**大连理工大学开源大作业**

**京东商城计算机销售页面数据获取与分析**

学 院（系）： 软件学院

专 业： 软件开发与测试

学 生 姓 名： 陈钦禹 201992030

刘永凯 201992121

陈嗣元 201995036

指 导 教 师： 任志磊

评 阅 教 师：

完 成 日 期： 2022/1/1

大连理工大学

Dalian University of Technology

# 摘 要

随着时代的进步，计算机逐渐成为人们日常生活工作中不可或缺的重要生产工具，同时品种繁多的电脑品牌也快速占领了市场，而这，也使得用户在购买计算机时面临这较大的困难，因为，为了辅助用户做出选择，我们在京东商城上调研了目前市场上市场份额占比较大的电脑品牌的平均价格以及具体市场份额占比，使用户可以更为轻松的做出选择。本次开发使用的是Python语言，首先从京东首页上搜索计算机，然后爬取页面上各商店售卖的电脑的价格及品牌，将爬取到的数据制作为饼状图以及条形图，以直观的得到市场占比和平均价格的具体信息，经过调查分析，目前京东商城上，市场占比份额最高的主流电脑品牌是Lenovo（联想），高达25.50%，而平均价格最高，最低的分别为HUAWEI（华为）和ASUS（华硕）。综合来看，如果您预算较为优先，更为推荐购买华硕（ASUS）和惠普（hp）两个品牌；而如果您预算较为充足，更为推荐虽然价格较高，但体验更好的华为（HUAWEI）和戴尔（DELL）；另外您更为相信大众的眼光，市场份额最大的联想（Lenovo）计算机，也是极佳的选择。

**关键词：市场份额；平均价格；计算机**

目 录

[摘 要 II](#_Toc92203025)

[1. 项目系统架构 1](#_Toc92203026)

[2. 数据获取 2](#_Toc92203027)

[2.1 请求得到目标网站的数据 2](#_Toc92203028)

[2.2 从解析器中截取需要的信息 4](#_Toc92203029)

[2.3 数据的分析 5](#_Toc92203030)

[3. 数据存储 5](#_Toc92203031)

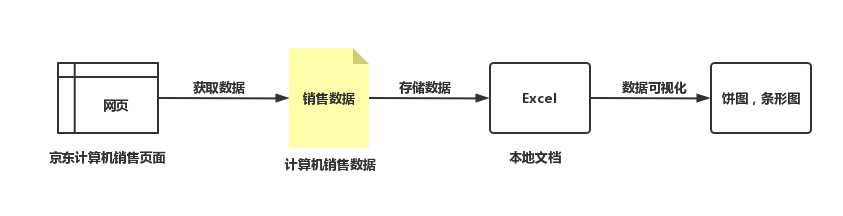
[4. 数据可视化及分析 7](#_Toc92203032)

[5. 词云生成 12](#_Toc92203033)

[6. 组内分工 14](#_Toc92203034)

[结 论 15](#_Toc92203035)

# 1. 项目系统架构



如上图，本项目的数据来自京东计算机销售页面，该网站包含众多店铺销售电脑的详细信息，本项目通过Python爬虫框架从该网页上获取数据，框架主要分为主要分为调度器、URL管理器、网页下载器、网页解析器、应用程序（爬取的有价值数据）。调度器主要来调度管理器、下载器和解析器；URL管理器主要用来管理URL，防止重复抓取或者循环抓取等；网页下载器用于下载网页，并转换成字符串；网页解析器用于解析下载下来的字符串。之后将获取到的数据存储为一个本地Excel文档，为了能够只管的分析数据，将数据进行可视化处理，此处使用了Python中的matplotlib来绘制2D图形（饼图，条形图）。

# 2. 数据获取

数据获取过程主要采用了Python的爬虫框架，框架的使用使得我们节省了大量工作，

因此该过程中，我们主要关注三个步骤：

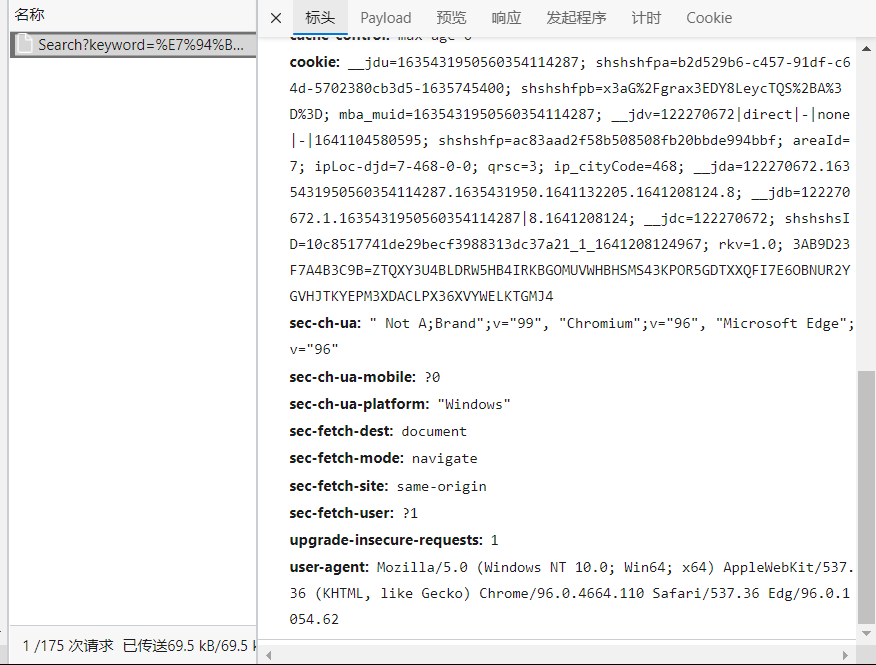
1、如何能请求得到目标网站的数据，

2、如何从解析器中截取我们想要数据，

3、得到数据后如何做分析。

## 2.1 请求得到目标网站的数据

直接访问，会被网站识别出，发送请求的是一个Python程序而不是浏览器，因此会被当做爬虫而无法访问网页，因此需要进行伪装：



核心代码如下

#得到指定一个url的网页信息

def askURL(url):

#用户代理表示不是爬虫，模拟头部信息

head={"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/95.0.4638.54 Safari/537.36"}

request=urllib.request.Request(url,headers=head)

html=""

try:

response=urllib.request.urlopen(request)

html=response.read().decode("utf-8")

#print(html)

#异常处理机制

except urllib.error.URLError as e:

if(hasattr(e,"code")):

print(e,"code")

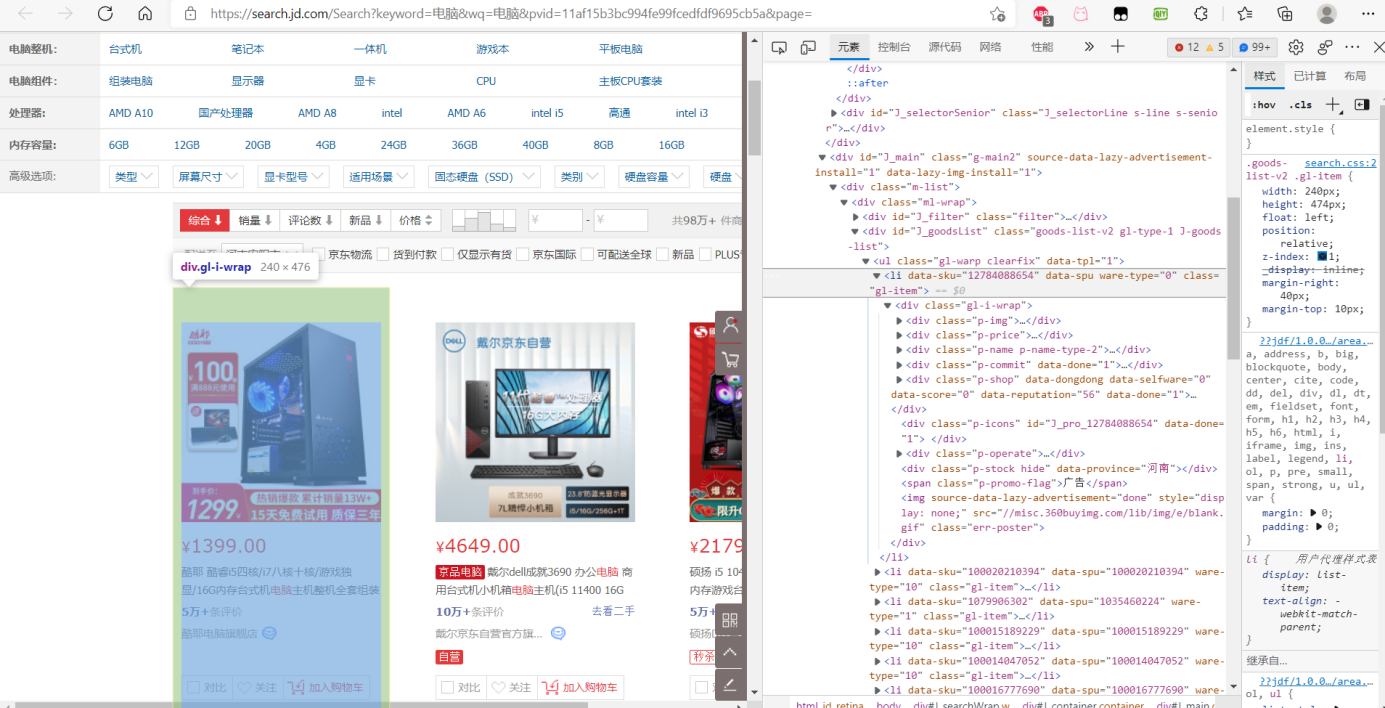
if(hasattr(e,"reason")):

print(e,"reason")

return html

## 2.2 从解析器中截取需要的信息

数据的截取，采用正则提取的方式



打开网页开发者模式，可以看到，我们需要的信息包含在一个开头为div.gl-i-wrap的内容之中，因为设置过滤规则，筛选所有符合规则的字符

具体代码如下：

soup =BeautifulSoup(html,"html.parser")

for item in soup.select("div.gl-i-wrap"):

data = []#保存一部电脑的全部信息

item = str(item)

........

然后通过正则表达式来查找我们所需要的内容，经过对内容的分析，我们所需内容的正则表达式为：

# 电脑信息查找规则

findName = re.compile(r'<em>(.\*?)<font class="skcolor\_ljg">') #生成正则表达式的对象，表示规则

price = re.compile(r'<i data-price="(.\*?)</i>')

shop = re.compile(r'<a class="curr-shop hd-shopname"(.\*?)</a>')

web = re.compile(r'<a href="(.\*?)" onclick="')

通过以上方式，可以从所有信息中，筛选出所有我们所需要的信息，从而完成有效信息的截取。

## 2.3 数据的分析

数据的分析会在数据的可视化阶段进行，用户在购买计算机时，一般最关心的是，计算机的价格和市场的分布情况（即大多数人购买的计算机时那种品牌的），这部分内容会在可视化阶段详细展开，在此不在过多赘述。

# 数据存储

数据的存储，在获取数据后，为方便进一步的分析处理，考虑将获取到的数据按照一定的规则存储在本地的文档中，文档分为4项内容：

-电脑标签：用于详细记录，计算机的具体品牌及参数

-电脑价格：用于记录获取数据时，计算机的价格

-网址：用于记录计算机销售店铺的链接

-品牌：用于记录计算的品牌，便于分析各公司的市场份额占比

考虑到数据统计的规范性以及分析的简便性，将数据整理为Excel表格，在保证所有信息得到保留的同时，方便下一步的分析统计；

具体代码如下：

def saveData(datalist,savepath):

book = xlwt.Workbook(encoding="utf-8")#utf-8编码方式

sheet = book.add\_sheet('information')#命名Excel文件sheet名字为information

col = ("电脑标签","电脑价格","网址","品牌")

for i in range(0,4):

sheet.write(0,i,col[i])#列名的写入

for i in range(0,1000):#

print("第%d条"%i)#将抓取的前1000条数据存入Excel中

data = datalist[i]#按行提取数据

for j in range(0,4):

sheet.write(i+1,j,data[j])#写入数据

book.save(savepath)#保存Excel路径

生成的Excel文档的局部截图如下：



# 数据可视化及分析

为了是数据分析结果能够更为直观的展现在用户面前，决定将数据整理为较为易于理解的饼状图和条形图：

* 1. 饼状图的绘制

饼状图用于展示京东商城上主要计算机品牌的市场占比份额，该数据根据京东商城上，各品牌的店铺开设数量了间接推导分析，根据开设某品牌计算机的店铺的数量与总店铺数量之比来表示该品牌计算机在市场上的占比，由于在数据存储阶段选择了将数据存储为Excel表格，因此，首先从Excel表格中读取出经过整理的数据，然后统计各种品牌的店铺数量，以此得出大致的市场份额占比：

具体代码如下：

data = xlrd.open\_workbook('computer.xls')#读取Excel文件

table = data.sheets()[0]#打开information名字的sheet

num1=0;num2=0;num3=0;num4=0;num5=0;num6=0

price1=0;price2=0;price3=0;price4=0;price5=0;price6=0

cols\_value = table.col\_values(3)#提取品牌(第3列) 列表形式 包括所有品牌

cols\_value\_price = table.col\_values(1)#提取价格(第1列)

for i in range(1,1001):#读取这1000个数据

#检测到对应品牌--对应数量加1 并且 对应总价钱加上搜索到的价钱

if(cols\_value[i]=="联想"):

num1=num1+1

price1=price1+float(cols\_value\_price[i])

elif(cols\_value[i]=="华为"):

num2=num2+1

price2 = price2 + float(cols\_value\_price[i])

elif (cols\_value[i] == "惠普"):

num3 = num3 + 1

price3 = price3 + float(cols\_value\_price[i])

elif (cols\_value[i] == "戴尔"):

num4 = num4 + 1

price4 = price4 + float(cols\_value\_price[i])

elif(cols\_value[i]=="华硕"):

num5=num5+1

price5 = price5 + float(cols\_value\_price[i])

else:

num6=num6+1

price6 = price6 + float(cols\_value\_price[i])

饼状图的绘制采用了Python的matplotlib库，

#关于市场份额的扇形图

X=[num1,num2,num3,num4,num5,num6]#数据成员

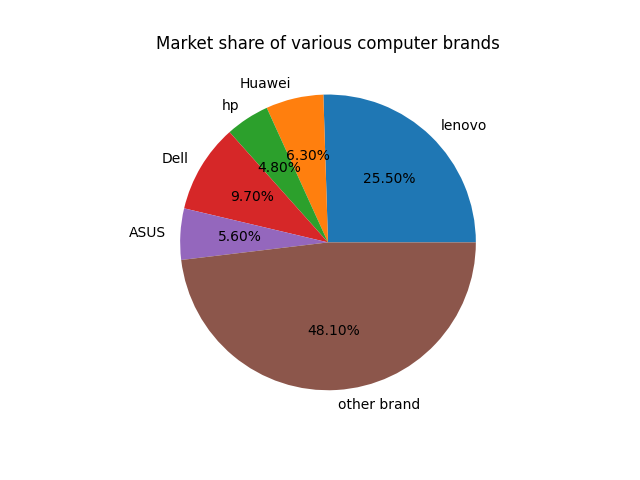
fig = plt.figure()

plt.pie(X, labels=labels, autopct='%1.2f%%') # 画饼图（数据，数据对应的标签，百分数保留两位小数点）

plt.title("Market share of various computer brands") #设置图片标题

plt.show()

最后得到的图形如下：



* 1. 条形图的绘制：

与饼状图类似，首先从Excel表格中提取需要的数据，计算每种品牌的平均价格，具体代码如下：

data = xlrd.open\_workbook('computer.xls')#读取Excel文件

table = data.sheets()[0]#打开information名字的sheet

num1=0;num2=0;num3=0;num4=0;num5=0;num6=0

price1=0;price2=0;price3=0;price4=0;price5=0;price6=0

cols\_value = table.col\_values(3)#提取品牌(第3列) 列表形式 包括所有品牌

cols\_value\_price = table.col\_values(1)#提取价格(第1列)

for i in range(1,1001):#读取这1000个数据

#检测到对应品牌--对应数量加1 并且 对应总价钱加上搜索到的价钱

if(cols\_value[i]=="联想"):

num1=num1+1

price1=price1+float(cols\_value\_price[i])

elif(cols\_value[i]=="华为"):

num2=num2+1

price2 = price2 + float(cols\_value\_price[i])

elif (cols\_value[i] == "惠普"):

num3 = num3 + 1

price3 = price3 + float(cols\_value\_price[i])

elif (cols\_value[i] == "戴尔"):

num4 = num4 + 1

price4 = price4 + float(cols\_value\_price[i])

elif(cols\_value[i]=="华硕"):

num5=num5+1

price5 = price5 + float(cols\_value\_price[i])

else:

num6=num6+1

price6 = price6 + float(cols\_value\_price[i])

print(num1,num2,num3,num4,num5,num6,num1+num2+num3+num4+num5+num6)

avg\_price1=price1/num1

avg\_price2=price2/num2

avg\_price3=price3/num3

avg\_price4=price4/num4

avg\_price5=price5/num5

avg\_price6=price6/num6

labels=['lenovo','Huawei','hp','Dell','ASUS','other brand']

条形图的绘制同样使用matplotlib库，代码如下：

#画关于平均价格的条形图

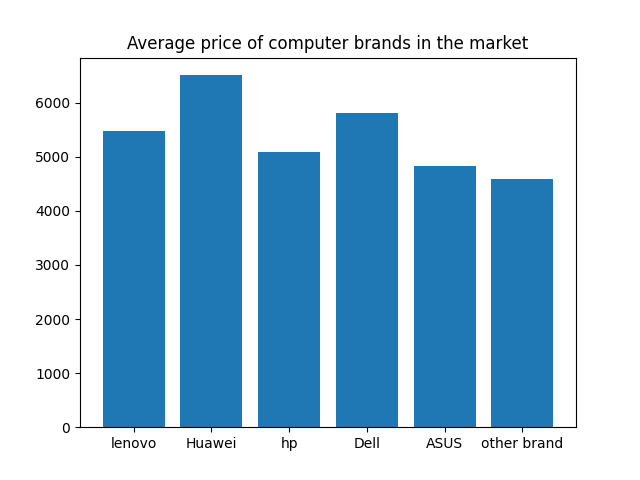
avg\_ = [avg\_price1,avg\_price2,avg\_price3,avg\_price4,avg\_price5,avg\_price6]

plt.bar(labels, avg\_)

plt.title('Average price of computer brands in the market')

plt.show()

绘制得到的条形图如下：



# 词云生成

此部分内容是在完成基础功能后，希望能够进一步分析京东商城上计算机品牌的特点所产生的，词云分析提取了京东商城计算机销售页面上个店铺销售计算机的关键形容词，由此来间接推断，用户购买计算机更希望选择何种款式，以及购买后，计算机的主要使用场景：

词云的生成与售价与份额的分析类似，首先从Excel文档中读取之前获取并整理保存后的信息，通过wordcloud库对数据进行分析，然后仍然使用matplotlib绘制2D图形，

具体代码如下：

import xlrd

from PIL import Image

from wordcloud import WordCloud, ImageColorGenerator

import matplotlib.pyplot as plt

import numpy as np

import jieba

def ToTxt():

# 将title导入到文本中

data = xlrd.open\_workbook('computer.xls') # 读取Excel文件

table = data.sheets()[0] # 打开information名字的sheet

text = table.col\_values(0)

print(text)

doc = open('computer.txt', 'w+', encoding='UTF-8')

# for i in range(1,1001):

print(text, file=doc)

doc.close()

def GetWordCloud():

path\_txt = 'computer.txt'

path\_img = "heart.png"

f = open(path\_txt, 'r', encoding='UTF-8').read()

background\_image = np.array(Image.open(path\_img))

# 结巴分词，生成字符串，如果不通过分词，无法直接生成正确的中文词云,感兴趣的朋友可以去查一下，有多种分词模式

#Python join()方法用于将序列中的元素以指定的字符连接生成一个新的字符串。

cut\_text = " ".join(jieba.cut(f))

wordcloud = WordCloud(

# 设置字体，不然会出现口字乱码，文字的路径是电脑的字体一般路径，可以换成别的

font\_path="C:/Windows/Fonts/simfang.ttf",

background\_color="white",

mask=background\_image).generate(cut\_text)

# 生成颜色值

image\_colors = ImageColorGenerator(background\_image)

# 下面代码表示显示图片

plt.imshow(wordcloud.recolor(color\_func=image\_colors), interpolation="bilinear")

plt.axis("off")

plt.show()

if \_\_name\_\_ == '\_\_main\_\_':

ToTxt()

GetWordCloud()

代码运行后，结果如下：



从上图可以分析得出，计算机的主要应用场景是商务和办公，店铺一般将电脑屏幕的大小作为主要卖点，计算机销售类型主要为笔记本电脑。

# 6. 组内分工

组长：陈钦禹

实现python爬虫，并将爬取到的数据存入到Excel中

组员：刘永凯

读取Excel中的数据，生存对应的价格条形图和数量饼状图，并完成报告的书写

组员：陈嗣元

使用python代码生成关于电脑标签的词云

# 结 论

经过调查分析，在京东商城计算机销售页面上，联想（Lenovo）品牌的店铺数量最多，占总店铺数量的25.50%，相比第二位的戴尔（Dell）品牌的9.70%，高出15.8%，为目前京东商城计算机市场中，份额占比最大的品牌；

在对于平均价格的分析中，华为（HUAWEI）品牌的平均价格最高，高达6400元，但考虑到华为品牌的质量一般令人满意，希望获得较好体验且预算充足的用户可以考虑购买；而市面主要品牌中，平均价格最低的是惠普（hp）品牌，以仅为不到5000的价格在低价区抢的一席之地，如果预算优先，可以优先考虑惠普品牌的计算机；

在对于店铺主要标签的分析中，英寸两个打字首先映入眼帘，说明，店铺在销售计算机时，屏幕尺寸一直是主要卖点；之后较为显眼的词语是商务和办公，这在一定程度上反应出，用户购买计算机的主要用途是用于商务办公。

以上即为此次项目对京东商城计算机销售页面做出的分析。希望能够为希望购买计算机的用户提供一定的参考。