古风商品的探索

CPDA 数据分析师[上海]　杨冰羽

十八飞骑，群雄授首
有公子摇扇轻笑
灵犀一指，花香满楼
你有鲜衣怒马，
我有蔽履寒裘
且拔刀来，看谁光寒三百州
你是庙堂衙内，
我是春江钓叟
袒腹而歌，一韵诗，一杯酒

江湖给少时的我们许了一场梦。我在梦外，看着梦里的人们清酒高歌，快意恩仇。不管是谁都会对江湖有一颗向往心吧，曾梦想仗剑走天涯，看一下世间的繁华。

说起江湖，在脑海中浮现的不仅是恩怨情仇的情节或者踏雪无痕的身影，还有他们那一层透露江湖意蕴的古风装束。

收起心绪，说说我们这次研究的主题：**淘宝下古风系列的商品探索。**

本次的文章研究方向：

1、创建数据库写入爬虫数据供后续分析
2、利用爬虫技术获取相应的数据，并进行试调优化
3、利用文本分析淘宝里的古风商品
4、古风商品价格，销售量，评论量等数据回归和可视化分析

实施步骤：

1、mysql 创建数据库并建立对应的表，为后续存储数据和分析提供前提
2、数据采集：Python 爬虫淘宝网的古风商品数据，并进行调错，反反爬
3、为分析清洗和处理所需的数据

4、利用文本分析技术：jieba 分词以及 wordcloud 可视化对高词频商品进行分析
5、古风商品销量，价格等数据的回归以及可视化分析
6、商品区域分布展示
7、总结以及指出不足之处

数据获取：

数据来源：淘宝网(www.taobao.com)
关键词搜索：古风
数据集：4303 个（爬取了 105 页左右的数据，仅爬取淘宝网站，对天猫店铺直接忽略）
工具以及相关库：python（urllib.request，re， pymysql，jieba，wordcloud，matplotlab，basemap 等）

数据集的定义：

字段	描述	应用
title	商品名称	商品的标识
itemloc	地址	店铺地址标识
nick	店铺	店铺名称标识
link	商品链接	产品的链接
detail	商品描述	商品详细描述
tagprice	吊牌价	网上原始价格
price	实售价	实际销售价格
mouthcount	月销量	截止爬虫当日销量
totalcount	总销量	店铺总体销量
comment	评论数	累计评论数目

一、创建数据库：

1.1 连接数据库之前的准备

1. 下载 mysql workbench;
2. 找到 anaconda\Lib\site-packages\pymysql\connections.py
   crt+F 搜索 charset='' 在''中加入 utf8;
3. mysql 创建数据和表详细代码如下;

```
'''[1.1]连接数据库之前的准备
>>>1.下载 mysql workbench;
>>>2.找到anaconda\Lib\site-packages\pymysql\connections.py
     crt+F 搜索charset=''  在''中加入utf8;
>>>3.mysql创建数据和表详细代码如下;
create database taobaoitem;
use taobaoitem;
create table zhiniaoku_goods(
id int(32) auto_increment primary key,
titles varchar(250),
nick  varchar(250),
itemloc varchar(250),
link varchar(250),
detail varchar(250),
tagprice varchar(100),
price varchar(100),
mounthcount varchar(100),
totalcount varchar(100),
comments varchar(250)
) character set = utf8;# 中文读入就不会出错了
```

1.2 可能会遇到的问题：

1、字节长度设置少了（解决方式：varchar(250)）
2、中文字段写入问题（解决方式： character set = utf8）
3、特殊符号的词处理（解决方式：replace 将特殊符号替换为空白）

二、撰写爬虫：

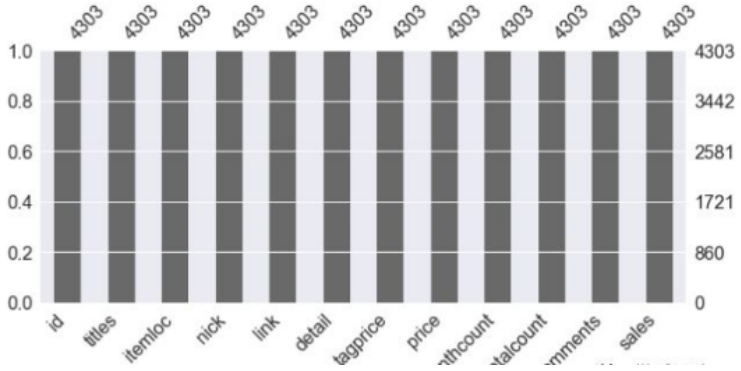
流程：（仅选取部分代码）

1、用户代理池，ip 代理池的建立（池越多越好，越深越赞，本次仅采用了 3 个 ip 切换）
2、设定要抓取的目标，对抓取的网站进行分析，
3、构建一级页面的抓取并获得详情页面的 Link 以及详情页面的对应信息
4、进行抓包分析（可以利用 fiddle 工具实施）
5、循环抓取，对程序错误进行试调，有错误的地方用 try-except 进行处理

三、数据处理：

1、连接数据库读入数据
2、处理缺失值，重复值
3、为后续数据分析对数据进行所需处理

数据缺失值图表（本次案例没有缺失值）：



四、jieba词频处理及词云展示：

4.1、对标题文本进行分词：

词云可视化：

安装模块 wordcloud 直接 pip install wordcloud 经常会遇到错误

所以选择 https://www.lfd.uci.edu/~gohlke/pythonlibs/#wordcloud，根据 python 版本下载对应的库

安装 cd D:pip install wordcloud-1.4.1-cp36-cp36m-win_amd64.whl

4.2、高频词云结果展示：

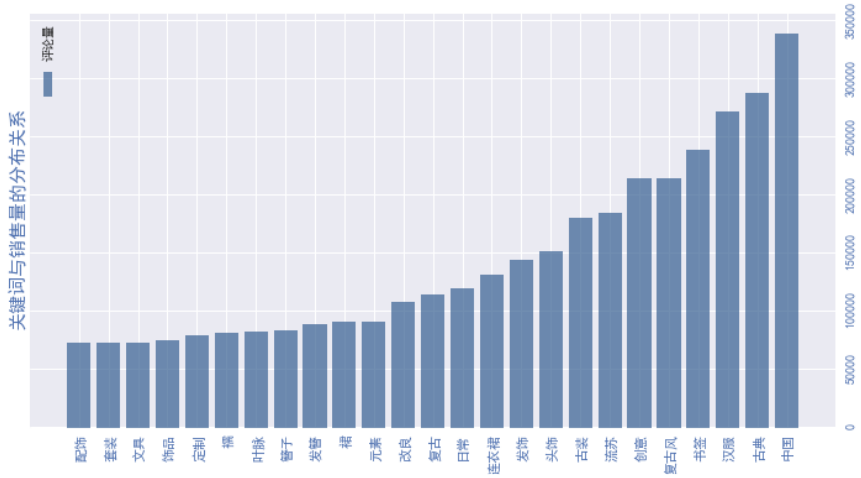
分析结论：

从商品上面来看：连衣裙，汉服，发饰，头饰，流苏等配饰比较多

从风格，元素来看：复古风，古典，淡雅，中国风

五、商品销量,关键词的关系探索

5.1、排名前 25 的关键词与总销量量的分布关系：

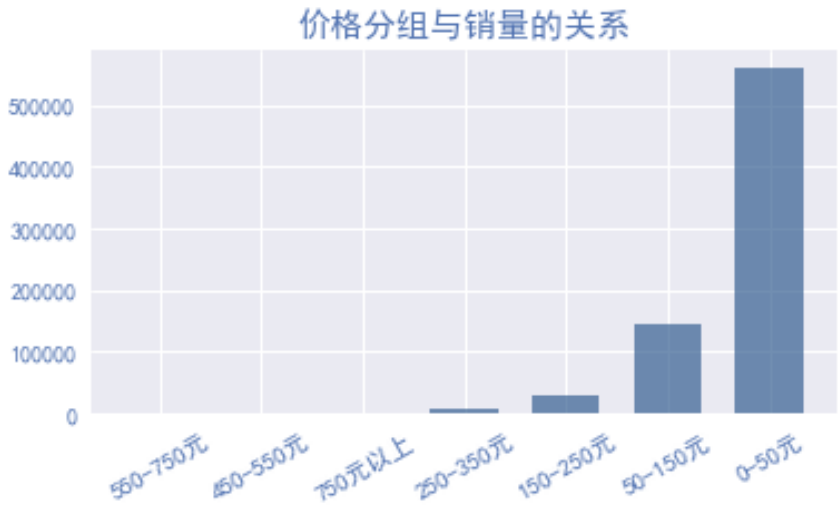


从图表上可知：

单个关键词来看：汉服（中国，古典不算商品）销量最高

配饰总计更高一点：比如发饰，流苏，步摇（原谅小编是个男孩子，第一次知道流苏，步摇是下面这东西？！涨知识了）

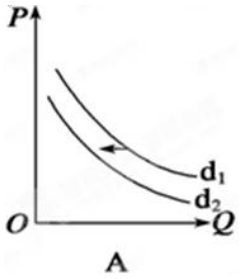
5.2、商品的价格与销售量的分布情况：



从图表上可知：

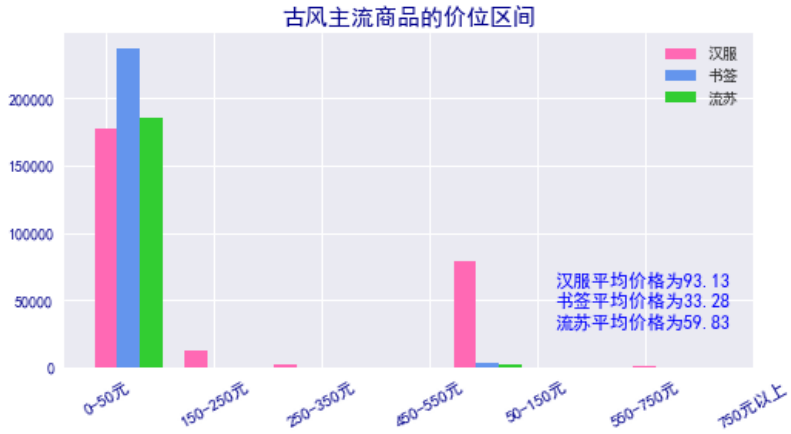
价格区间在 0-50 元的销量更受大家青睐，这些应该是饰品之类的小件，其次为 50-250 元这个价位，这部分大多数应该衣服，裙子，在 350 元以上的评价量就很少了。同样也能看出看出古风系列的关键词消费的客单价主要是应该在 0-350 元之间。

某种程度上实证了微观经济学中的概念：在同等效用下，价格和销量(评论量)是凹型曲线



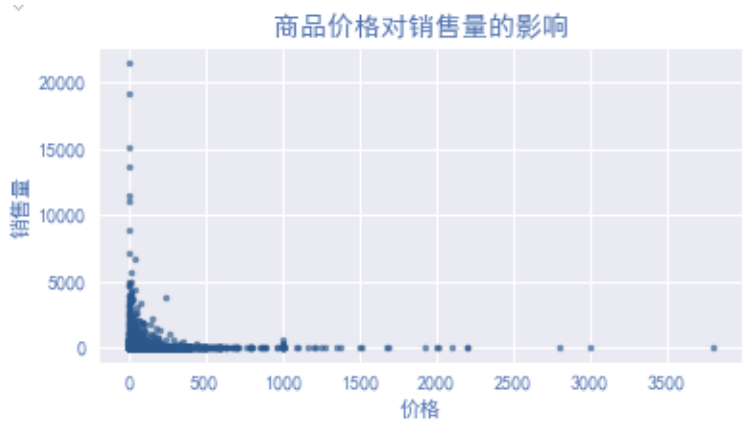
A

5.3、top3 古风商品的价格区间与销售量分布情况：



从上图可知：
0-50 元区间：书签销售占比更高，
50-250 元区间：汉服销售占比更高，和上面 0-50 元更多的是配饰的结论相符合。

5.4 价格与销售量的关系探索

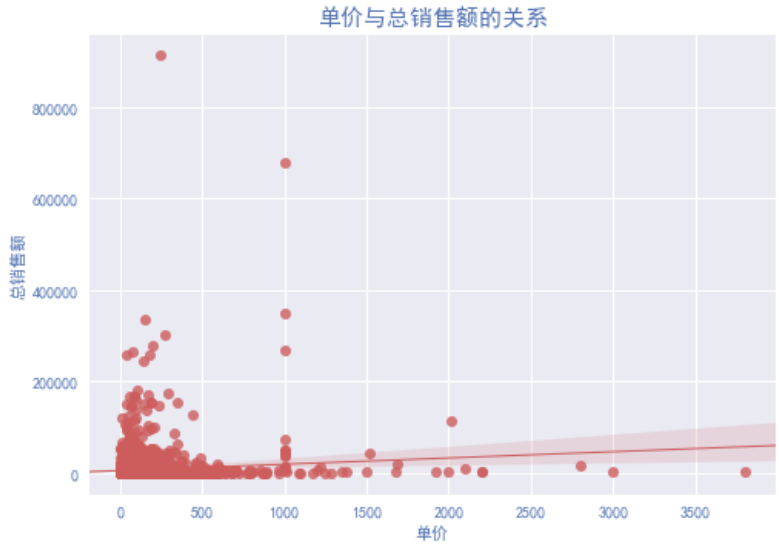


注：data 里面有几个离群值，价格大于 4000 元的为了方便观测将其去除

从图表上可知：

1. 商品数量随着价格总体呈现下降阶梯形势，价格越高，商品的销售就越少；也满足上面所说的在同等效用下，价格和销量是凹型曲线。
2. 低价位商品居多，价格在 0-250 元之间的销量最多，250-500 元之间的次之，价格 500 元以上的商品销售就相对较少；
3. 价格 500 元以上的商品，在售商品数量差异不大。

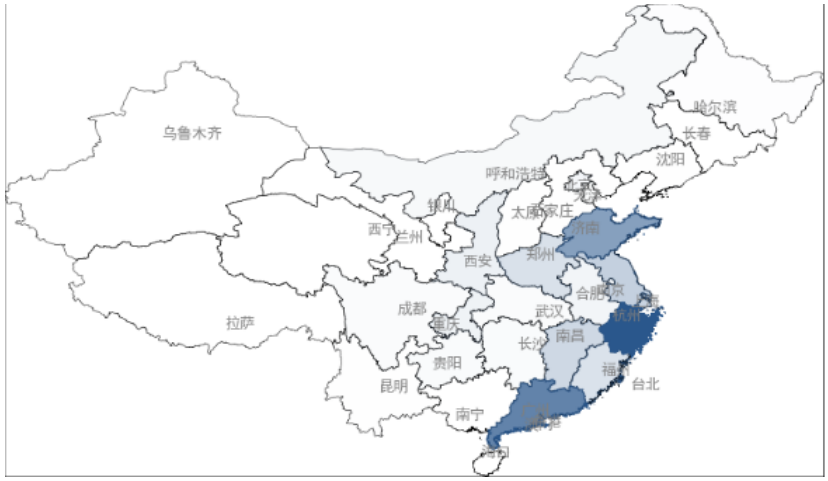
5.5、商品总金额与单价分布情况：：



从图表上可知：

1. 总体趋势：由线性回归拟合线，从总体可以看出，商品总金额与价格关系不大；
2. 多数商品的价格偏低，但是总金额很高，开句玩笑，薄利多销吗？
3. 价格在 0-250 元的商品总金额较高，价格 250-500 元的商品总金额最高，在 0-500 元区间价格越高收入也随之相对越高，有一种上升趋势。

六、不同省份的商品地图分布：



由地图可知：

各商品销量的主要区域主要分布浙江、广东、山东，其次为江苏，上海。怪不得江浙沪包邮，难道是因为商家都在这些地方，距离比较近？

总结：

在淘宝里的古风商品，服装、佩饰价格都普遍处于相对便宜的位置，如：佩饰价格普遍分布在 0-50 元，服装普遍分布在 50-150 元，这都是我们可以接受的范围。所以在生活中，买件古风装，圆个江湖梦的成本还是不算太高。^_^

不足：

1、抓取的速度方面可以优化，反爬方面可以准备更多尝试，因为在抓取到 100 多页后还是意料之内的被 forbid,后续可以考虑建立更多的 ip,和模拟账号登录。

2、仅对普通的淘宝店铺进行了抓取，未对天猫商家的数据进行抓取，后续技术提高后可以尝试增加天猫这一块的样本，毕竟常识来看同款商品天猫的价位比普通的淘宝店铺偏高，并且对于商品有最低和最高价的商品仅选择了最低价，综合来说，总体的价格结论可能会稍微偏低。

3、将关键词作为商品的分类，可能会有类型重叠部分，后续可以通过搜索类别+关键词作为依据。