温州大學瓯江學院

WENZHOU UNIVERSITY OUJIANG COLLEGE

《爬虫期中作业》

题　　目：网络爬虫的数据爬取

分　　院：理工分院

班　　级：16 计算机科学与技术 2 班

姓　　名：陈祥牛

学　　号：16219111204

完成日期：2019 年 4 月 26 日

温州大学瓯江学院教务部

二〇一二年十一月制

目录

一． 豆瓣 TOP250 电影网页的静态爬虫

1. 用 request 库将网页代码爬取下来：

```
def main():
    url='https://movie.douban.com/top250'
    flag=0
    while flag<250:
        html=getHTMLtext(url,flag)
        time.sleep(2)
        ulist=[]
        getText(ulist,html)
        Save(ulist)
        flag=flag+25
```

```
def getHTMLtext(url,flag):
    try:
        if(flag==0):
                kk={}
        else:
                kk={'start':flag,'filter':''}
        r=requests.get(url,params=kk)
        r.raise_for_status()
        r.encoding='utf-8'
        return r.text
    except:
        return ""
```

2. 用 BeautifulSoup 库将相应的内容爬去下来并保存到列表里：

```
def main():
    url='https://movie.douban.com/top250'
    flag=0
    while flag<250:
        html=getHTMLtext(url,flag)
        time.sleep(2)
        ulist=[]
        getText(ulist,html)
        Save(ulist)
        flag=flag+25
```

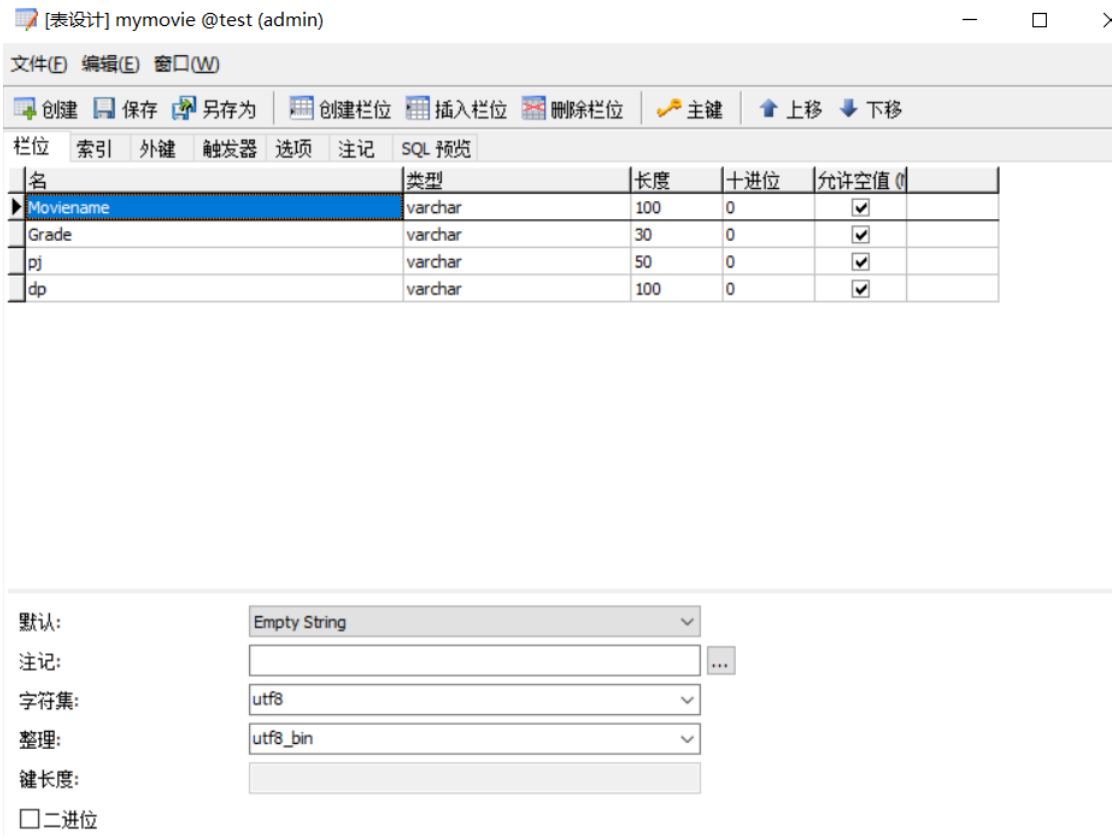
```
def getText(ulist,html):
    soup=BeautifulSoup(html,"html.parser")
    a=soup.find('ol',attrs={'class':'grid_view'})

    for flag in a.find_all('li'):
        s={}
        hd=flag.find('div',attrs={'class':'hd'})
        name=hd.find('span',attrs={'class':'title'}).getText()
        s['Moviename']=name
        grade=flag.find('span',attrs={'class':'rating_num'}).getText()
        s['Grade']=grade
        star=flag.find('div',attrs={'class':'star'})
        pj=star.find_all('span')[-1].getText()
        s['pj']=pj
        dp=flag.find('span',attrs={'class':'inq'})
        if(dp):
                s['dp']=dp.getText()
        else:
                s['dp']='无短评'

        ulist.append(s)
```

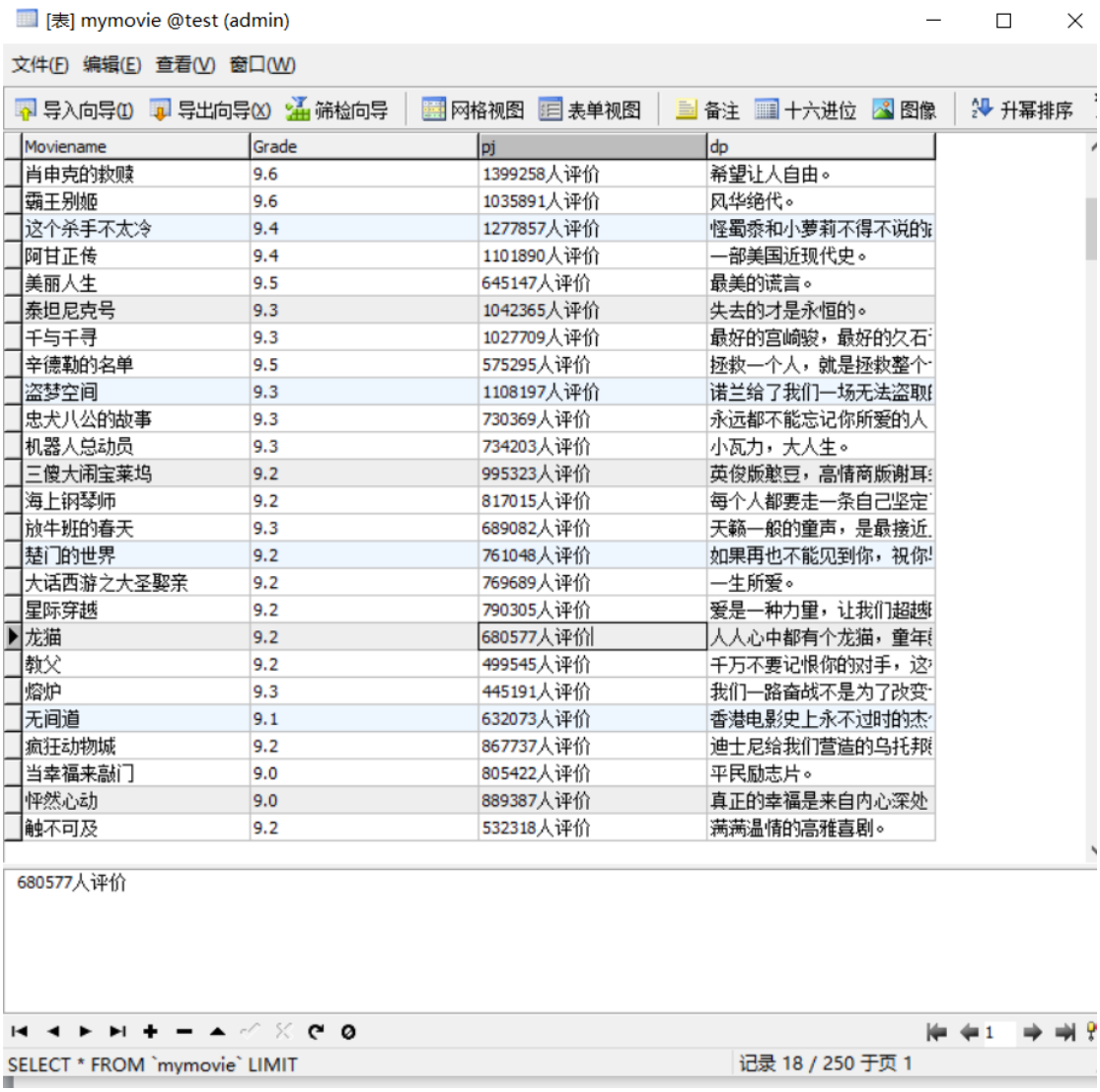
3. 将爬取的内容输入到数据库：

（1） 首先在 Mysql 中创建数据表名为 mymovie:



（2） 用代码将列表内数据输入到数据库中：

```python
def Save(ulist):
    try:
        db=pymysql.connect(host="localhost",user="root",password="admin",database="test",charset="utf8")
        cursor=db.cursor()
        ls={}
        for ls in ulist:
            cursor.execute("insert into mymovie(Moviename,Grade,pj,dp) values(%s,%s,%s,%s)",(ls['Moviename'],ls['Grade'],ls['
            db.commit()
        cursor.close()
        db.close()
    except:
        print("错误！")
```



Moviename	Grade	pj	dp
肖申克的救赎	9.6	1399258人评价	希望让人自由。
霸王别姬	9.6	1035891人评价	风华绝代。
这个杀手不太冷	9.4	1277857人评价	怪蜀黍和小萝莉不得不说的
阿甘正传	9.4	1101890人评价	一部美国近现代史。
美丽人生	9.5	645147人评价	最美的谎言。
泰坦尼克号	9.3	1042365人评价	失去的才是永恒的。
千与千寻	9.3	1027709人评价	最好的宫崎骏，最好的久石
辛德勒的名单	9.5	575295人评价	拯救一个人，就是拯救整个
盗梦空间	9.3	1108197人评价	诺兰给了我们一场无法盗取
忠犬八公的故事	9.3	730369人评价	永远都不能忘记你所爱的人
机器人总动员	9.3	734203人评价	小瓦力，大人生。
三傻大闹宝莱坞	9.2	995323人评价	英俊版憨豆，高情商版谢耳
海上钢琴师	9.2	817015人评价	每个人都要走一条自己坚定
放牛班的春天	9.3	689082人评价	天籁一般的童声，是最接近
楚门的世界	9.2	761048人评价	如果再也不能见到你，祝你
大话西游之大圣娶亲	9.2	769689人评价	一生所爱。
星际穿越	9.2	790305人评价	爱是一种力量，让我们超越
龙猫	9.2	680577人评价	人人心中都有个龙猫，童年
教父	9.2	499545人评价	千万不要记恨你的对手，这
熔炉	9.3	445191人评价	我们一路奋战不是为了改变
无间道	9.1	632073人评价	香港电影史上永不过时的杰
疯狂动物城	9.2	867737人评价	迪士尼给我们营造的乌托邦
当幸福来敲门	9.0	805422人评价	平民励志片。
怦然心动	9.0	889387人评价	真正的幸福是来自内心深处
触不可及	9.2	532318人评价	满满温情的高雅喜剧。

（3）。将数据库内的内容输出并且展示在 django 框架中：

```python
def Mysqlfind():
    db=pymysql.connect(host="localhost",user="root",password="admin",database="test",charset="utf8",cursorclass=pymysql.cursors.DictCurs
    cursor=db.cursor()
    cursor.execute("select * from mymovie")
    ulist=cursor.fetchall()
    cursor.close()
    db.close()
    return ulist
```

因为爬取的数据已经输入到数据库，所以输入数据库代码可以省略：

```python
def hello(request):
    ulist=Mysqlfind()
    return render(request, 'hello.html',{'movie':ulist})
def hello1(request):
```

url.py 文件：

```python
"""
from django.urls import path

from . import view

urlpatterns = [
    path('hello/', view.hello),
    path('hello1/', view.hello1),
]
```

Hello.html 文件：

view.py　base.html　s1.py　settings.py

```
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8">
<title>豆瓣电影爬取页面</title>
</head>
<body>
    <h1>豆瓣电影TOP250</h1>
    <table>
        <tr>
            <td>电影名称</td>
            <td>电影评分</td>
            <td>电影评价</td>
            <td>电影短评</td>
        </tr>
        {% for Mymovie in movie %}
        <tr>
            <td>{{Mymovie.Moviename}}</td>
            <td>{{Mymovie.Grade}}</td>
            <td>{{Mymovie.pj}}</td>
            <td>{{Mymovie.dp}}</td>
        </tr>
        {% endfor %}
    </table>
</body>
</html>
```

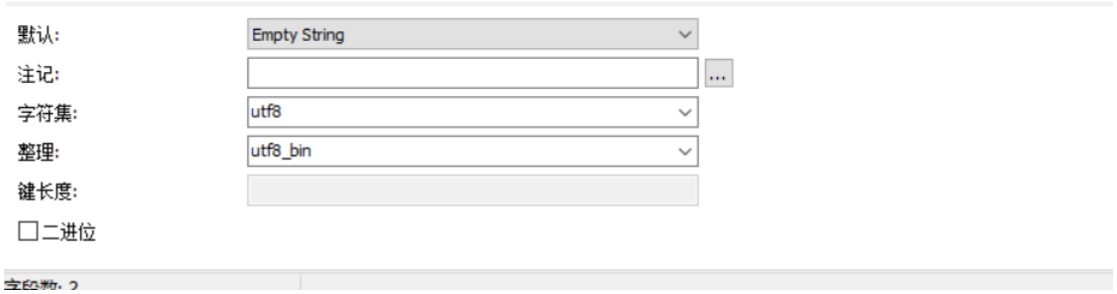
运行后的效果：

二. 动态爬取京东搜索手机信息

1. 用 selenium 直接获取所需信息并保存到列表中：

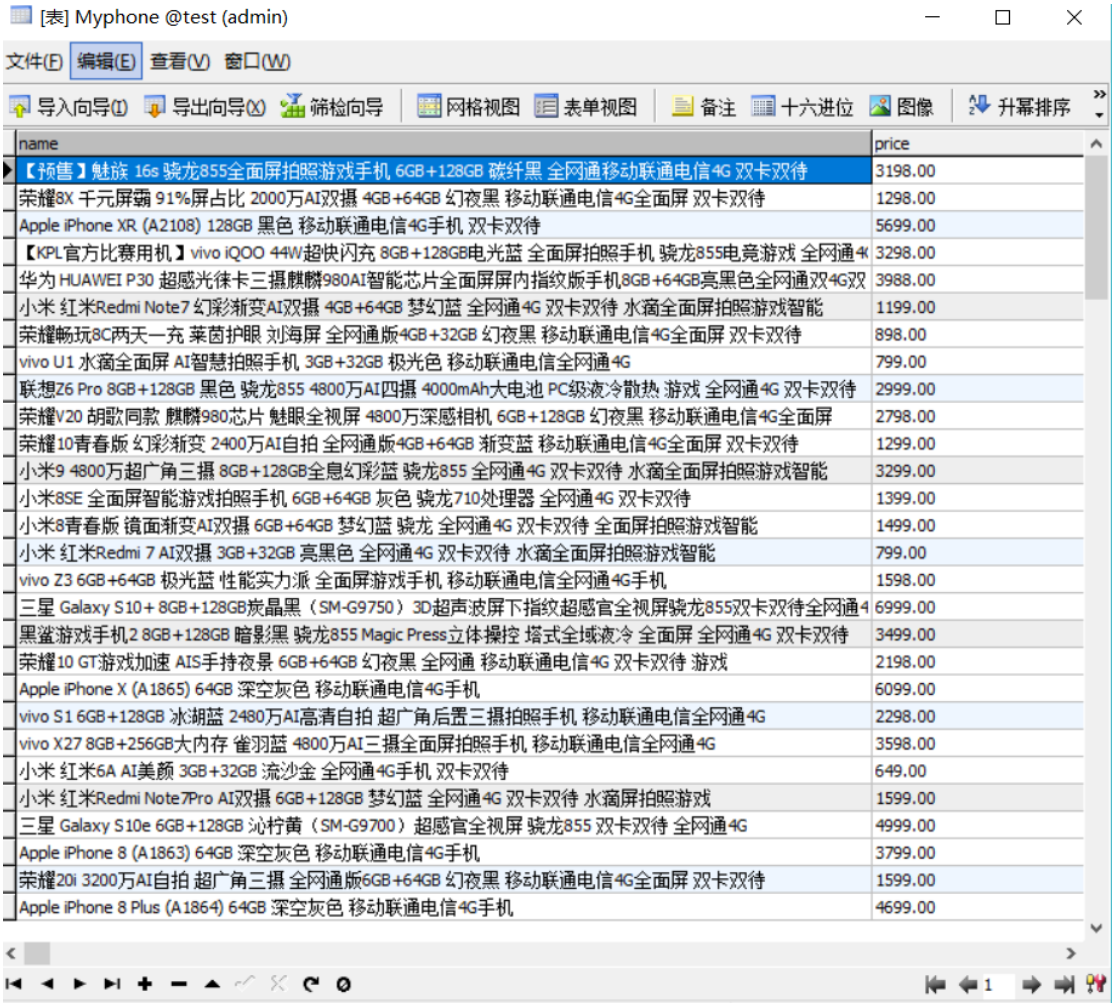
```
def dgetText(url):
    driver = webdriver.Chrome()
    driver.get(url)
    for i in range(10):
        driver.execute_script("var q=document.documentElement.scrollTop={0}".format(i*1000))
        time.sleep(1)
    searchprice=driver.find_elements_by_xpath('//*[@id="J_goodsList"]/ul/li/div/div[3]/strong/i')
    searchname=driver.find_elements_by_xpath('//*[@id="J_goodsList"]/ul/li/div/div[4]/a/em')
    names=[]
    for a in searchname:
        name=a.text
        names.append(name)
    prices=[]
    for c in searchprice:
        price=c.text
        prices.append(price)
    ulists=[]
    flag=0
    for b in names:
        ulist={}
        ulist['name']=b
        ulist['price']=prices[flag]
        flag=flag+1
        ulists.append(ulist)
    return ulists
```

2. 创建名为 Myphone 的数据表：



3. 将数据保存到 Myphone 数据表中：

```
def dSave(ulist):
        try:
                db=pymysql.connect(host="localhost",user="root",password="admin",database="test",charset="utf8")
                cursor=db.cursor()
                ls={}
                for ls in ulist:
                        cursor.execute("insert into Myphone(name,price) values(%s,%s)",(ls['name'],ls['price']))
                        db.commit()
                cursor.close()
                db.close()
        except:
                print("错误！")
def main():
    key='手机'
    url='https://search.jd.com/Search?keyword='+key+'&enc=utf-8'
    ulists=[]
    ulists=dgetText(url)
    dSave(ulists)
```



name	price
【预售】魅族 16s 骁龙855全面屏拍照游戏手机 6GB+128GB 碳纤黑 全网通移动联通电信4G 双卡双待	3198.00
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	1298.00
Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G手机 双卡双待	5699.00
【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4	3298.00
华为 HUAWEI P30 超感光徕卡三摄麒麟980AI智能芯片全面屏屏内指纹版手机8GB+64GB亮黑色全网通双4G双	3988.00
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
荣耀畅玩8C两天一充 莱茵护眼 刘海屏 全网通版4GB+32GB 幻夜黑 移动联通电信4G全面屏 双卡双待	898.00
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
联想Z6 Pro 8GB+128GB 黑色 骁龙855 4800万AI四摄 4000mAh大电池 PC级液冷散热 游戏 全网通4G 双卡双待	2999.00
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	2798.00
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
小米9 4800万超广角三摄 8GB+128GB全息幻彩蓝 骁龙855 全网通4G 双卡双待 水滴全面屏拍照游戏智能	3299.00
小米8SE 全面屏智能游戏拍照手机 6GB+64GB 灰色 骁龙710处理器 全网通4G 双卡双待	1399.00
小米8青春版 镜面渐变AI双摄 6GB+64GB 梦幻蓝 骁龙 全网通4G 双卡双待 全面屏拍照游戏智能	1499.00
小米 红米Redmi 7 AI双摄 3GB+32GB 亮黑色 全网通4G 双卡双待 水滴全面屏拍照游戏智能	799.00
vivo Z3 6GB+64GB 极光蓝 性能实力派 全面屏游戏手机 移动联通电信全网通4G手机	1598.00
三星 Galaxy S10+ 8GB+128GB炭晶黑（SM-G9750）3D超声波屏下指纹超感官全视屏骁龙855双卡双待全网通4	6999.00
黑鲨游戏手机2 8GB+128GB 暗影黑 骁龙855 Magic Press立体操控 塔式全域液冷 全面屏 全网通4G 双卡双待	3499.00
荣耀10 GT游戏加速 AIS手持夜景 6GB+64GB 幻夜黑 全网通 移动联通电信4G 双卡双待 游戏	2198.00
Apple iPhone X (A1865) 64GB 深空灰色 移动联通电信4G手机	6099.00
vivo S1 6GB+128GB 冰湖蓝 2480万AI高清自拍 超广角后置三摄拍照手机 移动联通电信全网通4G	2298.00
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
小米 红米6A AI美颜 3GB+32GB 流沙金 全网通4G手机 双卡双待	649.00
小米 红米Redmi Note7Pro AI双摄 6GB+128GB 梦幻蓝 全网通4G 双卡双待 水滴屏拍照游戏	1599.00
三星 Galaxy S10e 6GB+128GB 沁柠黄（SM-G9700）超感官全视屏 骁龙855 双卡双待 全网通4G	4999.00
Apple iPhone 8 (A1863) 64GB 深空灰色 移动联通电信4G手机	3799.00
荣耀20i 3200万AI自拍 超广角三摄 全网通版6GB+64GB 幻夜黑 移动联通电信4G全面屏 双卡双待	1599.00
Apple iPhone 8 Plus (A1864) 64GB 深空灰色 移动联通电信4G手机	4699.00

4. 将数据表内的内容输出到 django 框架：

```
def hello1(request):
        '''key='手机'
        url='https://search.jd.com/Search?keyword='+key+'&enc=utf-8'
        ulists=[]
        ulists=dgetText(url)
        dSave(ulists)'''
        ulist=dMysqlfind()
        return render(request, 'base.html',{'phone':ulist})
```

5. urls.py 文件：

```python
from django.urls import path

from . import view

urlpatterns = [
    path('hello/', view.hello),
    path('hello1/', view.hello1),
]
```

Base.html 文件：

view.py | base.html | s1.py | settings.py

```html
<!DOCTYPE html>
<html>
<head>
<meta charset="utf-8">
<title>动态爬取京东手机页面</title>
</head>
<body>
    <h1>京东手机</h1>
    <table>
        <tr>
            <td>手机名称</td>
            <td>手机价格</td>
        </tr>
        {% for Myphone in phone %}
        <tr>
            <td>{{Myphone.name}}</td>
            <td>{{Myphone.price}}</td>
        </tr>
        {% endfor %}
    </table>
</body>
</html>
```

运行后输出的网页：

127.0.0.1:8000/hello1/

京东手机

手机名称	手机价格
【预售】魅族 16s 骁龙855全面屏拍照游戏手机 6GB+128GB 碳纤黑 全网通移动联通电信4G 双卡双待	3198.00
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	1298.00
Apple iPhone XR (A2108) 128GB 黑色 移动联通电信4G手机 双卡双待	5699.00
【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
华为 HUAWEI P30 超感光徕卡三摄麒麟980AI智能芯片全面屏屏内指纹版手机8GB+64GB亮黑色全网通双4G双	3988.00
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
荣耀畅玩8C两天一充 莱茵护眼 刘海屏 全网通版4GB+32GB 幻夜黑 移动联通电信4G全面屏 双卡双待	898.00
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
联想Z6 Pro 8GB+128GB 黑色 骁龙855 4800万AI四摄 4000mAh大电池 PC级液冷散热 游戏 全网通4G 双卡双待	2999.00
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	2798.00
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
小米9 4800万超广角三摄 8GB+128GB全息幻彩蓝 骁龙855 全网通4G 双卡双待 水滴全面屏拍照游戏智能	3299.00
小米8SE 全面屏智能游戏拍照手机 6GB+64GB 灰色 骁龙710处理器 全网通4G 双卡双待	1399.00
小米8青春版 镜面渐变AI双摄 6GB+64GB 梦幻蓝 骁龙 全网通4G 双卡双待 全面屏拍照游戏智能	1499.00
小米 红米Redmi 7 AI双摄 3GB+32GB 亮黑色 全网通4G 双卡双待 水滴全面屏拍照游戏智能	799.00
vivo Z3 6GB+64GB 极光蓝 性能实力派 全面屏游戏手机 移动联通电信全网通4G手机	1598.00
三星 Galaxy S10+ 8GB+128GB炭晶黑（SM-G9750）3D超声波屏下指纹超感官全视屏骁龙855双卡双待全网通4G	6999.00
黑鲨游戏手机2 8GB+128GB 暗影黑 骁龙855 Magic Press立体操控 塔式全域液冷 全面屏 全网通4G 双卡双待	3499.00
荣耀10 GT游戏加速 AIS手持夜景 6GB+64GB 幻夜黑 全网通 移动联通电信4G 双卡双待 游戏	2198.00
Apple iPhone X (A1865) 64GB 深空灰色 移动联通电信4G手机	6099.00
vivo S1 6GB+128GB 冰湖蓝 2480万AI高清自拍 超广角后置三摄拍照手机 移动联通电信全网通4G	2298.00
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
小米 红米6A AI美颜 3GB+32GB 流沙金 全网通4G手机 双卡双待	649.00
小米 红米Redmi Note7Pro AI双摄 6GB+128GB 梦幻蓝 全网通4G 双卡双待 水滴屏拍照游戏	1599.00
三星 Galaxy S10e 6GB+128GB 沁柠黄（SM-G9700）超感官全视屏 骁龙855 双卡双待 全网通4G	4999.00
Apple iPhone 8 (A1863) 64GB 深空灰色 移动联通电信4G手机	3799.00