

第5讲--淘宝、天猫数据挖掘

任务目标

相关知识

- 1、注册和登录淘数据平台
- 2、数据的清洗，选择"冲饮制品"。
- 3、二值化处理
- 4、通过淘数据下载数据

第5讲--淘宝、天猫数据挖掘

任务目标

- 1、登录淘数据网站，选择商品，普通用户可以浏览数据，无下载数据的权限
- 2、安装Selenium工具的驱动程序Webdriver
- 3、通过Selenium框架进行搜索

相关知识

- 1、Selenium模块可以自动提交信息、自动跳转页面。

1、注册和登录淘数据平台

- 1、登录<https://www.taosj.com/>淘数据（行业数据）注册，用自己的手机号码注册，然后登录系统。

The screenshot shows the Taosj.com website interface. At the top, there is a navigation bar with links for '产品' (Products), '解决方案' (Solutions), '购买服务' (Buy Services), '电商工具' (E-commerce Tools), '帮助中心' (Help Center), and 'COVID-19'. There are also links for '登录' (Login) and '注册' (Register). The main content area features a large banner with the text '淘宝天猫全行业、品牌、店铺、直播、预售数据' and '抖音快手数据，跨境数据，以及更多电商解决方案'. Below the banner, there is a section for '提交定制需求' (Submit Custom Request) with input fields for '您的职位或称呼' (Your position or title), '您的手机号码' (Your mobile phone number), '您的公司名称' (Your company name), and '您的数据需求' (Your data requirements). There is also a section for '您是什么平台用户?' (Which platform are you a user of?) with checkboxes for '淘宝天猫' (Taobao/Tmall), '淘宝直播' (Taobao Live), '小红书' (Xiaohongshu), '京东' (JD.com), '抖音' (Douyin), '快手' (Kuaishou), '细分市场?' (Sub-market?), '品牌库?' (Brand library?), '跨境电商' (Cross-border e-commerce), and '定制报告' (Custom report). A '在线提交' (Submit online) button is at the bottom of the form. The footer contains two sections: '电商数据' (E-commerce data) with links for '淘宝数据' (Taobao data), '淘宝直播' (Taobao Live), '跨境电商' (Cross-border e-commerce), '抖音数据' (Douyin data), and '快手数据' (Kuaishou data); and '解决方案' (Solutions) with links for '品牌公司' (Brand company), '电商企业' (E-commerce company), '金融证券' (Financial securities), '投资行研' (Investment research), '咨询调研' (Consulting research), and '政府高校' (Government and universities).

- 2、通过电话号码登录淘数据平台



3、免费用户是无法导出数据，只能浏览数据。

淘数据的数据全部来自公开网页，提供的数据仅供参考，不作为投资、经营决策的依据。

淘数据 行业数据 [淘数据交流群](#) [帮助中心](#)

当前结果: 分类: 百货食品 一级行业: 咖啡/麦片/冲饮 子行业: 纯牛奶 品牌: 搜索品牌关键词 平台: 全网

行业数据已更新到2020年10月，11月部分数据已更新到2020年11月7日，11月全部数据需在2020年12月4日后全量更新。

删除宝贝排行

日期筛选: 2020/07

关键词: 请输入宝贝ID、宝贝名称

免费用户无法导出数据

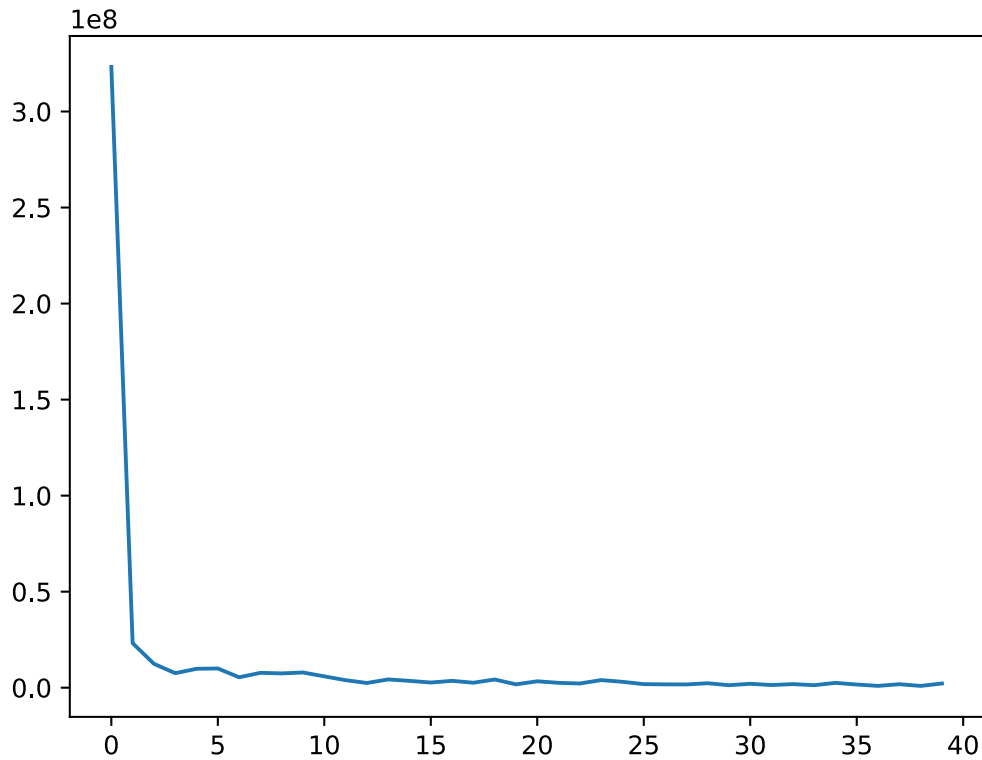
宝贝图	宝贝名称/掌柜/信用	品牌	店铺名称	类目	营销推广	标价	成交均价	销售量	销售金额	操作
	包邮光明纯奶250mL*24盒/箱 掌柜: 天猫超市	光明	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内	¥59.9	¥59.9	210258	¥12594454.2	查看详情 直接详情 查看淘宝地址 >>
	超定制 伊利金典纯牛奶 250mL*24盒/箱 掌柜: 天猫超市	伊利	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内 外	¥109	¥109	174979	¥19072711	查看详情 直接详情 查看淘宝地址 >>
	包邮伊利高钙低脂奶 250mL*24盒/箱 掌柜: 天猫超市	伊利	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内	¥74.4	¥74.4	163024	¥12128985.6	查看详情 直接详情 查看淘宝地址 >>
	包邮蒙牛纯牛奶 PU RE MILK 250mL*16 掌柜: 天猫超市	蒙牛	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内	¥49.9	¥49.9	162071	¥8087342.9	查看详情 直接详情 查看淘宝地址 >>
	包邮蒙牛特仑苏纯牛奶 48250mL*16和/全 掌柜: 天猫超市	蒙牛	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内 外	¥86	¥86	142234	¥12232124	查看详情 直接详情 查看淘宝地址 >>
	包邮伊利无糖砖纯牛奶 48250mL*24盒/箱 掌柜: 天猫超市	伊利	天猫超市	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	聚 内	¥72	¥72	135670	¥9768240	查看详情 直接详情 查看淘宝地址 >>
	蒙牛特仑苏纯牛奶 250mL*12盒 掌柜: 好食期超市旗舰店	蒙牛	好食期超市旗舰店	咖啡/麦片/冲饮 > 乳制品 > 纯牛奶	外	¥68	¥48.62	127658	¥6206348.93	查看详情 直接详情 查看淘宝地址 >>

2、数据的清洗，选择"冲饮制品"。

1、数据的整理和清洗，淘数据不支持爬虫获取数据，由于采集的数据量并不大，采用手工的方式下载数据。

2、读取淘数据下载 冲饮制品 的商品，通过sublime text3将复制的数据拷贝到bao.csv文件，数据要保存为UTF-8的编码。

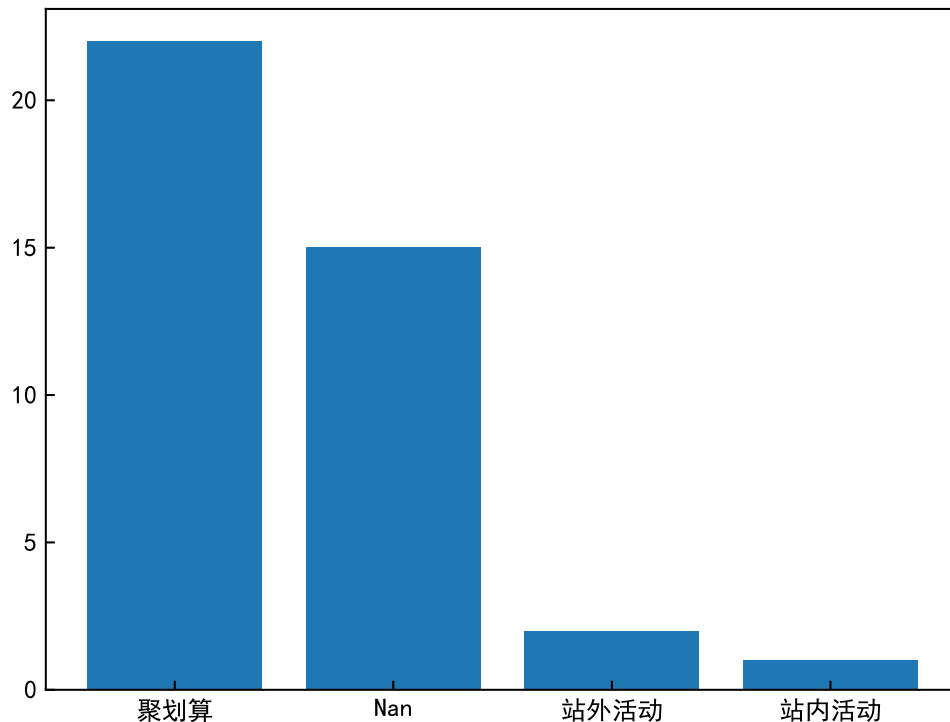
```
import pandas as pd
data = pd.read_csv('tao.csv', sep=',');
data.head();
money = data['销售金额'].apply(lambda x: x.split('¥')[1])
money = money.astype('float');
plt.plot(money);
plt.show();
```



```
import numpy as np;
plt.rcParams['font.sans-serif'] = ['SimHei'];
plt.rcParams['axes.unicode_minus'] = False;
plt.rcParams['xtick.direction'] = 'in';
plt.rcParams['ytick.direction'] = 'in';
money = np.log(money);
store = data['店铺名称/掌柜/信用'].apply(lambda x: x.split('/')[0]);
ax = plt.gca();
plt.plot(money);
plt.xlabel('商家', fontsize=15);
plt.ylabel('销售金额(对数变换)', fontsize=15);
ax.set_xticks(range(store.count()));
ax.set_xticklabels(store, rotation=50);
plt.tick_params(axis='both', labelsize=12, color='red');
plt.show();
```

3、销量第一名的商家的销量太高，可以使用 $\ln x$ 函数对数据进行对数变换。


```
x = data['营销推广'];
x = x.astype('str');
t.value_counts()    #对聚划算，站内活动，站外活动三种营销方式进行计数
plt.bar(['聚划算', 'Nan', '站外活动', '站内活动'], [22, 15, 2, 1]);
plt.show();
```



3、直方图量化

```
import time
from selenium import webdriver
from selenium.webdriver.support.ui import Select
import matplotlib.pyplot as plt;
import numpy as np;
import math

plt.rcParams['font.sans-serif'] = ['SimHei'] # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负号

wd = webdriver.Chrome()
url = 'http://cas.nbut.edu.cn/cas/login'
wd.get(url)
time.sleep(2)
we_account = wd.find_element_by_css_selector('#username')
we_account.clear()
we_account.send_keys("用户名")

we_password = wd.find_element_by_css_selector('#password')
we_password.clear()
we_password.send_keys("密码")

wd.find_element_by_css_selector('.login_box_landing_btn').click()
```

```
time.sleep(10)
wd.get('http://i.nbut.edu.cn')
wd.close()
```

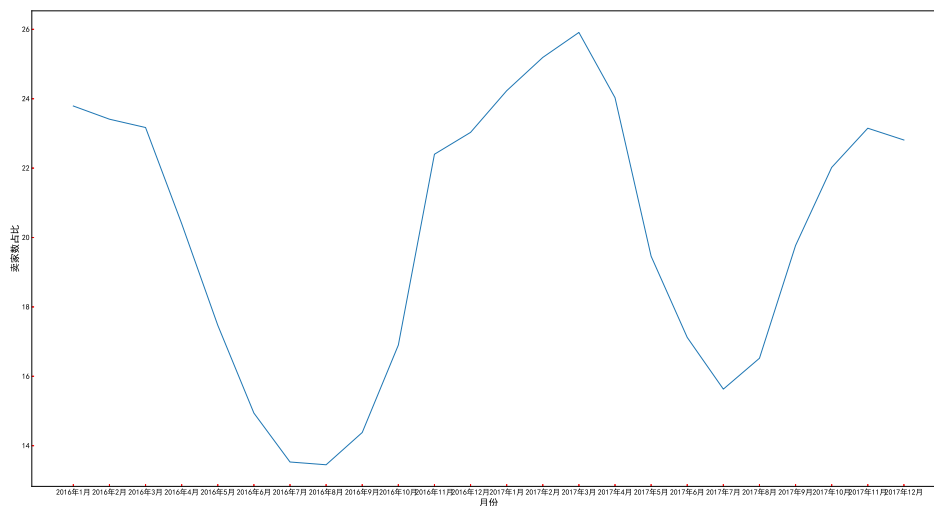
4、通过淘数据下载数据

1、淘宝数据的处理，选择'风衣'数据，时间选择2017年数据。

```
import matplotlib.pyplot as plt;
import numpy as np;

plt.rcParams['font.sans-serif'] = ['SimHei'] # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负号
plt.rcParams['xtick.direction'] = 'in' #x的刻度向内
plt.rcParams['ytick.direction'] = 'in' #y的刻度向内

data = pd.read_csv('taobao.csv', sep=',');
x = data[(data['行业名称']=='风衣') & (data['年份']=='2017')];
x = data[(data['行业名称']=='风衣') & (data['年份']=='2017')];
x1 = x['卖家数占比'].apply(lambda x: x.split('%')[0]);
x1 = x1.astype('float');
month = x['年份'].map(str) + '年' + x['月份'].map(str);
plt.plot(month, x1);
plt.xlabel('月份', fontsize=15);
plt.ylabel('卖家数占比', fontsize=15);
plt.tick_params(axis='both', labelsize=12, color='red');
plt.show();
```



2、筛选羽绒服数据。

```
import matplotlib.pyplot as plt;
import numpy as np;

plt.rcParams['font.sans-serif'] = ['SimHei'] # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负号
plt.rcParams['xtick.direction'] = 'in' #x的刻度向内
plt.rcParams['ytick.direction'] = 'in' #y的刻度向内
```

```

data = pd.read_csv('taobao.csv', sep=',');
x = data[data['行业名称']=='羽绒服']
x1 = x['卖家数占比'].apply(lambda x:x.split('%')[0]);
x1 = x1.astype('float');
month = x['年份'].map(str)+'年'+x['月份'].map(str);
plt.plot(month,x1);
plt.xlabel('月份', fontsize=15);
plt.ylabel('卖家数占比', fontsize=15);
plt.title('羽绒服');
plt.tick_params(axis='both', labelsize=12, color='red');
plt.show();

```

