数据持久化

- csv

```
import csv
with open('xxx.csv','w',encoding='utf-8',newline='') as f:
    writer = csv.writer(f)
    writer.writerow([])
```

- MongoDB

```
import pymongo

# __init__(self):

    self.conn = pymongo.MongoClient('IP',27017)
    self.db = self.conn['cardb']
    self.myset = self.db['car_set']

# save_html(self,r_list):

    self.myset.insert_one(dict)
```

MySQL

- pymysql

```
import pymysql

db = pymysql.connect('localhost','root','123456','maoyandb',charset='utf8')
cursor = db.cursor()

ins = 'insert into filmtab values(%s,%s,%s)'
cursor.execute(ins,['霸王别姬','张国荣','1993'])

db.commit()
cursor.close()
db.close()
```

```
# __init__(self):
    self.db = pymysql.connect('IP',... ...)
    self.cursor = self.db.cursor()

# save_html(self,r_list):
    self.cursor.execute('sql',[data1])
    self.db.commit()

# run(self):
    self.cursor.close()
    self.db.close()
```

- 练习 - 将电影信息存入MySQL数据库

```
【1】提前建库建表
mysql -h127.0.0.1 -uroot -p123456
create database maoyandb charset utf8;
use maoyandb;
create table maoyantab(
name varchar(100),
star varchar(300),
time varchar(100)
)charset=utf8;

【2】 使用excute()方法将数据存入数据库思路
    2.1) 在 __init__() 中连接数据库并创建游标对象
    2.2) 在 save_html() 中将所抓取的数据处理成列表，使用execute()方法写入
    2.3) 在 run() 中等数据抓取完成后关闭游标及断开数据库连接
```

- 汽车之家二手车信息抓取

```
【1】URL地址
    进入汽车之家官网，点击 二手车
    即: https://www.che168.com/beijing/a0_0msdgscncgpi1lto1cspexx0/

【2】抓取目标
    每辆汽车的
    2.1) 汽车名称
    2.2) 行驶里程
    2.3) 城市
    2.4) 个人还是商家
    2.5) 价格

【3】抓取前5页
```

- 参考答案

```
import requests
import re
import time
```

```python
import random

class CarSpider:
    def __init__(self):
        self.url = 'https://www.che168.com/beijing/a0_0msdgscncgpi1lto1csp{}exx0/?pvareaid=102179#currengpostion'
        self.headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.138 Safari/537.36'}

    def get_html(self, url):
        html = requests.get(url=url, headers=self.headers).content.decode('gb2312', 'ignore')
        self.parse_html(html)

    def parse_html(self, html):
        pattern = re.compile('<li class="cards-li list-photo-li".*?<div class="cards-bottom">.*?<h4 class="card-name">(.*?)</h4>.*?<p class="cards-unit">(.*?)</p>.*?<span class="pirce"><em>(.*?)</em>', re.S)
        car_list = pattern.findall(html)
        self.save_html(car_list)

    def save_html(self, car_list):
        for car in car_list:
            print(car)

    def run(self):
        for i in range(1,6):
            page_url = self.url.format(i)
            self.get_html(page_url)
            time.sleep(random.randint(1,2))

if __name__ == '__main__':
    spider = CarSpider()
    spider.run()
```

MongoDB

- MongoDB特点

```
【1】非关系型数据库,数据以键值对方式存储，端口27017
【2】MongoDB基于磁盘存储
【3】MongoDB数据类型单一,值为JSON文档,而Redis基于内存,
  3.1> MySQL数据类型：数值类型、字符类型、日期时间类型、枚举类型
  3.2> Redis数据类型：字符串、列表、哈希、集合、有序集合
  3.3> MongoDB数据类型：值为JSON文档
【4】MongoDB：库 -> 集合 -> 文档
    MySQL  ：库 -> 表  ->  表记录
```

- MongoDB常用命令

```
Linux进入: mongo
>show dbs                    - 查看所有库
>use 库名                     - 切换库
>show collections            - 查看当前库中所有集合
>db.集合名.find().pretty()   - 查看集合中文档
>db.集合名.count()           - 统计文档条数
>db.集合名.drop()            - 删除集合
>db.dropDatabase()           - 删除当前库
# MongoDB - Commmand - 库->集合->文档

  mongo

>   show dbs
>   use db_name
>   show collections
>   db.集合名.find().pretty()
>   db.集合名.count()
>   db.集合名.drop()
>   db.dropDatabase()
```

- pymongo模块使用

```
import pymongo

# 1.连接对象
conn = pymongo.MongoClient(host = 'localhost',port = 27017)
# 2.库对象
db = conn['maoyandb']
# 3.集合对象
myset = db['maoyanset']
# 4.插入数据库
myset.insert_one({'name':'赵敏'})
```

- 练习 - 将电影信息存入MongoDB数据库

```
"""
猫眼电影top100抓取（电影名称、主演、上映时间）
存入mongodb数据库中
"""
import requests
import re
import time
import random
import pymongo

class MaoyanSpider:
    def __init__(self):

        self.url = 'https://maoyan.com/board/4?offset={}'
```

```
        self.headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/81.0.4044.113 Safari/537.36'}
        # 三个对象：连接对象、库对象、集合对象
        self.conn = pymongo.MongoClient('127.0.0.1', 27017)
        self.db = self.conn['maoyandb']
        self.myset = self.db['maoyanset2']

    def get_html(self, url):
        html = requests.get(url=url, headers=self.headers).text
        # 直接调用解析函数
        self.parse_html(html)

    def parse_html(self, html):
        """解析提取数据"""
        regex = '<div class="movie-item-info">.*?title="(.*?)".*?<p class="star">(.*?)</p>.*?<p class="releasetime">(.*?)</p>'
        pattern = re.compile(regex, re.S)
        r_list = pattern.findall(html)
        # r_list: [('活着','牛犇','2000-01-01'),(),(),...,()]
        self.save_html(r_list)

    def save_html(self, r_list):
        """数据处理函数"""
        for r in r_list:
            item = {}
            item['name'] = r[0].strip()
            item['star'] = r[1].strip()
            item['time'] = r[2].strip()
            print(item)
            # 存入到mongodb数据库
            self.myset.insert_one(item)

    def run(self):
        """程序入口函数"""
        for offset in range(0, 91, 10):
            url = self.url.format(offset)
            self.get_html(url=url)
            # 控制数据抓取频率:uniform()生成指定范围内的浮点数
            time.sleep(random.uniform(0,1))

if __name__ == '__main__':
    spider = MaoyanSpider()
    spider.run()
```

csv

- csv描述

```
【1】作用
   将爬取的数据存放到本地的csv文件中

【2】使用流程
    2.1> 打开csv文件
    2.2> 初始化写入对象
    2.3> 写入数据(参数为列表)

【3】示例代码
    import csv
    with open('sky.csv','w') as f:
        writer = csv.writer(f)
        writer.writerow([])
```

- 示例

```
【1】题目描述
   创建 test.csv 文件，在文件中写入数据

【2】数据写入 - writerow([])方法
   import csv
   with open('test.csv','w') as f:  # with open('test.csv','w',newline='') as f:-
----->windows里面的写法，因为再wiondows中每条数据会有一个空行
        writer = csv.writer(f)
        writer.writerow(['超哥哥','25'])

```

- 练习 - 使用 writerow() 方法将猫眼电影数据存入本地 maoyan.csv 文件

```
【1】在 __init__() 中打开csv文件，因为csv文件只需要打开和关闭1次即可
【2】在 save_html() 中将所抓取的数据处理成列表，使用writerow()方法写入
【3】在run() 中等数据抓取完成后关闭文件
```

- 代码实现

```
"""
猫眼电影top100抓取（电影名称、主演、上映时间）
存入csv文件,使用writerow()方法
"""
import requests
import re
import time
import random
import csv

class MaoyanSpider:
    def __init__(self):
        self.url = 'https://maoyan.com/board/4?offset={}'
```

```
        self.headers = {'User-Agent':'Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Win64; x64; Trident/5.0; .NET CLR 2.0.50727; SLCC2; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0; InfoPath.3; .NET4.0C; Tablet PC 2.0; .NET4.0E)'}
        # 打开文件,初始化写入对象
        self.f = open('maoyan.csv', 'w', newline='', encoding='utf-8')
        self.writer = csv.writer(self.f)

    def get_html(self, url):
        html = requests.get(url=url, headers=self.headers).text
        # 直接调用解析函数
        self.parse_html(html)

    def parse_html(self, html):
        """解析提取数据"""
        regex = '<div class="movie-item-info">.*?title="(.*?)".*?<p class="star">(.*?)</p>.*?<p class="releasetime">(.*?)</p>'
        pattern = re.compile(regex, re.S)
        r_list = pattern.findall(html)
        # r_list: [('活着','牛犇','2000-01-01'),(),(),...,()]
        self.save_html(r_list)

    def save_html(self, r_list):
        """数据处理函数"""
        for r in r_list:
            li = [ r[0].strip(), r[1].strip(), r[2].strip() ]
            self.writer.writerow(li)
            print(li)

    def run(self):
        """程序入口函数"""
        for offset in range(0, 91, 10):
            url = self.url.format(offset)
            self.get_html(url=url)
            # 控制数据抓取频率:uniform()生成指定范围内的浮点数
            time.sleep(random.uniform(1,2))

        # 所有数据抓取并写入完成后关闭文件
        self.f.close()

if __name__ == '__main__':
    spider = MaoyanSpider()
    spider.run()
```