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

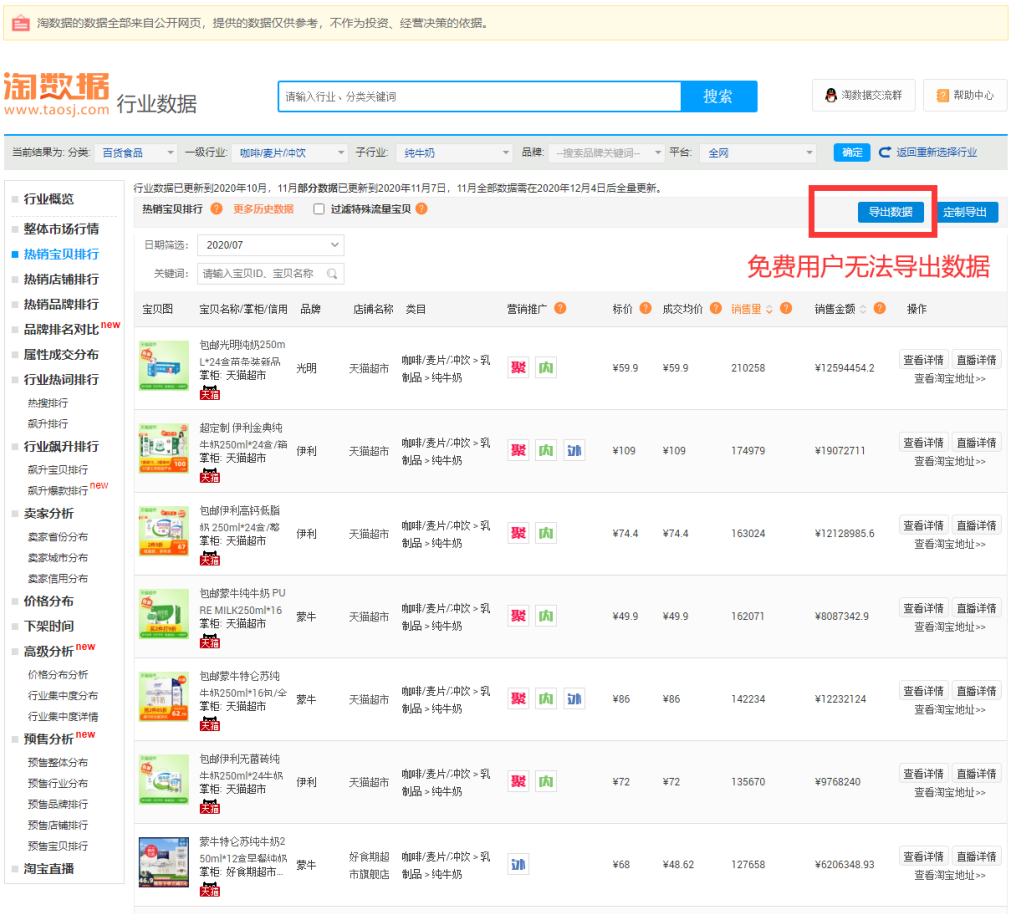
1、Selenium模块可以自动提交信息、自动跳转页面。

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。

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。

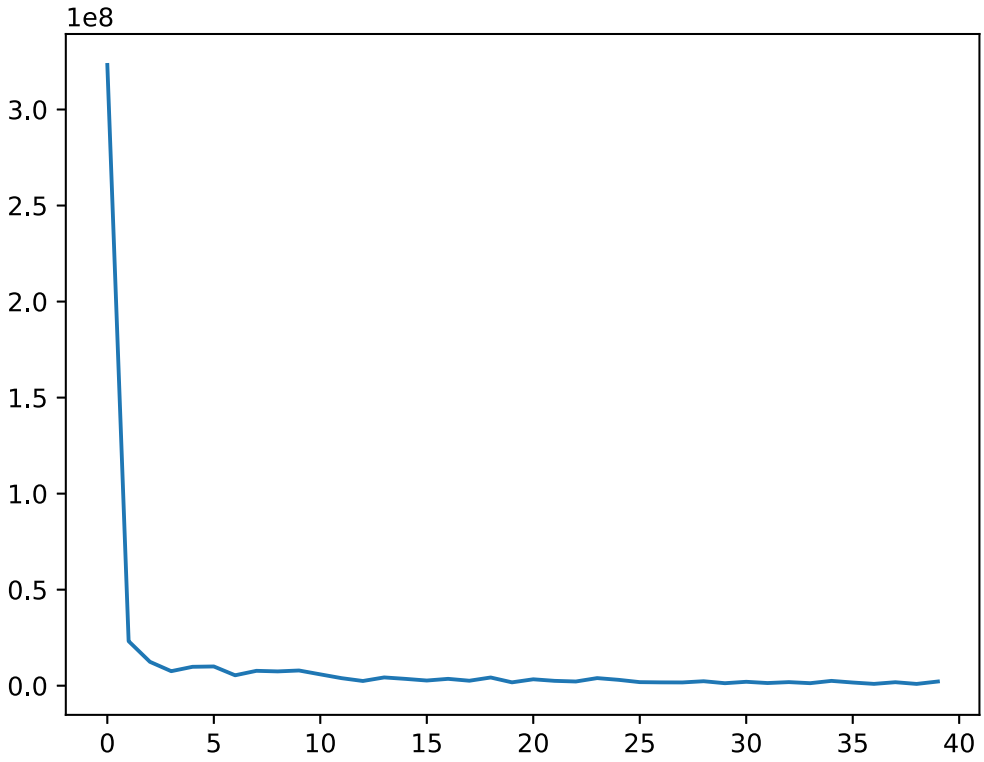


2、数据的清洗，选择"冲饮制品"。

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。

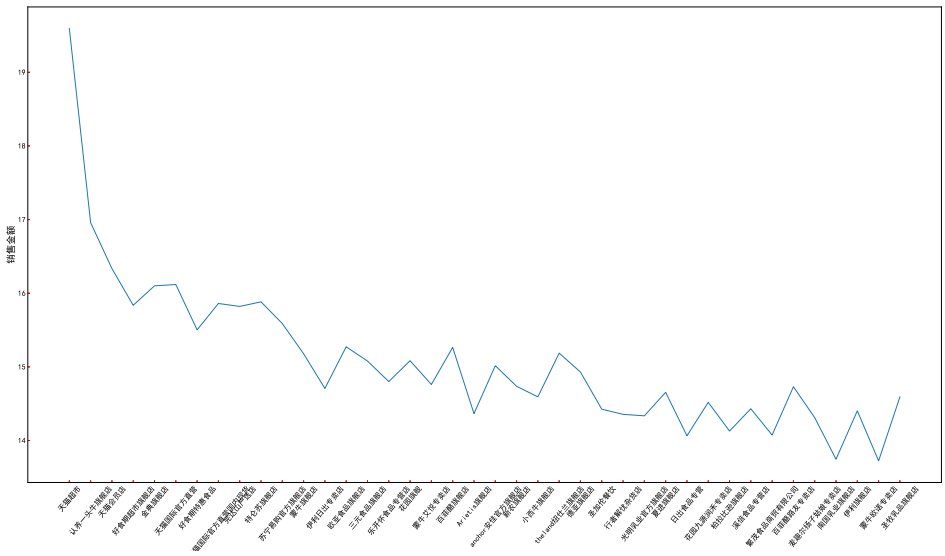
2、读取淘数据下载 冲饮制品 的商品，通过sublime text3将复制的数据拷贝到bao.csv文件，数据要保存为UTF-8的编码。

```
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();
```

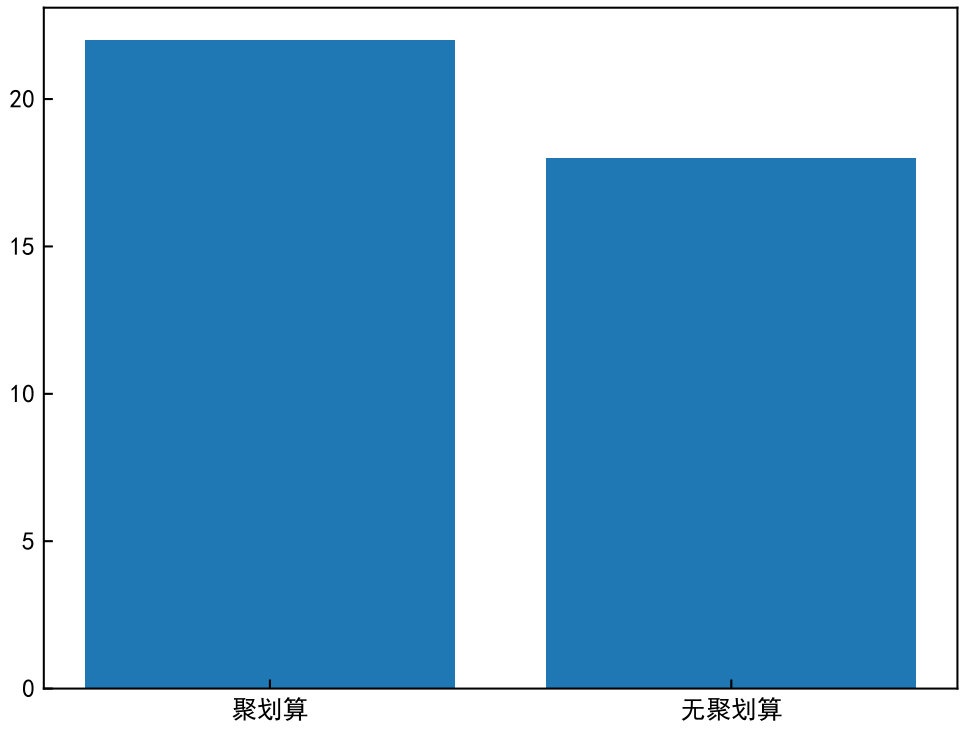
3、销量第一名的商家的销量太高，可以使用$\ln x$函数对数据进行对数变换。



3、二值化处理

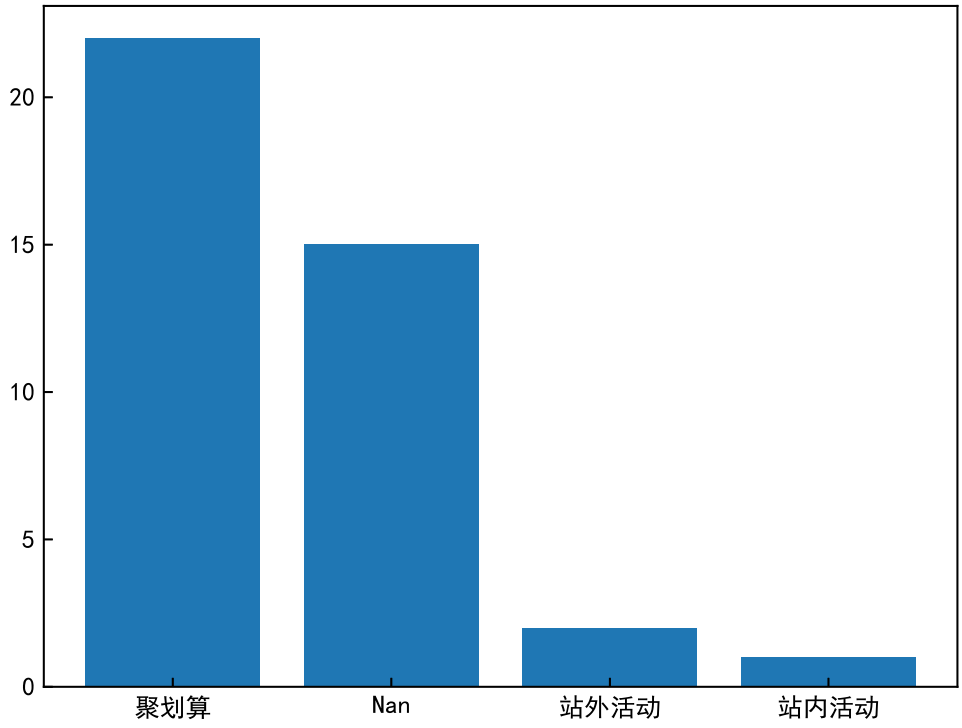
1、根据 热销店铺 的“营销推广”方式，主要有"聚划算"、"站内活动"、"站外活动"三种营销推广方式。

```
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','无聚划算'],[22,18]);
plt.show();
```



2、区间量化（分箱）

```
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()
```

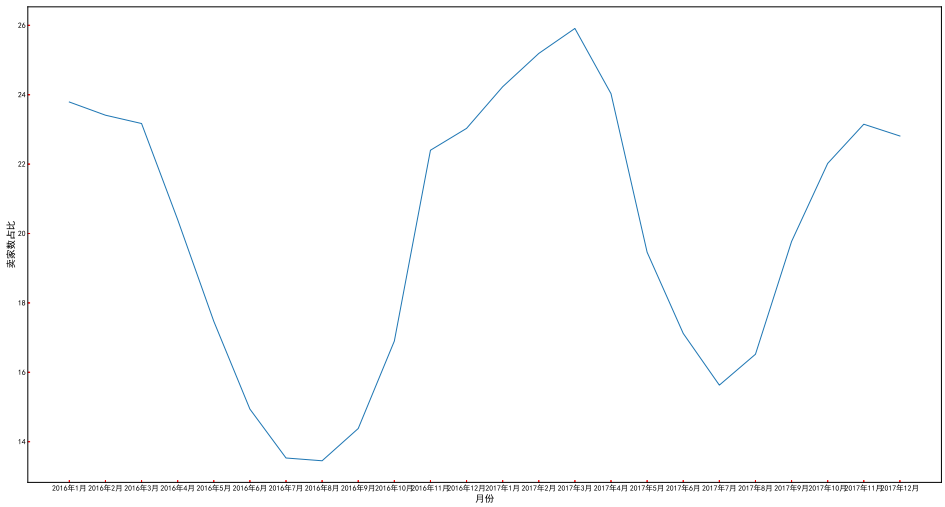
4、通过淘数据下载数据

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();
```