

练习1 - 电影天堂二级页面抓取

■ 领取任务

```
1  ## html = requests.get(url=url,headers=headres).content.decode('gb2312','ignore')
2  【1】地址
3      电影天堂 - 2019年新片精品 - 更多
4  【2】目标
5      电影名称、下载链接
6
7  【3】分析
8      *****一级页面需抓取*****
9          1、电影详情页链接
10
11      *****二级页面需抓取*****
12          1、电影名称
13          2、电影下载链接
14
15  【4】要求
16      4.1) 所抓数据存入MySQL数据库
17      4.2) 所抓数据存入mongodb数据库
18      4.3) 所抓数据存入csv文件
19      4.4) redis实现增量爬虫
20      4.5) MySQL实现增量爬虫
```

实现步骤

- 1、确定响应内容中是否存在所需抓取数据
- 2、找URL规律

```
1  第1页 : https://www.dytt8.net/html/gndy/dyzz/list_23_1.html
2  第2页 : https://www.dytt8.net/html/gndy/dyzz/list_23_2.html
3  第n页 : https://www.dytt8.net/html/gndy/dyzz/list_23_n.html
```

■ 3、写正则表达式

```
1  1、一级页面正则表达式
2      <table width="100%" .*?<td width="5%" .*?<a href="(.*?)".*?</a>.*?</td>.*?</tr>.*?</table>
3  2、二级页面正则表达式
4      <div class="title_all"><h1><font color=#07519a>(.*?)</font></h1></div>.*?<td style="WORD-
    WRAP.*?>.*?>(.*?)</a>
```

■ 4、代码实现

```

1 import requests
2 import re
3 import time
4 import random
5 from fake_useragent import UserAgent
6
7 class FilmSkySpider(object):
8     def __init__(self):
9         # 一级页面url地址
10        self.url = 'https://www.dytt8.net/html/gndy/dyzz/list_23_{}.html'
11
12    # 获取html功能函数
13    def get_html(self,url):
14        headers = {'User-Agent':UserAgent().random}
15        # 通过网站查看网页源码,查看网站charset='gb2312'
16        # 如果遇到解码错误,识别不了某些字符,则 ignore 忽略掉
17        html = requests.get(url=url, headers=headers).content.decode('gb2312', 'ignore')
18
19        return html
20
21    # 正则解析功能函数
22    def re_func(self,re_bds,html):
23        pattern = re.compile(re_bds,re.S)
24        r_list = pattern.findall(html)
25
26        return r_list
27
28    # 获取数据函数 - html是一级页面响应内容
29    def parse_page(self,one_url):
30        html = self.get_html(one_url)
31        re_bds = r'<table width="100%" .*?<td width="5%" .*?<a href="(.*?)".*?</td>.*?</table>'
32        # one_page_list: ['/html/xxx', '/html/xxx', '/html/xxx']
33        one_page_list = self.re_func(re_bds,html)
34
35        for href in one_page_list:
36            two_url = 'https://www.dytt8.net' + href
37            self.parse_two_page(two_url)
38            # uniform: 浮点数,爬取1个电影信息后sleep
39            time.sleep(random.uniform(1, 3))
40
41
42    # 解析二级页面数据
43    def parse_two_page(self,two_url):
44        item = {}
45        html = self.get_html(two_url)
46        re_bds = r'<div class="title_all"><h1><font color=#07519a>(.*?)</font></h1></div>.*?<td style="WORD-WRAP.*?>.*?>(.*?)</a>'
47        # two_page_list: [('名称1', 'ftp://xxxx.mkv')]
48        two_page_list = self.re_func(re_bds,html)
49
50        item['name'] = two_page_list[0][0].strip()
51        item['download'] = two_page_list[0][1].strip()
52
53        print(item)
54
55
56    def main(self):

```

```

57     for page in range(1,201):
58         one_url = self.url.format(page)
59         self.parse_page(one_url)
60         # uniform: 浮点数
61         time.sleep(random.uniform(1,3))
62
63 if __name__ == '__main__':
64     spider = FilmSkySpider()
65     spider.main()

```

■ 5、练习

```

1  # 请使用两种方式实现
2  【1】使用redis实现增量爬虫
3  【2】使用MySQL实现增量爬虫
4      2.1) MySQL中新建表 urltab,存储所有爬取过的链接的指纹
5      2.2) 在爬取之前,先判断该指纹是否爬取过,如果爬取过,则不再继续爬取

```

练习代码实现 - MySQL

```

1  # 建库建表
2  create database filmskydb charset utf8;
3  use filmskydb;
4  create table request_finger(
5  finger char(32)
6  )charset=utf8;
7  create table filmtab(
8  name varchar(200),
9  download varchar(500)
10 )charset=utf8;

```

```

1  import requests
2  import re
3  from fake_useragent import UserAgent
4  import time
5  import random
6  import pymysql
7  from hashlib import md5
8  import sys
9
10 class FilmSkySpider(object):
11     def __init__(self):
12         # 一级页面url地址
13         self.url = 'https://www.dytt8.net/html/gndy/dyzz/list_23_{}.html'
14         self.db = pymysql.connect('localhost','root','123456','filmskydb',charset='utf8')
15         self.cursor = self.db.cursor()
16
17     # 获取html功能函数
18     def get_html(self,url):
19         headers = {'User-Agent':UserAgent().random}
20         # 通过网站查看网页源码,查看网站charset='gb2312'
21         # 如果遇到解码错误,识别不了一些字符,则 ignore 忽略掉
22         html = requests.get(url=url, headers=headers).content.decode('gb2312', 'ignore')
23

```

```

24     return html
25
26 # 正则解析功能函数
27 def re_func(self, re_bds, html):
28     pattern = re.compile(re_bds, re.S)
29     r_list = pattern.findall(html)
30
31     return r_list
32
33 # 获取数据函数 - html是一级页面响应内容
34 def parse_page(self, one_url):
35     html = self.get_html(one_url)
36     re_bds = r'<table width="100%" .*?<td width="5%" .*?<a href="(.*?)".*?</td>.*?</table>'
37     # one_page_list: ['/html/xxx', '/html/xxx', '/html/xxx']
38     one_page_list = self.re_func(re_bds, html)
39
40     for href in one_page_list:
41         two_url = 'https://www.dytt8.net' + href
42         # 判断在数据库中是否存在此链接, 一旦存在, 直接break, 新更新的链接都在上面
43         sel = 'select finger from request_finger where finger=%s'
44         s = md5()
45         s.update(two_url.encode())
46         finger = s.hexdigest()
47         result = self.cursor.execute(sel, [finger])
48         if not result:
49             self.parse_two_page(two_url)
50             # uniform: 浮点数, 爬取1个电影信息后sleep
51             time.sleep(random.uniform(1, 3))
52             ins = 'insert into request_finger values(%s)'
53             self.cursor.execute(ins, [finger])
54             self.db.commit()
55         else:
56             sys.exit('更新完成')
57
58 # 解析二级页面数据
59 def parse_two_page(self, two_url):
60     item = {}
61     html = self.get_html(two_url)
62     re_bds = r'<div class="title_all"><h1><font color=#07519a>(.*?)</font></h1></div>.*?<td style="WORD-WRAP.*?>.*?>(.*?)</a>'
63     # two_page_list: [('名称1', 'ftp://xxxx.mkv')]
64     two_page_list = self.re_func(re_bds, html)
65
66     item['name'] = two_page_list[0][0].strip()
67     item['download'] = two_page_list[0][1].strip()
68     ins = 'insert into filmtab values(%s,%s)'
69     film_list = [
70         item['name'], item['download']
71     ]
72     self.cursor.execute(ins, film_list)
73     self.db.commit()
74     print(film_list)
75
76
77
78 def run(self):
79     for page in range(1, 201):

```

```
80         one_url = self.url.format(page)
81         self.parse_page(one_url)
82         # uniform: 浮点数
83         time.sleep(random.uniform(1,3))
84
85 if __name__ == '__main__':
86     spider = FilmSkySpider()
87     spider.run()
```