温州大学瓯江学院理工分院

爬虫与数据分析 实验报告

实验名称：	爬虫与 django				
班　　级：	16 计算机四班	姓　　名：	魏梦佳	学　　号：	16219111427
实验地点：	403	日　　期：	2019.4.23		

一、实验目的：

1、掌握 python
2、掌握 jango
3、新闻添加和读取

二、实验环境：

有 myeclipse、tomcat 等环境的计算机若干台

三、实验内容和要求：

1.爬静态网页，如豆瓣电影top250
2.爬动态网页 ，如京东商城
3.jango展示内容
4.上传github

四、实验步骤：

五、实验结果与分析（含程序、数据记录及分析和实验总结等）：

1.

代码：

```
from bs4 import BeautifulSoup
import pymysql
import requests
import re
import os

def connect_db():
    connect = pymysql.connect(
        user="root",
        password="admin",
        host="localhost",
        db="aaa",
        port=3306,
        charset=("utf8"),
        use_unicode=True,
    )
    return connect
```

```
def get_html(web_url):
    header = {
        "User-Agent": "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.108 Safari/537.36 2345Explorer/8.5.1.15355"}
    html = requests.get(url=web_url, headers=header).text
    Soup = BeautifulSoup(html, "lxml")
    data = Soup.find("ol").find_all("li")
    return data

def get_info(all_move, connect, cursor):
    for info in all_move:

        nums = re.findall(r'<em class="">\d+</em>', str(info), re.S | re.M)
        nums = re.findall(r'\d+', str(nums), re.S | re.M)
        num = nums[0]


        names = info.find("span")
        name = names.get_text()


        charactors = info.find("p")
        charactor = charactors.get_text().replace(" ", "").replace("\n", "")
        charactor = charactor.replace("\xa0", "").replace("\xee", "").replace("\xf6", "").replace("\u0161", "").replace("\xf4", "").replace("\xfb", "").replace("\u2027", "")


        scores = info.find_all("span", {"class": "rating_num"})
        score = scores[0].get_text()

        data = {'num':num, 'name':name, 'charactor':charactor,  'score':score}
        print(data)

        cursor.execute("insert into douban(num,name,charactor,score)values(%s,%s,%s,%s)",
                       [data['num'], data['name'], data['charactor'],data['score']])

        connect.commit()
    return

if __name__ == "__main__":
    connect = connect_db()
    cursor = connect.cursor()
    page = 0
    while page <= 225:
        web_url = "https://movie.douban.com/top250?start=%s&filter=" % page
        all_move = get_html(web_url)
        data = get_info(all_move, connect, cursor)
        page += 25

    connect.close()
```

截图：

id	name	charactor	score
1	肖申克的救	导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗宾斯TimRobbi	9.60
2	霸王别姬	导演:陈凯歌KaigeChen主演:张国荣LeslieCheung/张丰毅Fengyi	9.60
3	这个杀手不	导演:吕克·贝松LucBesson主演:让·雷诺JeanReno/娜塔莉·波特曼..	9.40
4	阿甘正传	导演:罗伯特·泽米吉斯RobertZemeckis主演:汤姆·汉克斯TomHan	9.40
5	美丽人生	导演:罗伯托·贝尼尼RobertoBenigni主演:罗伯托·贝尼尼Roberto	9.50
6	泰坦尼克号	导演:詹姆斯·卡梅隆JamesCameron主演:莱昂纳多·迪卡普里奥Le	9.30
7	千与千寻	导演:宫崎骏HayaoMiyazaki主演:柊瑠美RumiHragi/入野自由Mi	9.30
8	辛德勒的名	导演:史蒂文·斯皮尔伯格StevenSpielberg主演:连姆·尼森LiamNee	9.50
9	盗梦空间	导演:克里斯托弗·诺兰ChristopherNolan主演:莱昂纳多·迪卡普里	9.30
10	忠犬八公的	导演:莱塞·霍尔斯道姆LasseHallstrm主演:理查·基尔RichardGer...	9.30
11	机器人总动	导演:安德鲁·斯坦顿AndrewStanton主演:本·贝尔特BenBurtt/艾丽	9.30
12	三傻大闹宝	导演:拉库马·希拉尼RajkumarHirani主演:阿米尔·汗AamirKhan/卡	9.20
13	海上钢琴师	导演:朱塞佩·托纳多雷GiuseppeTornatore主演:蒂姆·罗斯TimRot	9.20
14	放牛班的春	导演:克里斯托夫·巴拉蒂ChristopheBarratier主演:热拉尔·朱尼奥(	9.30
15	楚门的世界	导演:彼得·威尔PeterWeir主演:金·凯瑞JimCarrey/劳拉·琳妮Lau...	9.20
16	大话西游之	导演:刘镇伟JeffreyLau主演:周星驰StephenChow/吴孟达ManTat	9.20
17	星际穿越	导演:克里斯托弗·诺兰ChristopherNolan主演:马修·麦康纳Matthe	9.20
18	龙猫	导演:宫崎骏HayaoMiyazaki主演:日高法子NorikoHidaka/坂本千	9.20
19	教父	导演:弗朗西斯·福特·科波拉FrancisFordCoppola主演:马龙·白兰度	9.20
20	熔炉	导演:黄东赫Dong-hyukHwang主演:孔侑YooGong/郑有美Yu-mi	9.30
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	9.10
22	疯狂动物城	导演:拜伦·霍华德ByronHoward/瑞奇·摩尔RichMoore主演:金妮弗	9.20

id	name	charactor	score
229	勇闯夺命岛	导演:迈克尔·贝MichaelBay主演:肖恩·康纳利SeanConnery/尼古拉	8.60
230	变脸	导演:吴宇森JohnWoo主演:约翰·特拉沃尔塔JohnTravolta/尼古拉	8.40
231	发条橙	导演:StanleyKubrick主演:MalcolmMcDowell/PatrickMagee/M	8.50
232	功夫	导演:周星驰StephenChow主演:周星驰StephenChow/元秋QiuYu	8.30
233	秒速5厘米	导演:新海诚MakotoShinkai主演:水桥研二KenjiMizuhashi/近藤	8.30
234	黄金三镖客	导演:SergioLeone主演:ClintEastwood/EliWallach/LeeVanCleef	9.10
235	黑鹰坠落	导演:雷德利·斯科特RidleyScott主演:乔什·哈奈特JoshHartnett/...	8.60
236	非常嫌疑犯	导演:布莱恩·辛格BryanSinger主演:史蒂芬·鲍德温StephenBaldwi	8.60
237	卡萨布兰卡	导演:迈克尔·柯蒂兹MichaelCurtiz主演:亨弗莱·鲍嘉HumphreyBo	8.60
238	我爱你	导演:秋昌民Chang-minChoo主演:宋在河Jae-hoSong/李顺载So	9.00
239	国王的演讲	导演:汤姆·霍珀TomHooper主演:柯林·菲尔斯ColinFirth/杰弗里·...	8.40
240	千钧一发	导演:安德鲁·尼科尔AndrewNiccol主演:伊桑·霍克EthanHawke/乌	8.70
241	疯狂的麦克	导演:乔治·米勒GeorgeMiller主演:汤姆·哈迪TomHardy/查理兹·塞	8.60
242	美国丽人	导演:萨姆·门德斯SamMendes主演:凯文·史派西KevinSpacey/安妮	8.50
243	遗愿清单	导演:罗伯·莱纳RobReiner主演:杰克·尼科尔森JackNicholson/摩	8.60
244	奇迹男孩	导演:斯蒂芬·卓博斯基StephenChbosky主演:雅各布·特伦布莱Jac	8.60
245	碧海蓝天	导演:LucBesson主演:让·马克·巴尔Jean-MarcBarr/让·雷诺JeanR	8.70
246	驴得水	导演:周申ShenZhou/刘露LuLiu主演:任素汐SuxiRen/大力DaLi...2	8.30
247	荒岛余生	导演:罗伯特·泽米吉斯RobertZemeckis主演:汤姆·汉克斯TomHan	8.50
248	枪火	导演:杜琪峰JohnnieTo主演:吴镇宇FrancisNg/任达华SimonYam	8.70
249	英国病人	导演:安东尼·明格拉AnthonyMinghella主演:拉尔夫·费因斯RalphF	8.50
250	荒野生存	导演:西恩·潘SeanPenn主演:埃米尔·赫斯基EmileHirsch/马西娅·...	8.60

SELECT * FROM `gizhongzuoye_movie` LIMIT 0, 1000

2.

代码：

```
from selenium.webdriver import Chrome
from config import *
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from time import sleep
from selenium.common.exceptions import NoSuchElementException
import pymysql
```

```python
def next_page(client, wait, page_num):
    END_PAGE = 1

    while len(client.find_elements_by_class_name('gl-item')) < 60:
        client.execute_script('window.scrollTo(0, document.body.scrollHeight)')
        sleep(1)
    print("[+] 第{}加载完成".format(page_num))

    parse_page(page_num, client)


    page_num += 1
    if page_num > END_PAGE:
        print('前{}页爬取成功'.format(END_PAGE))
        return

    wait.until(
        EC.presence_of_element_located(
            (By.CSS_SELECTOR, '#J_bottomPage > span.p-skip > input')
        )
    )

    wait.until(
        EC.element_to_be_clickable(
            (By.CSS_SELECTOR, '#J_bottomPage.> span.p-skip > a')
        )
    )

    input_ = client.find_element_by_css_selector('#J_bottomPage > span.p-skip > input')
    input_.clear()
    input_.send_keys(page_num)
    input_.send_keys(Keys.ENTER)
    wait.until(
        EC.text_to_be_present_in_element(
            (By.CSS_SELECTOR, '#J_bottomPage > span.p-num > a.curr'),
            str(page_num)
        )
    )
    next_page(client, wait, page_num)
```

```python
def connect_db():
    connect = pymysql.connect(
        user="root",
        password="admin",
        host="localhost",
        db="aaa",
        port=3306,
        charset=("utf8"),
        use_unicode=True,
    )
    return connect
def parse_page(page_num, client):
    print("开始解析第{}页数据".format(page_num))
    items = client.find_elements_by_class_name('gl-item')
    index = 1

    for item in items:
        print("[{}] ".format(index), end="")
         try:
            title = item.find_element_by_css_selector("div.p-name > a > em").text
        except NoSuchElementException:
            title = None
        try:
            price = item.find_element_by_css_selector("div.p-price > strong > i").text
        except NoSuchElementException:
            price = None


        try:
            comment = item.find_element_by_css_selector(".p-commit a").text
        except NoSuchElementException:
            comment = None

        print("{} >>> {}   >>> {}".format(title, price, comment))

        connect=connect_db()
        cursor = connect.cursor()
        cursor.execute("insert into jd(id,title,price,comment)values (%s,%s,%s,%s)",(index,title, price, comment))
        connect.commit()
        index += 1
    print("解析第{}页数据完成".format(page_num))
    connect.close()
```

```
def search(client, url, keyword,wait):

    client.get(url)

    wait.until(
        EC.presence_of_element_located(
            (By.ID, 'key')
        )
    )

    wait.until(
        EC.element_to_be_clickable(
            (By.CSS_SELECTOR, '#search > div > div.form > button > i')
        )
    )

    input_ = client.find_element_by_id('key')
    input_.send_keys(keyword)

    botton = client.find_element_by_css_selector('#search > div > div.form > button > i')
    botton.click()
    print("搜索完成")


    page_num = 1
    next_page(client, wait, page_num)

def main():

    client = Chrome()
    url = "http://www.jd.com"

    KEYWORD = '手机'

    wait = WebDriverWait(client, 10)
    search(client, url, KEYWORD, wait)

if __name__ == '__main__':
    main()
```

截图：

id	title	price	comment
1	荣耀V20	2799.00	22万+
2	Apple iP	5899.00	二手有售
3	【KPL官	3298.00	二手有售
4	vivo U1 :	799.00	13万+
5	荣耀10青	1299.00	二手有售
6	荣耀8X 手	1299.00	二手有售
7	vivo X27	3598.00	二手有售
8	小米 红米	1199.00	55万+
9	OPPO R	2999.00	1200+
10	小米 红米	799.00	78万+
11	小米8SE	1399.00	二手有售
12	荣耀畅玩	899.00	40万+
13	小米 红米	799.00	5.2万+
14	Apple iP	6199.00	二手有售
15	vivo Z3 i	1598.00	二手有售
16	Apple iP	3799.00	二手有售
17	OPPO R	2999.00	0
18	黑鲨游戏	3499.00	3.9万+
19	Apple iP	9699.00	二手有售
20	小米9 xia	3299.00	11万+
21	Apple iP	4699.00	二手有售
22	华为 HU	3999.00	二手有售

id	title	price	comment
39	华为 HU	5488.00	1.5万+
40	三星 Gal	1449.00	1.7万+
41	Apple iP	4199.00	二手有售
42	华为 HU	1499.00	1.3万+
43	荣耀畅玩	799.00	16万+
44	三星 Gal	6999.00	二手有售
45	小米8屏	2499.00	二手有售
46	魅族 Not	1398.00	1.5万+
47	Apple iP	7599.00	二手有售
48	【销量20	599.00	2000+
49	小米9 SE	1999.00	二手有售
50	华为 HU	3088.00	二手有售
51	HUAWEI	949.00	二手有售
52	荣耀 畅玩	599.00	二手有售
53	荣耀Not	2599.00	二手有售
54	华为 HU	4588.00	二手有售
55	华为 HU	4499.00	二手有售
56	vivo【新	3598.00	2400+
57	Apple iP	3099.00	二手有售
58	小米6X 红	1299.00	二手有售
59	OPPO R	2999.00	二手有售
60	酷派（Co	599.00	3700+

3.

代码：

Cmd

创建一个项目

```
D:\360downloads>django-admin startproject mysite
```

创建一个 app

```
D:\360downloads\mysite>python manage.py startapp zuoye
```

运行

```
D:\360downloads\mysite>python manage.py runserver
```

创建更改的文件

```
D:\360downloads\mysite>python manage.py makemigrations
```

在数据库中生成对应与 models 对应的的表

```
D:\360downloads\mysite>python manage.py migrate
```

setting.py

settings.py

```
75  # https://docs.djangoproject.com/en/2.1/ref/settings/#databases
76
77  DATABASES = {
78      'default': {
79          'ENGINE': 'django.db.backends.mysql',
80          'NAME':'aaa', #数据库名称
81          'USER':'root',
82          'PASSWORD':'admin',
83          'HOST':'localhost',
84          'PORT':'3306',
85      }
86  }
87
```

settings.py

```
55  TEMPLATES = [
56      {
57          'BACKEND': 'django.template.backends.django.DjangoTemplates',
58          'DIRS': [os.path.join(BASE_DIR,'templates')],
59          'APP_DIRS': True,
60          'OPTIONS': {
61              'context_processors': [
62                  'django.template.context_processors.debug',
63                  'django.template.context_processors.request',
64                  'django.contrib.auth.context_processors.auth',
65                  'django.contrib.messages.context_processors.messages',
66              ],
67          },
68      },
69  ]
```

settings.py

```
32
33  INSTALLED_APPS = [
34      'django.contrib.admin',
35      'django.contrib.auth',
36      'django.contrib.contenttypes',
37      'django.contrib.sessions',
38      'django.contrib.messages',
39      'django.contrib.staticfiles',
40      'qizhongzuoye',
41  ]
42
```

project 下的 urls.py

urls.py

```
12  Including another URLconf
13      1. Import the include() function: from django.urls import include, path
14      2. Add a URL to urlpatterns:  path('blog/', include('blog.urls'))
15  """
16  from django.contrib import admin
17  from django.urls import path
18  from django.conf.urls import include,url
19  from qizhongzuoye import views
20
21  urlpatterns = [
22      path('admin/', admin.site.urls),
23      url(r'^qizhongzuoye/',include('qizhongzuoye.urls'))
24  ]
25
```

app 下的 urls.py

urls.py

```
1  from django.contrib import admin
2  from django.urls import path
3  from django.conf.urls import url
4  from qizhongzuoye import views
5  urlpatterns = [
6      url(r'^douban/', views.douban),#第一个douban表示html，第二个表示views.py的函数
7      url(r'^jingdong/', views.jingdong),
8      url(r'^index/',views.index),
9  ]
```

models.py

models.py

```python
from django.db import models

class MOVIE(models.Model):
    id=models.BigIntegerField
    name = models.CharField(max_length=255)
    charactor = models.CharField(max_length=255, blank=True, null=True)
    score=models.CharField(max_length=255)
    def __str__(self):
        return self.id

class Mobeil(models.Model):
    id=models.BigIntegerField
    title = models.CharField(max_length=255)
    price = models.CharField(max_length=255)
    comment = models.CharField(max_length=255, blank=True, null=True)
    def __str__(self):
        return self.id
# Create your models here.

```

在 models.py 创建的表插入数据

查询创建工具 查询编辑器

```
insert into zuoye_name select * from blog_top250
```

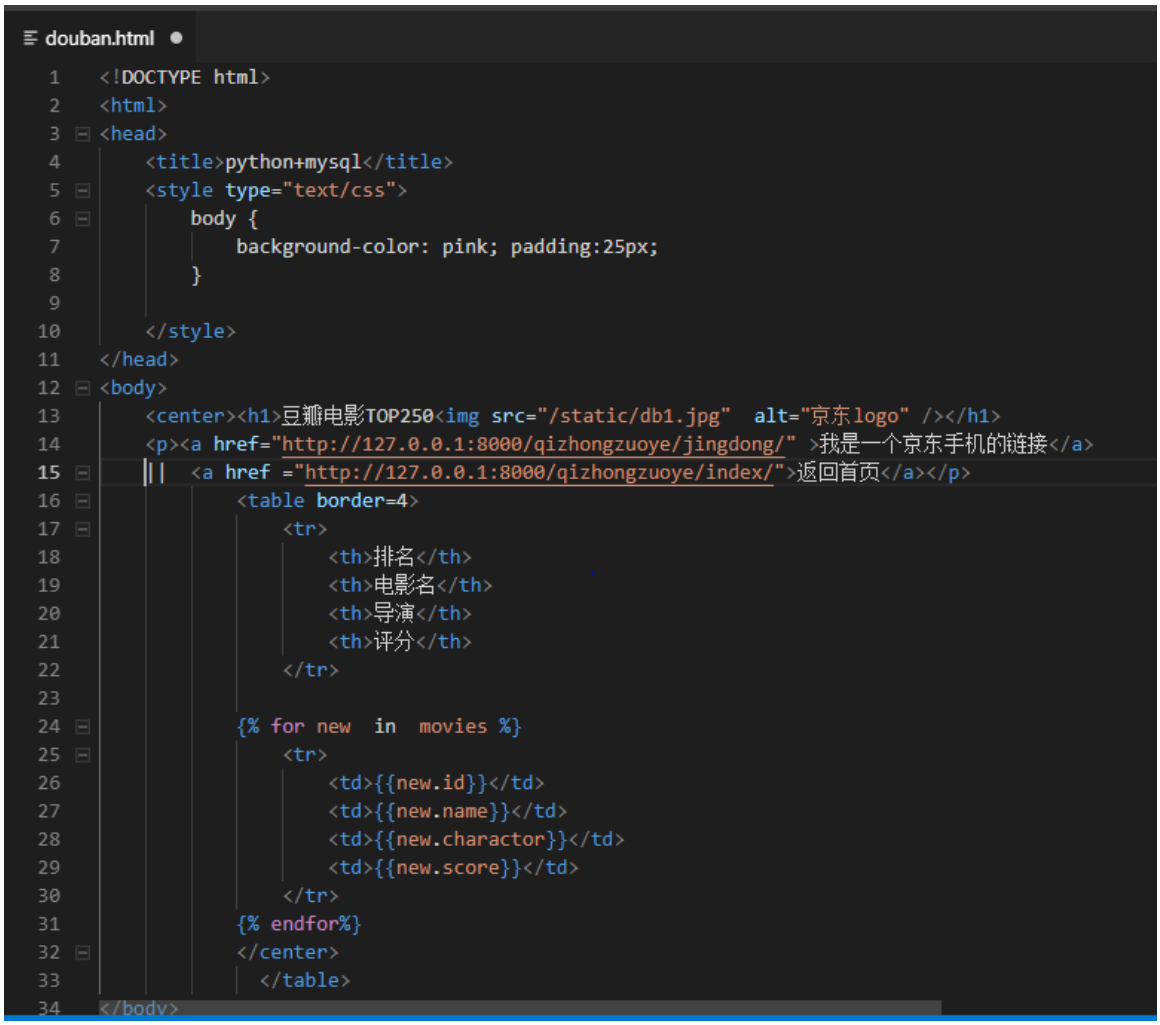
Views.py

views.py

```python
from django.shortcuts import render
from django.http import HttpResponse
from qizhongzuoye.models import MOVIE
from django.db import models
from qizhongzuoye.models import Mobeil
import MySQLdb
import datetime
from django.template.loader import get_template
from django.core.paginator import Paginator
# Create your views he

def douban(request):
  # movies=movie.objects.all()
   movies=MOVIE.objects.all()
   return render(request,'qizhongzuoye/douban.html',
   {
      'movies':movies,
   })
def jingdong(request):

  # movies=movie.objects.all()
   mobeils=Mobeil.objects.all()
   return render(request,'qizhongzuoye/jingdong.html',
   {
      'mobeils':mobeils,
   })

def index(request):
   return render(request,'qizhongzuoye/index.html')
# Create your views here.

```

Douban.html

douban.html



```
<!DOCTYPE html>
<html>
<head>
    <title>python+mysql</title>
    <style type="text/css">
        body {
            background-color: pink; padding:25px;
        }

    </style>
</head>
<body>
    <center><h1>豆瓣电影TOP250<img src="/static/db1.jpg"  alt="京东logo" /></h1>
    <p><a href="http://127.0.0.1:8000/qizhongzuoye/jingdong/" >我是一个京东手机的链接</a>
    <a href ="http://127.0.0.1:8000/qizhongzuoye/index/">返回首页</a></p>
            <table border=4>
                <tr>
                    <th>排名</th>
                    <th>电影名</th>
                    <th>导演</th>
                    <th>评分</th>
                </tr>

            {% for new  in  movies %}
                <tr>
                    <td>{{new.id}}</td>
                    <td>{{new.name}}</td>
                    <td>{{new.charactor}}</td>
                    <td>{{new.score}}</td>
                </tr>
            {% endfor%}
            </center>
              </table>
</body>
```

Jingdong.html

douban.html ● jingdong.html ●

```
<!DOCTYPE html>
<html>
<head>
    <title>python+selenium+mysql</title>
</head>
<body>
    <center><h1>京东手机排行<img src="/static/jd.jpg"  alt="京东logo" /></h1>
    <p><a href="http://127.0.0.1:8000/qizhongzuoye/douban/" >我是一个豆瓣电影top250的链接</a>
     ||  <a href ="http://127.0.0.1:8000/qizhongzuoye/index/">返回首页</a></p>

            <table border="6">

                    <tr>
                       <th>排名</th>
                       <th>手机名</th>
                       <th>价格</th>
                       <th>销量</th>
                       </tr>

                       {% for a  in  mobeils %}
                           <tr>
                               <td>{{a.id}}</td>
                               <td>{{a.title}}</td>
                               <td>{{a.price}}</td>
                               <td>{{a.comment}}</td>
                           </tr>
                       {% endfor%}
                       </table></center>
</body>

```

Index.html

douban.html ● jingdong.html ● index.html ●

```
    <title>
        This is my index
    </title>
</head>
<body>
    <center>
        <h1>
        <a href="http://127.0.0.1:8000/qizhongzuoye/jingdong/" >查看京东手机排行</a><br>
        <a href="http://127.0.0.1:8000/qizhongzuoye/douban/" >查看豆瓣电影top250</a>
        </h1>
    </center>
</body>
</html>
```

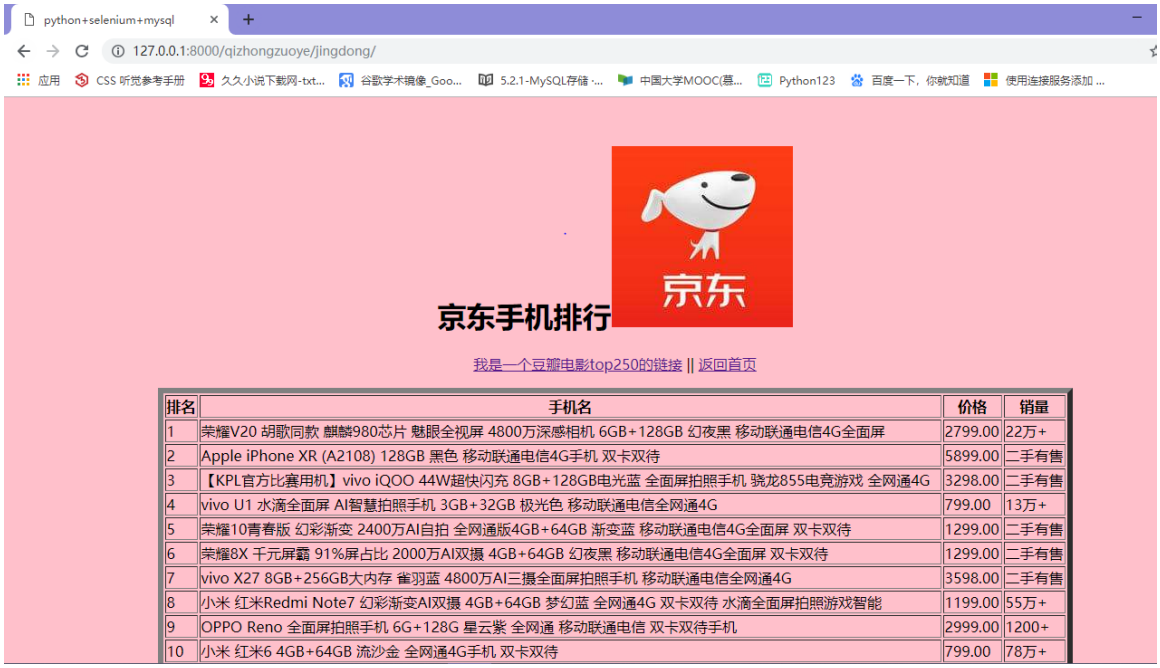
截图：

Index. html

Douban. html

排名	电影名	导演	评分
1	肖申克的救赎	导演:弗兰克·德拉邦特FrankDarabont主演:蒂姆·罗宾斯TimRobbins/...1994/美国/犯罪剧情	9.60
2	霸王别姬	导演:陈凯歌KaigeChen主演:张国荣LeslieCheung/张丰毅FengyiZha...1993/中国大陆香港/剧情爱情同性	9.60
3	这个杀手不太冷	导演:吕克·贝松LucBesson主演:让·雷诺JeanReno/娜塔莉·波特曼...1994/法国/剧情动作犯罪	9.40
4	阿甘正传	导演:罗伯特·泽米吉斯RobertZemeckis主演:汤姆·汉克斯TomHanks/...1994/美国/剧情爱情	9.40
5	美丽人生	导演:罗伯托·贝尼尼RobertoBenigni主演:罗伯托·贝尼尼RobertoBeni...1997/意大利/剧情喜剧爱情战争	9.50
6	泰坦尼克号	导演:詹姆斯·卡梅隆JamesCameron主演:莱昂纳多·迪卡普里奥Leonardo...1997/美国/剧情爱情灾难	9.30
7	千与千寻	导演:宫崎骏HayaoMiyazaki主演:柊瑠美RumiHragi/入野自由Miy...2001/日本/剧情动画奇幻	9.30
8	辛德勒的名单	导演:史蒂文·斯皮尔伯格StevenSpielberg主演:连姆·尼森LiamNeeson...1993/美国/剧情历史战争	9.50
9	盗梦空间	导演:克里斯托弗·诺兰ChristopherNolan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	9.30
10	忠犬八公的故事	导演:莱塞·霍尔斯道姆LasseHallstrm主演:理查·基尔RichardGer...2009/美国英国/剧情	9.30

Jingdong.html



排名	手机名	价格	销量
1	荣耀V20 胡歌同款 麒麟980芯片 魅眼全视屏 4800万深感相机 6GB+128GB 幻夜黑 移动联通电信4G全面屏	2799.00	22万+
2	Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G手机 双卡双待	5899.00	二手有售
3	【KPL官方比赛用机】vivo iQOO 44W超快闪充 8GB+128GB电光蓝 全面屏拍照手机 骁龙855电竞游戏 全网通4G	3298.00	二手有售
4	vivo U1 水滴全面屏 AI智慧拍照手机 3GB+32GB 极光色 移动联通电信全网通4G	799.00	13万+
5	荣耀10青春版 幻彩渐变 2400万AI自拍 全网通版4GB+64GB 渐变蓝 移动联通电信4G全面屏 双卡双待	1299.00	二手有售
6	荣耀8X 千元屏霸 91%屏占比 2000万AI双摄 4GB+64GB 幻夜黑 移动联通电信4G全面屏 双卡双待	1299.00	二手有售
7	vivo X27 8GB+256GB大内存 雀羽蓝 4800万AI三摄全面屏拍照手机 移动联通电信全网通4G	3598.00	二手有售
8	小米 红米Redmi Note7 幻彩渐变AI双摄 4GB+64GB 梦幻蓝 全网通4G 双卡双待 水滴全面屏拍照游戏智能	1199.00	55万+
9	OPPO Reno 全面屏拍照手机 6G+128G 星云紫 全网通 移动联通电信 双卡双待手机	2999.00	1200+
10	小米 红米6 4GB+64GB 流沙金 全网通4G手机 双卡双待	799.00	78万+

六：思考题：

七、教师评语：

实验成绩：　　教师：邱儒　　2019 年 04 月 08 日