

温州大学瓯江学院

爬虫与数据分析实验报告

实验名称:	爬虫期末作业				
班 级:	16 计算机科学与技术三班	姓 名:	刘骋豪	学 号:	16219111333
实验地点:		日 期:			

一、实验目的:

二、实验环境:

- 1.VS code
- 2.python
- 3.Django

三、实验内容和要求:

四、实验步骤:

爬取温州大学新闻页

主要代码:

```
import requests
import datetime
from bs4 import BeautifulSoup
from lxml import etree
from urllib import parse
import MySQLdb
#from sqlcon.models import Wdu

conn=MySQLdb.connect(host="localhost",user="root",passwd="123456",db="testsql",charset="utf8")
cur=conn.cursor()
cur.execute("delete from final_result")
```

```
def getcontent():
for i in range(1640,1648):
link="http://news.wzu.edu.cn/wdxw/"+str(i)+".htm"
```

```
r=requests.get(link)
r.encoding="UTF-8"
```

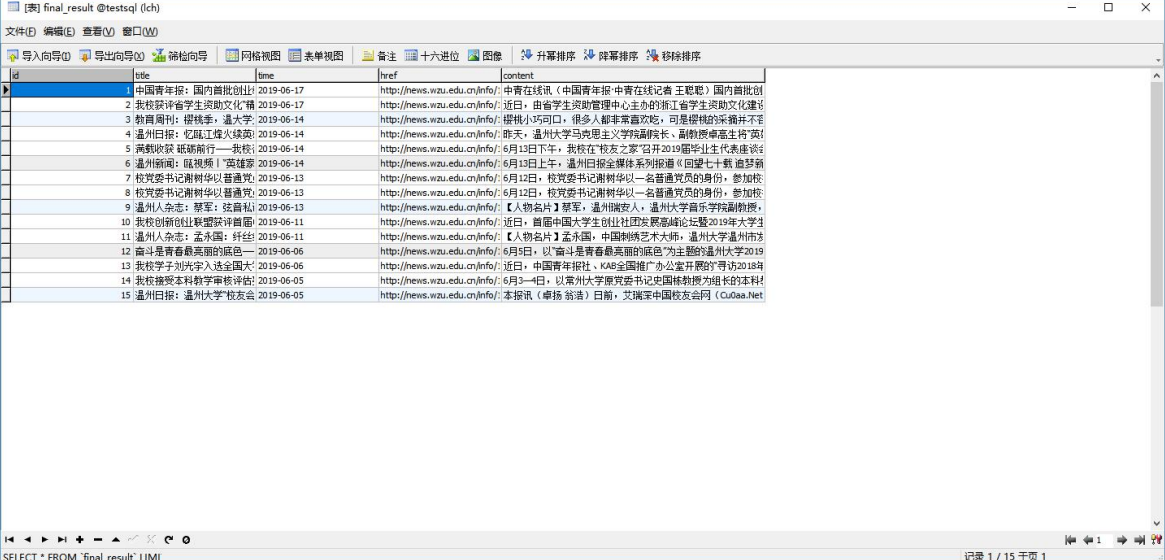
```
tree=etree.HTML(r.text)
time1=tree.xpath("//div[@class='ym']/text()")
time2=tree.xpath("//div[@class='d']/text()")
time=[]
for i in range(len(time1)):
time.append(time1[i]+"-"+time2[i])
```

```
href=tree.xpath("//div[@class='tit']/a/@href")
href=[parse.urljoin(link,i) for i in href]
```

```
title=tree.xpath("//div[@class='tit']/a/@title")
content=tree.xpath("//div[@class='jj']/text()")
```

```
for i in range(len(title)-1):
Title=title[i]
Href=href[i]
Time=time[i]
Content=content[i]
#record=Wdu(title=Title,time=Time,href=Href,content=Content)
#record.save()
cur.execute("insert into final_result(title,time,href,content)
values(\"%s\",\"%s\",\"%s\",\"%s\")"%(Title,Time,Href,Content))
getcontent()
conn.commit()
conn.close()
```

爬取结果（取了 15 条）



The screenshot shows a web browser window with a table containing 15 rows of data. The table has five columns: id, title, time, href, and content. The data represents various news items from Wuzhou University, including awards, student activities, and faculty information.

id	title	time	href	content
1	中国青年报：国内首批创业...	2019-06-17	http://news.wzu.edu.cn/info/...	中青在线讯（中国青年报 中青在线记者 王聪聪）国内首批创...
2	我校获评省学生资助文化建...	2019-06-17	http://news.wzu.edu.cn/info/...	近日，由省学生资助管理中心主办的浙江省学生资助文化建...
3	教育周刊：留桃李，通大学...	2019-06-14	http://news.wzu.edu.cn/info/...	留校小径可口，很多人都非常喜欢吃，可是留校的保藕并不...
4	温州日报：记工匠烽火线...	2019-06-14	http://news.wzu.edu.cn/info/...	昨天，温州大学马克思主义学院院长、副教授傅延生将英...
5	温州晚报 祝福前行——我...	2019-06-14	http://news.wzu.edu.cn/info/...	6月13日下午，我校在“校友之家”召开2019届毕业生代表座谈...
6	温州新闻：陈视频 “陈雄...	2019-06-14	http://news.wzu.edu.cn/info/...	6月13日上午，温州日报全媒体系列报道《回望七十载 追梦新...
7	校党委书记谢树华以普通党...	2019-06-13	http://news.wzu.edu.cn/info/...	6月12日，校党委书记谢树华以一名普通党员的身份，参加校...
8	校党委书记谢树华以普通党...	2019-06-13	http://news.wzu.edu.cn/info/...	6月12日，校党委书记谢树华以一名普通党员的身份，参加校...
9	温州人杂志：蔡军：弦音私...	2019-06-13	http://news.wzu.edu.cn/info/...	【人物名片】蔡军，温州瑞安人，温州大学音乐学院副教授...
10	我校创新创业联盟获评首...	2019-06-11	http://news.wzu.edu.cn/info/...	近日，首届中国大学生创业社团发展高峰论坛暨2019年大学生...
11	温州人杂志：孟永国：丝线...	2019-06-11	http://news.wzu.edu.cn/info/...	【人物名片】孟永国，中国刺绣艺术大师，温州大学温州市...
12	奋斗是青春最美丽的底色—...	2019-06-06	http://news.wzu.edu.cn/info/...	6月5日，以“奋斗是青春最美丽的底色”为主题的温州大学2019...
13	我校学子刘兴学入选全国大...	2019-06-06	http://news.wzu.edu.cn/info/...	近日，中国青年报社、KAP全国推广办公室开展的“寻访2018年...
14	我校播教本科教学审核评估...	2019-06-05	http://news.wzu.edu.cn/info/...	6月3—4日，以温州大学原党委书记史国栋教授为组长的本科...
15	温州日报：温州大学“校友...	2019-06-05	http://news.wzu.edu.cn/info/...	本报讯（廖扬 翁浩）日前，艾瑞安中国校友会网（Cuoaa.Net...

五、实验结果与分析：

Django 配置

创建好 django 项目后打开 setting.py 文件,编辑数据库配置

```
DATABASES = {  
'default': {  
'ENGINE': 'django.db.backends.mysql',  
'NAME': 'python_crawler',  
'USER': 'root',  
'PASSWORD': 'cyb123',  
'HOST': 'localhost',  
'PORT': '3306',  
}  
}
```

使用 cmd 命令, cd 到项目根目录

运行命令 python manage.py migrate 创建相关数据表输入 python manage.py
runserver +本机 ip+ --insecure 即可启动 django 项目

Django 源代码

爬虫以爬取豆瓣 TOP250 电影为例 getmovies.py

```
import requests from bs4 import  
BeautifulSoup import MySQLdb  
import time from crawler.models  
import Movies def get_movies():  
#conn=MySQLdb.connect(host="localhost",user="gjj",passwd="gjj8897",db="python_cra  
wler",charset="utf8") #cur=conn.cursor()  
headers={'user-agent': 'Mozilla/5.0 (Windows NT 6.1;  
Win64;x64) AppleWebKit/537.36 (KHTML,like Gecko) Chrome/52..02743.82  
Safari/537.36','Host': 'movie.douban.com'} for i in range(0,10):  
link='https://movie.douban.com/top250?start='+str(i*25)  
r=requests.get(link,headers=headers,timeout=10)  
soup=BeautifulSoup(r.text,"xml")  
div_list=soup.find_all('div',class_='info') for each  
in div_list:  
url=each.div.a['href']  
title=each.div.a.span.text.strip()  
synopsis=each.contents[3].p.get_text().strip()  
now=time.strftime("%Y-%m-%d %H:%M:%S",time.localtime(time.time()))  
record=Movies(name=title,url=url,synopsis=synopsis,time=now)  
record.save()  
# cur.execute("insert into crawler_movies(name,url,synopsis,time)  
values(%s,%s,%s,%s)",(title,url,synopsis,now)) print("电影爬取成功! ")  
# cur.close()  
# conn.commit()  
# conn.close()
```

引入了 django 的模型, 所以无需配置数据库连接, 直接在 setting.py 修改即可, 但也


```
<a href="http://127.0.0.1:8000/updatabase">更新</a>&nbsp;&nbsp;&nbsp;&nbsp;&nbsp;</p>
<div style="margin:0 auto">
{% block mainbody %}
{{hello}}
{% endblock %}
</div>
</body>
</html>
```

其余 html 继承 index，只需编辑<body>内容例如 movie.html

```
{%extends "index.html" %}
{% block mainbody %}
<h1>豆瓣前 250 部电影</h1>
<div style="text-align:left">
<ol>
{% for movie in movies %}
<li><a
href="{{movie.url}}">{{movie.name}}</a>&nbsp;&nbsp;{{movie.synopsis}}&nbsp;&nbsp;{{movie.time}}</li>
{% endfor %}
</ol>
</div>
{% endblock %}
```

control.py 调用爬虫，实现数据更新

```
from . import getmovies,getphones,getweathers,view import
time
from crawler.models import Movies from
crawler.models import Weathers from
crawler.models import Phones from
django.shortcuts import render
from django.http import HttpResponse
def deletelall(request):
context = {} try:
Movies.objects.all().delete()
Weathers.objects.all().delete()
Phones.objects.all().delete()
context['hello']='删除成功！' except:
context['hello']='删除失败，请先写入数据！'
return render(request, 'index.html', context)
def insertdata(request):
context = {} try:
getmovies.get_movies() time.sleep(2)
getweathers.get_weather() time.sleep(2)
getphones.get_phones('手机')
time.sleep(2) context['hello']='插入成
功！' except:
context['hello']='插入失败！'
return render(request, 'index.html', context)
```

```
def updatabase(request):
deletelall(request)
insertdata(request) context = {}
context['hello']='更新成功！'
return render(request, 'index.html', context)
```

[首页](#) | [豆瓣前250部电影](#) | [各地天气情况](#) | [京东手机](#) | [删除](#) | [插入](#) | [更新](#) |

豆瓣前250部电影

1. 肖申克的救赎 导演: 弗兰克·德拉邦特 Frank Darabont 主演: 蒂姆·罗宾斯 Tim Robbins / ... 1994 / 美国 / 犯罪 剧情 2019-04-25 14:11:25
2. 霸王别姬 导演: 陈凯歌 Kaige Chen 主演: 张国荣 Leslie Cheung / 张丰毅 Fengyi Zha... 1993 / 中国大陆 香港 / 剧情 爱情 同性 2019-04-25 14:11:26
3. 这个杀手不太冷 导演: 吕克·贝松 Luc Besson 主演: 让·雷诺 Jean Reno / 娜塔莉·波特曼 ... 1994 / 法国 / 剧情 动作 犯罪 2019-04-25 14:11:26
4. 阿甘正传 导演: 罗伯特·泽米吉斯 Robert Zemeckis 主演: 汤姆·汉克斯 Tom Hanks / ... 1994 / 美国 / 剧情 爱情 2019-04-25 14:11:26
5. 美丽人生 导演: 罗伯托·贝尼尼 Roberto Benigni 主演: 罗伯托·贝尼尼 Roberto Beni... 1997 / 意大利 / 剧情 喜剧 爱情 战争 2019-04-25 14:11:26
6. 泰坦尼克号 导演: 詹姆斯·卡梅隆 James Cameron 主演: 莱昂纳多·迪卡普里奥 Leonardo... 1997 / 美国 / 剧情 爱情 灾难 2019-04-25 14:11:26
7. 千与千寻 导演: 宫崎骏 Hayao Miyazaki 主演: 柊瑠美 Rumi Hiragi / 入野自由 Miy... 2001 / 日本 / 剧情 动画 奇幻 2019-04-25 14:11:26
8. 辛德勒的名单 导演: 史蒂文·斯皮尔伯格 Steven Spielberg 主演: 连姆·尼森 Liam Neeson... 1993 / 美国 / 剧情 历史 战争 2019-04-25 14:11:26
9. 盗梦空间 导演: 克里斯托弗·诺兰 Christopher Nolan 主演: 莱昂纳多·迪卡普里奥 Le... 2010 / 美国 英国 / 剧情 科幻 悬疑 冒险 2019-04-25 14:11:26
10. 忠犬八公的故事 导演: 莱塞·霍尔斯特姆 Lasse Hallström 主演: 理查·基尔 Richard Ger... 2009 / 美国 英国 / 剧情 2019-04-25 14:11:26
11. 机器人总动员 导演: 安德鲁·斯坦顿 Andrew Stanton 主演: 本·贝尔特 Ben Burt / 艾丽... 2008 / 美国 / 爱情 科幻 动画 冒险 2019-04-25 14:11:26
12. 三傻大闹宝莱坞 导演: 拉库马·希拉尼 Rajkumar Hirani 主演: 阿米尔·汗 Aamir Khan / 卡... 2009 / 印度 / 剧情 喜剧 爱情 歌舞 2019-04-25 14:11:26
13. 海上钢琴师 导演: 朱塞佩·托纳多雷 Giuseppe Tornatore 主演: 蒂姆·罗斯 Tim Roth / ... 1998 / 意大利 / 剧情 音乐 2019-04-25 14:11:26
14. 放牛班的春天 导演: 克里斯托夫·巴拉蒂 Christophe Barratier 主演: 热拉尔·朱尼奥 Gé... 2004 / 法国 瑞士 德国 / 剧情 音乐 2019-04-25 14:11:26
15. 楚门的世界 导演: 彼得·威尔 Peter Weir 主演: 金·凯瑞 Jim Carrey / 劳拉·琳妮 Lau... 1998 / 美国 / 剧情 科幻 2019-04-25 14:11:26
16. 大话西游之大圣娶亲 导演: 刘镇伟 Jeffrey Lau 主演: 周星驰 Stephen Chow / 吴孟达 Man Tat Ng... 1995 / 香港 中国大陆 / 喜剧 爱情 奇幻 古装 2019-04-25 14:11:26
17. 星际穿越 导演: 克里斯托弗·诺兰 Christopher Nolan 主演: 马修·麦康纳 Matthew Mc... 2014 / 美国 英国 加拿大 冰岛 / 剧情 科幻 冒险 2019-04-25 14:11:26
18. 龙猫 导演: 宫崎骏 Hayao Miyazaki 主演: 日高法子 Noriko Hidaka / 坂本千夏 Ch... 1988 / 日本 / 动画 奇幻 冒险 2019-04-25 14:11:26
19. 教父 导演: 弗朗西斯·福特·科波拉 Francis Ford Coppola 主演: 马龙·白兰度 M... 1972 / 美国 / 剧情 犯罪 2019-04-25 14:11:26
20. 熔炉 导演: 黄东赫 Dong-hyuk Hwang 主演: 孔侑 Yoo Gong / 郑有美 Yu-mi Jeong ... 2011 / 韩国 / 剧情 2019-04-25 14:11:26
21. 无间道 导演: 刘伟强 / 麦兆辉 主演: 刘德华 / 梁朝伟 / 黄秋生 2002 / 香港 / 剧情 犯罪 悬疑 2019-04-25 14:11:26
22. 疯狂动物城 导演: 拜伦·霍华德 Byron Howard / 瑞奇·摩尔 Rich Moore 主演: 金妮弗·... 2016 / 美国 / 喜剧 动画 冒险 2019-04-25 14:11:26
23. 当幸福来敲门 导演: 加布里埃尔·穆奇 Gabriele Muccino 主演: 威尔·史密斯 Will Smith ... 2006 / 美国 / 剧情 传记 家庭 2019-04-25 14:11:26

```
#selenium 验证 12306
```

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.action_chains import ActionChains
import requests
import base64
import re
import time

class Demo():
    def __init__(self,username,password):
        self.coordinate=[[-105,-20],[-35,-20],[40,-20],[110,-20],[-105,50],[-35,50],[40,50],[110,50]]
        self.username=username
        self.password=password
    def login(self):
        login_url="https://kyfw.12306.cn/otn/resources/login.html"
        driver = webdriver.Chrome("python_crawler\chromedriver.exe")
        driver.set_window_size(1200, 900)
        driver.get(login_url)
```

```

account=driver.find_element_by_class_name("login-hd-account")
account.click()
userName=driver.find_element_by_id("J-username")
userName.send_keys(self.username)
password=driver.find_element_by_id("J-password")
password.send_keys(self.password)
self.driver=driver
def getVerifyImage(self):
    try:

        img_element =WebDriverWait(self.driver, 100).until(
            EC.presence_of_element_located((By.ID, "J-loginImg"))
        )

```

```

except Exception as e:
    print(u"网络开小差,请稍后尝试")
    base64_str=img_element.get_attribute("src").split(",")[-1]
    imgdata=base64.b64decode(base64_str)
    with open('verify.jpg','wb') as file:
        file.write(imgdata)
    self.img_element=img_element
def getVerifyResult(self):
    url="http://littlebigluo.qicp.net:47720/"
    response=requests.request("POST",url,data={"type":"1"},files={'pic_x
xfile':open('verify.jpg','rb')})
    result=[]
    print(response.text)
    for i in re.findall("<B>(.*?)</B>",response.text)[0].split(" "):
        result.append(int(i)-1)
    self.result=result
    print(result)
def moveAndClick(self):
    try:
        Action=ActionChains(self.driver)
        for i in self.result:
            Action.move_to_element(self.img_element).move_by_offset(self
.coordinate[i][0],self.coordinate[i][1]).click()
            Action.perform()
    except Exception as e:
        print(e.message())
def submit(self):
    self.driver.find_element_by_id("J-login").click()
def __call__(self):
    self.login()
    time.sleep(5)
    self.getVerifyImage()

```

```
time.sleep(3)
self.getVerifyResult()
time.sleep(3)
self.moveAndClick()
time.sleep(3)
self.submit()
time.sleep(10000)
```



六：思考题：

七、教师评语：

实验成绩：

教师：（签名要全称）

年 月 日