练习1 - 电影天堂二级页面抓取

- **领取任务**

```
## html = requests.get(url=url,headers=headres).content.decode('gb2312','ignore')
【1】地址
    电影天堂 - 2019年新片精品 - 更多
【2】目标
    电影名称、下载链接

【3】分析
    *********一级页面需抓取***********
            1、电影详情页链接

    *********二级页面需抓取***********
            1、电影名称
            2、电影下载链接

【4】要求
    4.1) 所抓数据存入MySQL数据库
    4.2) 所抓数据存入mongodb数据库
    4.3) 所抓数据存入csv文件
    4.4) redis实现增量爬虫
    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**

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

- **3、写正则表达式**

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

- **4、代码实现**

```python
import requests
import re
import time
import random
from fake_useragent import UserAgent

class FilmSkySpider(object):
  def __init__(self):
    # 一级页面url地址
    self.url = 'https://www.dytt8.net/html/gndy/dyzz/list_23_{}.html'

  # 获取html功能函数
  def get_html(self,url):
    headers = {'User-Agent':UserAgent().random}
    # 通过网站查看网页源码,查看网站charset='gb2312'
    # 如果遇到解码错误,识别不了一些字符,则 ignore 忽略掉
    html = requests.get(url=url, headers=headers).content.decode('gb2312', 'ignore')

    return html

  # 正则解析功能函数
  def re_func(self,re_bds,html):
    pattern = re.compile(re_bds,re.S)
    r_list = pattern.findall(html)

    return r_list

  # 获取数据函数 - html是一级页面响应内容
  def parse_page(self,one_url):
    html = self.get_html(one_url)
    re_bds = r'<table width="100%".*?<td width="5%".*?<a href="(.*?)".*?ulink">.*?</table>'
    # one_page_list: ['/html/xxx','/html/xxx','/html/xxx']
    one_page_list = self.re_func(re_bds,html)

    for href in one_page_list:
      two_url = 'https://www.dytt8.net' + href
      self.parse_two_page(two_url)
      # uniform: 浮点数,爬取1个电影信息后sleep
      time.sleep(random.uniform(1, 3))


  # 解析二级页面数据
  def parse_two_page(self,two_url):
    item = {}
    html = self.get_html(two_url)
    re_bds = r'<div class="title_all"><h1><font color=#07519a>(.*?)</font></h1></div>.*?<td style="WORD-WRAP.*?>.*?>(.*?)</a>'
    # two_page_list: [('名称1','ftp://xxxx.mkv')]
    two_page_list = self.re_func(re_bds,html)

    item['name'] = two_page_list[0][0].strip()
    item['download'] = two_page_list[0][1].strip()

    print(item)


  def main(self):
```

```
    for page in range(1,201):
      one_url = self.url.format(page)
      self.parse_page(one_url)
      # uniform: 浮点数
      time.sleep(random.uniform(1,3))

if __name__ == '__main__':
  spider = FilmSkySpider()
  spider.main()
```

- **5、练习**

```
# 请使用两种方式实现
【1】使用redis实现增量爬虫
【2】使用MySQL实现增量爬虫
    2.1) MySQL中新建表 urltab,存储所有爬取过的链接的指纹
    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

```
# 建库建表
create database filmskydb charset utf8;
use filmskydb;
create table request_finger(
finger char(32)
)charset=utf8;
create table filmtab(
name varchar(200),
download varchar(500)
)charset=utf8;
```

```
import requests
import re
from fake_useragent import UserAgent
import time
import random
import pymysql
from hashlib import md5
import sys

class FilmSkySpider(object):
  def __init__(self):
    # 一级页面url地址
    self.url = 'https://www.dytt8.net/html/gndy/dyzz/list_23_{}.html'
    self.db = pymysql.connect('localhost','root','123456','filmskydb',charset='utf8')
    self.cursor = self.db.cursor()

  # 获取html功能函数
  def get_html(self,url):
    headers = {'User-Agent':UserAgent().random}
    # 通过网站查看网页源码,查看网站charset='gb2312'
    # 如果遇到解码错误,识别不了一些字符,则 ignore 忽略掉
    html = requests.get(url=url, headers=headers).content.decode('gb2312', 'ignore')

```

```
        return html

    # 正则解析功能函数
    def re_func(self,re_bds,html):
        pattern = re.compile(re_bds,re.S)
        r_list = pattern.findall(html)

        return r_list

    # 获取数据函数 - html是一级页面响应内容
    def parse_page(self,one_url):
        html = self.get_html(one_url)
        re_bds = r'<table width="100%".*?<td width="5%".*?<a href="(.*?)".*?ulink">.*?</table>'
        # one_page_list: ['/html/xxx','/html/xxx','/html/xxx']
        one_page_list = self.re_func(re_bds,html)

        for href in one_page_list:
            two_url = 'https://www.dytt8.net' + href
            # 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
            sel = 'select finger from request_finger where finger=%s'
            s = md5()
            s.update(two_url.encode())
            finger = s.hexdigest()
            result = self.cursor.execute(sel,[finger])
            if not result:
                self.parse_two_page(two_url)
                # uniform: 浮点数,爬取1个电影信息后sleep
                time.sleep(random.uniform(1, 3))
                ins = 'insert into request_finger values(%s)'
                self.cursor.execute(ins,[finger])
                self.db.commit()
            else:
                sys.exit('更新完成')


    # 解析二级页面数据
    def parse_two_page(self,two_url):
        item = {}
        html = self.get_html(two_url)
        re_bds = r'<div class="title_all"><h1><font color=#07519a>(.*?)</font></h1></div>.*?<td style="WORD-WRAP.*?>.*?>(.*?)</a>'
        # two_page_list: [('名称1','ftp://xxxx.mkv')]
        two_page_list = self.re_func(re_bds,html)

        item['name'] = two_page_list[0][0].strip()
        item['download'] = two_page_list[0][1].strip()
        ins = 'insert into filmtab values(%s,%s)'
        film_list = [
            item['name'],item['download']
        ]
        self.cursor.execute(ins,film_list)
        self.db.commit()
        print(film_list)


    def run(self):
        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()
```