------------------------------
2：《西方美学史》
作者及出品方：  朱光潜，酷威文化 出品
出版社：    江苏凤凰文艺出版社
现价：¥28.40，  原价：¥78.00，  折扣：3.6折
--------------------------------------------------
--------------------------------------------------
3：《论语译注》
作者及出品方：  杨伯峻 译注
出版社：    中华书局
现价：¥16.90，  原价：¥26.00，  折扣：6.5折
--------------------------------------------------
--------------------------------------------------
4：《简单逻辑学 改变思维方式第一书 》
作者及出品方：  吴昱荣
出版社：    中国华侨出版社
现价：¥14.90，  原价：¥29.80，  折扣：5.0折
--------------------------------------------------
--------------------------------------------------
```

图 16 当当网爬取结果

保存的 csv 内容（节选）如图 17 所示：

	A	B	C	D	E	F	G	H
1	排名	书名	作者及出	出版社	出版年份	现价	原价	折扣
2	1	基本演绎	刘洪波, 李	四川文艺	2020/8/1	¥17.10	¥39.80	4.3折
3	2	西方美学	朱光潜,	江苏凤凰	2019/11/1	¥28.40	¥78.00	3.6折
4	3	论语译注	杨伯峻 译	中华书局	2017/8/1	¥16.90	¥26.00	6.5折
5	4	简单逻辑	吴昱荣	中国华侨	2013/7/1	¥14.90	¥29.80	5.0折
6	5	季羡林谈	季羡林 著	当代中国	2009/3/1	¥13.60	¥19.00	7.2折
7	6	资本论	马克思	上海三联	2009/4/1	¥74.40	¥158.00	4.7折
8	7	单独中的	张方宇,	四川文艺	2018/7/1	¥19.00	¥38.00	5.0折
9	8	易经杂说	南怀瑾	复旦大学	2013/9/1	¥16.50	¥21.00	7.9折
10	9	陈果：好	陈果 著;	江苏凤凰	2017/4/19	¥18.00	¥36.00	5.0折
11	10	老子他说	南怀瑾 著	复旦大学	2005/12/1	¥18.90	¥24.00	7.9折
12	11	美的历程	李泽厚 著	生活.读书	2009/7/1	¥33.90	¥43.00	7.9折
13	12	每时每刻	费勇著	江苏文艺	2014/7/1	¥35.00	¥35.00	10.0折
14	13	懂你	陈果	山东画报	2016/6/1	¥15.60	¥24.00	6.5折
15	14	南怀瑾选	南怀瑾	复旦大学	2006/6/1	¥296.60	¥412.00	7.2折
16	15	尼采的心	(德) 尼采	江苏文艺	2013/10/1	¥22.10	¥28.00	7.9折
17	16	菜根谭	（明）洪	当代世界	2008/1/1	¥9.30	¥20.00	4.7折
18	17	给快节奏	[英]阿兰·	四川文艺	2020/2/1	¥19.90	¥39.80	5.0折
19	18	做才是得	索达吉堪	读者出版	2012/11/1	¥18.80	¥38.00	4.9折
20	19	天堂的证	(美)亚历山	百花洲文	2013/7/1	¥25.20	¥35.00	7.2折
21	20	次第花开	希阿荣博	海南出版	2017/2/1	¥19.90	¥39.80	5.0折
22	21	沉思录 买	〔古罗马〕	上海三联	2008/5/1	¥13.20	¥22.00	6.0折
23	22	宽心-心灵	星云大师	江苏文艺	2009/6/1	¥13.90	¥28.00	5.0折
24	23	金刚经说	南怀瑾	复旦大学	2012/10/1	¥23.70	¥30.00	7.9折
25	24	于丹《庄	于丹 著	中国民主	2007/2/1	¥10.00	¥20.00	5.0折
26	25	故道白云	一行禅师,	线装书局	2007/6/1	¥12.40	¥38.00	3.3折
27	26	不抑郁的	None	华东师范	2013/11/1	¥24.00	¥32.00	7.5折

图 17　保存的 CSV 内容

四、实验结果总结

（对实验结果进行分析，完成思考题目，总结实验的新的体会，并提出实验的改进意见）

小题分：

本次爬虫实验相对来讲比较简单，信息几乎都被存储在静态页面上，爬虫实际工作的时候也没有遇到反爬措施的限制。事实上，编写爬虫时其实应当考虑 robot 协议；有些网页会采用异步加载的形式，并不会直接将信息写在网页源码中；有的网页会检查一个访问请求是否来自爬虫，并对频繁访问的请求进行一定的限制和验证措施，因此很多时候要考虑采用反反爬措施，包括但不限于随机 UA、停等时间、设置 cookie 等。

页面的选取对爬虫的编写成本及工作效率影响较大。很多购物网页往往会采用大图模式为消费者提供更多的信息，然而这对于爬虫来讲有时会造成一定的不便。页面的 URL 中往往会带有很多参数，有些参数是必要的而有些是不必要的，因此，选择合适的页面、合适的参数将大大提高爬虫的工作效率、简化工作流程。

在从 html 源码中提取目标元素时有多种方法可以使用，虽然原理大差不差，但合适的方法可以为工作提供更多的便利，实际工作中可以考虑多种方法联合使用。

正则表达式在对提取到的数据进行过滤清洗的时候会起到至关重要的作用，编写合适的正则表达式可以更有效地得到结构化数据。