**电影票房爬虫**

# 课程设计内容概述

## 开发环境和开发工具说明

使用Python语言为开发工具，开发环境为Pycharm。

## 课程设计的任务

完成基于Python的十年（2009年~2018年）电影票房前25位数据信息爬虫。  
网址为“[http://www.cbooo.cn](http://www.cbooo.cn/" \t "https://blog.csdn.net/weixin_33951761/article/details/_blank)/year?year=”， 通过代码爬取2009-2018每年票房排名前25的影片名称、类型、总票房（万）、平均票价、场均人次和国家及地区等信息。

## 程序运行顺序说明

在Pycharm中打开程序，首先得安装requests, time, re, csv, bs4等库，接着直接运行pachong.py就行。

# 系统设计方案

对于爬虫获取信息，按照请求获得数据，数据处理，保存数据的步骤进行。



请求获得数据部分：包括获取网站的地址；通过获取User-Agent 来将自己的爬虫程序伪装成由人亲自来完成信息的获取；获取我们所需求的网址的源码，便于我们获取数据；获取响应；获得指定数据。

数据处理部分：有些数据会十分的杂乱，有多余的空格和一些标签等，要将数据中的不需要的东西给去掉。

数据保存部分：通过保存程序将数据存为文件，保存在磁盘中。

# 程序主要代码分析

（1）请求获取数据信息部分：

#循环抓取列表页信息

for year in range(2009,2019):

if year == 2009:

year=str(year)

a=(url+year)

r=requests.get(url=a,headers=headers)

html=r.content

else:

year=str(year)

a=(url+year)

r=requests.get(url=a,headers=headers)

html2=r.content

html = html + html2

（2）数据处理部分

#提取名称、类型、总票房（万）、平均票价、场均人次及国家及地区

result=lj.find\_all('td')

mname=[]

title=""

index=1

year=2009

for i in result:

i=str(i)

title=re.findall(r'</span>(.\*?)</p>',i,re.I|re.M)

if len(title)>0:

mname.append(index)

index=index+1

mname.append(title[0])

else:

info=re.findall(r'<td>(.\*?)</td>',i,re.I|re.M)

mname.append(info[0])

k=0

data=[]

while k<2000:

year=2009

year=year+(k//200)

data.append([mname[k],mname[k+1],mname[k+2],mname[k+3],mname[k+4],

mname[k+5],mname[k+6],mname[k+7],year,1])

k=k+8

1. 数据保存部分

with open('./data.csv','w') as fout:

cin= csv.writer(fout,lineterminator='\n')

for item in data:

cin.writerow(item)

# 程序调试中出现的问题及解决办法

在运行程序的时候，报了SyntaxError: invalid syntax的错误，检查了一下，发现是url=网址这报错了，通过网上查找得知，这里url=的网址应该是以字符串形式，缺少了” ”,加上之后，错误消失了。

在第二次运行程序时，出现了PermissionError: [Errno 13] Permission denied: './data.csv'的错误，由于第一次运行程序时完全正确，没有报错，这就有可能是其他原因导致的错误。错误名为data权限被拒绝，分析可能是没法向文件中写数据的意思，这是发现文件被打开了，而第一次运行程序时，由于没有文件，所以就没有打开，所以错误可能是data.csv被打开了。这时关闭data.csv，重新运行程序，没有报错，说明分析正确。所以，运行程序时，要保存的文件不能被打开。

第一年的表页信息要注意和后面的表页信息分开，第一次是类似于赋值，后面是依次在原基础上加上，与求和相似。

# 主要参考资料

Python网络爬虫实战

Python爬虫开发与项目实战

# 程序及解释

#!/usr/bin/python

# -\*- coding: utf-8 -\*-

import requests

import time

import re

import csv

from bs4 import BeautifulSoup

#设置URL固定部分

url= "http://www.cbooo.cn/year?year="

#设置请求头部信息

headers = {

'User-Agent':

'Mozilla/5.0 (Windows NT 10.0; WOW64) '

'AppleWebKit/537.36 (KHTML, like Gecko) '

'Chrome/55.0.2883.87 Safari/537.36'

}

#循环抓取列表页信息

for year in range(2009,2019):

if year == 2009:

year=str(year)

a=(url+year)

r=requests.get(url=a,headers=headers)

html=r.content

else:

year=str(year)

a=(url+year)

r=requests.get(url=a,headers=headers)

html2=r.content

html = html + html2

#每次间隔0.5秒

time.sleep(0.5)

lj=BeautifulSoup(html,'html.parser')

#提取名称、类型、总票房（万）、平均票价、场均人次及国家及地区

result=lj.find\_all('td')

mname=[]

title=""

index=1

year=2009

for i in result:

i=str(i)

title=re.findall(r'</span>(.\*?)</p>',i,re.I|re.M)

if len(title)>0:

mname.append(index)

index=index+1

mname.append(title[0])

else:

info=re.findall(r'<td>(.\*?)</td>',i,re.I|re.M)

mname.append(info[0])

k=0

data=[]

while k<2000:

year=2009

year=year+(k//200)

data.append([mname[k],mname[k+1],mname[k+2],mname[k+3],mname[k+4],mname[k+5],

mname[k+6],mname[k+7],year,1])

k=k+8

#print(data)

print(len(data))#一共250条数据

#将结果存到CSV文件

with open('./data.csv','w') as fout:

cin= csv.writer(fout,lineterminator='\n')

for item in data:

cin.writerow(item)