使用selenium爬取天猫近视眼镜片评论和销售情况并分析

内容简介：先进行数据采集，获取商品数据后再根据商品url拼接评论内容url，通过评论url获得评论数据，最后进行数据分析

1. 数据收集

采用selenium

得到的商品url

https://list.tmall.com/search_product.htm?q=%BD%FC%CA%D3%D1%DB%BE%B5%C6%AC&type=p&spm=a220m.1000858.a2227oh.d100&from=.list.pc_1_searchbutton (https://list.tmall.com/search_product.htm?q=%BD%FC%CA%D3%D1%DB%BE%B5%C6%AC&type=p&spm=a220m.1000858.a2227oh.d100&from=.list.pc_1_searchbutton)

换页条

点击第三页后的商品url

https://list.tmall.com/search_product.htm?spm=a220m.1000858.0.0.3888589de9xg9o&s=120&q=%BD%FC%CA%D3%D1%DB%BE%B5%C6%AC&sort=s&style=g&from=mallfp..pc_1_searchbutton&type=pc#J_Filter (https://list.tmall.com/search_product.htm?spm=a220m.1000858.0.0.3888589de9xg9o&s=120&q=%BD%FC%CA%D3%D1%DB%BE%B5%C6%AC&sort=s&style=g&from=mallfp..pc_1_searchbutton&type=pc#J_Filter)

分析：相比第一页多了一个s参数，s=120，接着跳转到80页，得到的url中参数s=4740
所以可以通过s参数换页获取到不同商品信息，s=(页数 - 1)×60

```
for page in range(1, 78):
    s = (page - 1)*60
    search_url = "https://list.tmall.com/search_product.htm" + "?
cat=50041298&s=" + str(s)\
                 +
"&q=%BD%FC%CA%D3%D1%DB%BE%B5%C6%AC&sort=s&style=g&from=mallfp..pc_1_s
earchbutton&active=2" \

"&industryCatId=50041298&spm=a220m.1000858.0.0.2f4e58981G3iTJ&type=pc
#J_Filter "
    cat_spider = CatSpider(url=search_url)
    cat_spider.get_goods_info(page=page)
    print(search_url)
    sleep(6 + randint(1, 4))
```

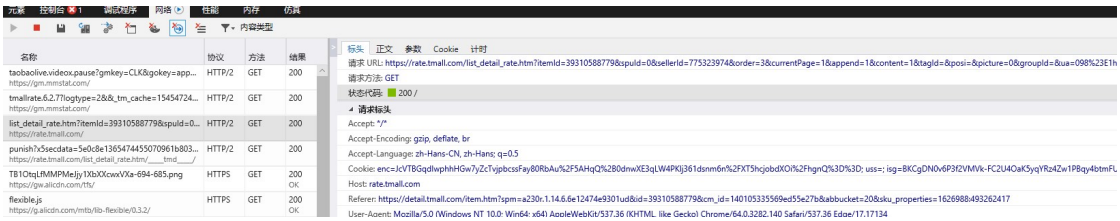
为了防止触发天猫的反爬虫，加了一个随机等待时间

在获取到商品数据后，根据商品地址拼接商品评论url

商品url如下
http://detail.tmall.com/item.htm?id=565334203629&skuId=4063956831708&areaId=430100&user_id=3355869104&cat_id=50041298&is_b=1&rn=1f4a2c261ac23e5987794fcc03e7a9c7 (http://detail.tmall.com/item.htm?id=565334203629&skuId=4063956831708&areaId=430100&user_id=3355869104&cat_id=50041298&is_b=1&rn=1f4a2c261ac23e5987794fcc03e7a9c7)
分析
id：商品id
user_id：卖家id

当点击商品评论的下一页时，商品的url并没有变化，说明评论数据的更新是通过Ajax获取

利用开发者工具查看Ajax请求的地址



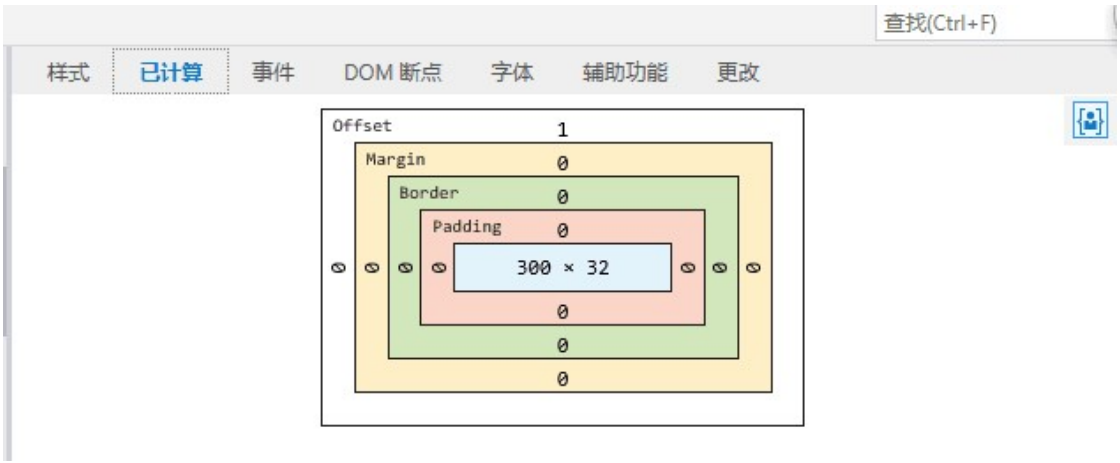
评论请求的url

根据网上的资料分析：
评论数据url https://rate.tmall.com/list_detail_rate.htm?itemId=39310588779&sellerId=775323974&currentPage=1&callback=jsonp1210 (https://rate.tmall.com/list_detail_rate.htm?itemId=39310588779&sellerId=775323974&currentPage=1&callback=jsonp1210)
其中
itemId：商品id 商品url 的 id
sellerId：卖家id 商品url中的user_id
currentPage：页码
callback：
作为回调函数的一部分，返回的是json格式数据，“callback=jsonp”这部分是固定不变的，后面的数字利用random函数生成一个随机数拼接上去

利用parse解析出商品url中的参数

```
url_param = good_info.split('?')
res = parse.parse_qs(url_param[1])
url_str = 'https://rate.tmall.com/list_detail_rate.htm?itemId=' +
''.join(res['id']) + '&sellerId=' \
          + ''.join(res['user_id']) + '&currentPage=1&callback=jsonp'
+ str(randint(1000, 20000))
urls_str.append(url_str)
```

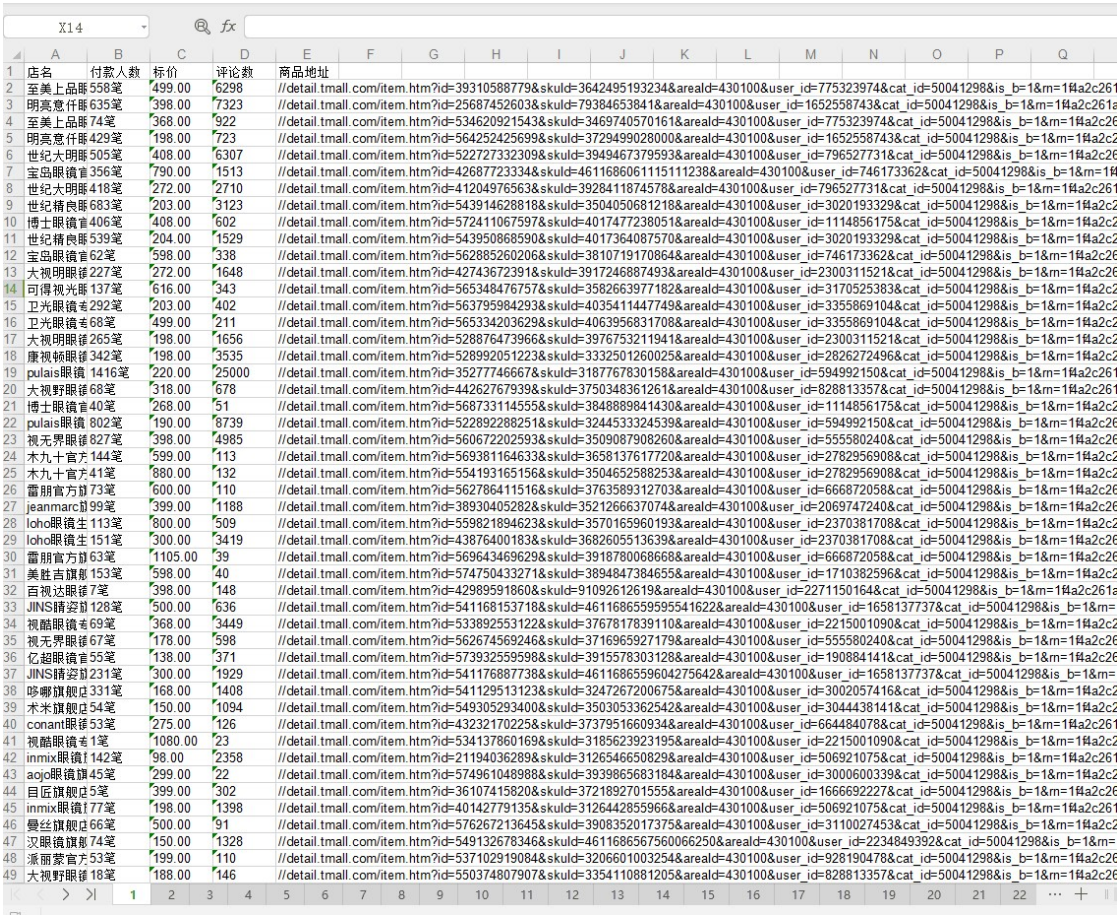
访问100多个页面后，会遇到访问验证，此时服务器返回的json是固定的，包含极验验证的地址，获取地址后，跳转到该地址，获得滑块对象，通过开发者工具可以很容易看到滑动条长度为300px，接着杜撰一个滑动轨迹，（这里我是参考了《python网络爬虫实战》里的模拟验证，但其实没什么用，最后还是自己根据我验证通过的经验，模拟了一个轨迹），开始验证直至验证通过(最开始成功过几次，后来再也没成功过)。



遇到的问题：
多次访问后，（大概访问100多个页面）就会触发阿里的反爬虫，第一次需要登录后，再滑动验证，天猫和淘宝采用的都是第三代极验验证，号称采用机器学习识别轨迹，实测真的很坑，在前几次验证还可以通过，验证多次后感觉难度越来越大，基本不可能通过，就算是人工验证也不鸟你。只有等过几个小时，才会解封。

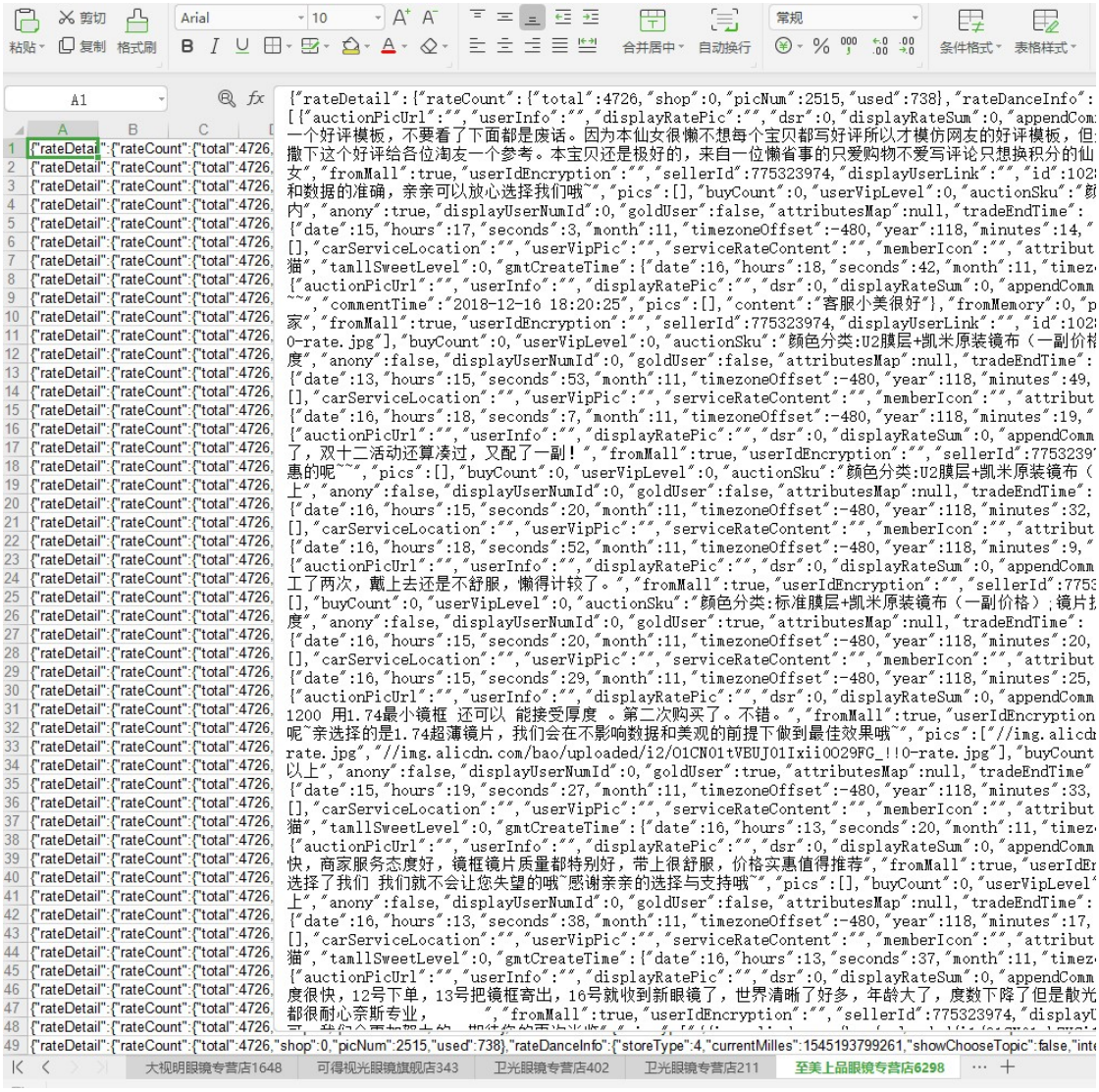
2. 数据处理

爬取到的商品数据



店名	付款人数	标价	评论数	商品地址
至美上品眼	558笔	499.00	6298	//detail.tmall.com/item.htm?id=39310588779&skuId=3642495193234&areaId=430100&user_id=775323974&cat_id=50041298&is_b=1&rn=1f4a2c261
明亮意仟眼	635笔	398.00	7323	//detail.tmall.com/item.htm?id=25687452603&skuId=79384653841&areaId=430100&user_id=1652558743&cat_id=50041298&is_b=1&rn=1f4a2c261a
至美上品眼	74笔	368.00	922	//detail.tmall.com/item.htm?id=534620921543&skuId=3469740570161&areaId=430100&user_id=775323974&cat_id=50041298&is_b=1&rn=1f4a2c26
明亮意仟眼	429笔	198.00	723	//detail.tmall.com/item.htm?id=564252425699&skuId=37294990280000&areaId=430100&user_id=1652558743&cat_id=50041298&is_b=1&rn=1f4a2c2
世纪大明眼	505笔	408.00	6307	//detail.tmall.com/item.htm?id=522727332309&skuId=39494673795938areaId=430100&user_id=796527731&cat_id=50041298&is_b=1&rn=1f4a2c26
宝岛眼镜官	356笔	790.00	1513	//detail.tmall.com/item.htm?id=42687723334&skuId=461168606111511112388areaId=430100&user_id=746173362&cat_id=50041298&is_b=1&rn=1f4
世纪大明眼	418笔	272.00	2710	//detail.tmall.com/item.htm?id=41204976563&skuId=3928411874578&areaId=430100&user_id=796527731&cat_id=50041298&is_b=1&rn=1f4a2c261
世纪精良眼	683笔	203.00	3123	//detail.tmall.com/item.htm?id=543914628818&skuId=3504050681218&areaId=430100&user_id=3020193329&cat_id=50041298&is_b=1&rn=1f4a2c2
博士眼镜官	406笔	408.00	602	//detail.tmall.com/item.htm?id=572411067597&skuId=4017477238051&areaId=430100&user_id=1148561758&cat_id=50041298&is_b=1&rn=1f4a2c2
世纪精良眼	539笔	204.00	1529	//detail.tmall.com/item.htm?id=543950868590&skuId=4017364087570&areaId=430100&user_id=3020193329&cat_id=50041298&is_b=1&rn=1f4a2c2
宝岛眼镜官	62笔	598.00	338	//detail.tmall.com/item.htm?id=562885260206&skuId=3810719170864&areaId=430100&user_id=746173362&cat_id=50041298&is_b=1&rn=1f4a2c26
大视明眼镜	227笔	272.00	1648	//detail.tmall.com/item.htm?id=42743672391&skuId=3917246887493&areaId=430100&user_id=2300311521&cat_id=50041298&is_b=1&rn=1f4a2c26
可得视光眼	137笔	616.00	343	//detail.tmall.com/item.htm?id=565348476757&skuId=3582663977182&areaId=430100&user_id=3170525383&cat_id=50041298&is_b=1&rn=1f4a2c2
卫光眼镜专	292笔	203.00	402	//detail.tmall.com/item.htm?id=563795984293&skuId=4035411447749&areaId=430100&user_id=3355869104&cat_id=50041298&is_b=1&rn=1f4a2c2
卫光眼镜专	68笔	499.00	211	//detail.tmall.com/item.htm?id=565334203629&skuId=4063956831708&areaId=430100&user_id=3355869104&cat_id=50041298&is_b=1&rn=1f4a2c2
大视明眼镜	265笔	198.00	1656	//detail.tmall.com/item.htm?id=528876473966&skuId=3976753211941&areaId=430100&user_id=2300311521&cat_id=50041298&is_b=1&rn=1f4a2c2
康视帧眼镜	342笔	198.00	3535	//detail.tmall.com/item.htm?id=528992051223&skuId=3332501260025&areaId=430100&user_id=2826272496&cat_id=50041298&is_b=1&rn=1f4a2c2
pulais眼镜	1416笔	220.00	25000	//detail.tmall.com/item.htm?id=35277746667&skuId=3187767830158&areaId=430100&user_id=594992150&cat_id=50041298&is_b=1&rn=1f4a2c261
大视野眼镜	68笔	318.00	678	//detail.tmall.com/item.htm?id=442627679398skuId=3750348361261&areaId=430100&user_id=828813357&cat_id=50041298&is_b=1&rn=1f4a2c261
博士眼镜官	40笔	268.00	51	//detail.tmall.com/item.htm?id=568733114555&skuId=3848889841430&areaId=430100&user_id=1114856175&cat_id=50041298&is_b=1&rn=1f4a2c2
pulais眼镜	802笔	190.00	8739	//detail.tmall.com/item.htm?id=522892288251&skuId=3244533324539&areaId=430100&user_id=594992150&cat_id=50041298&is_b=1&rn=1f4a2c26
视无界眼镜	827笔	398.00	4985	//detail.tmall.com/item.htm?id=560672202593&skuId=3509087908260&areaId=430100&user_id=555580240&cat_id=50041298&is_b=1&rn=1f4a2c26
木九十官方	144笔	599.00	113	//detail.tmall.com/item.htm?id=569381164633&skuId=3658137617720&areaId=430100&user_id=2782956908&cat_id=50041298&is_b=1&rn=1f4a2c2
木九十官方	41笔	880.00	132	//detail.tmall.com/item.htm?id=554193165156&skuId=3504652588253&areaId=430100&user_id=2782956908&cat_id=50041298&is_b=1&rn=1f4a2c2
雷朋官方旗	73笔	600.00	110	//detail.tmall.com/item.htm?id=562786411516&skuId=3763589312703&areaId=430100&user_id=666872058&cat_id=50041298&is_b=1&rn=1f4a2c26
jeanmarc旗	99笔	399.00	1188	//detail.tmall.com/item.htm?id=38930405282&skuId=3521266637074&areaId=430100&user_id=2069747240&cat_id=50041298&is_b=1&rn=1f4a2c26
loho眼镜生	113笔	800.00	509	//detail.tmall.com/item.htm?id=559821894623&skuId=3570165960193&areaId=430100&user_id=2370381708&cat_id=50041298&is_b=1&rn=1f4a2c2
loho眼镜生	151笔	300.00	3419	//detail.tmall.com/item.htm?id=43876400183&skuId=3682605513639&areaId=430100&user_id=2370381708&cat_id=50041298&is_b=1&rn=1f4a2c26
雷朋官方旗	63笔	1105.00	39	//detail.tmall.com/item.htm?id=569643469629&skuId=3918780068668&areaId=430100&user_id=666872058&cat_id=50041298&is_b=1&rn=1f4a2c26
美胜吉旗舰	153笔	598.00	40	//detail.tmall.com/item.htm?id=574750433271&skuId=3894847384655&areaId=430100&user_id=1710382596&cat_id=50041298&is_b=1&rn=1f4a2c2
百视达眼镜	7笔	398.00	148	//detail.tmall.com/item.htm?id=42989591860&skuId=91092612619&areaId=430100&user_id=2271150164&cat_id=50041298&is_b=1&rn=1f4a2c261a
JINS睛姿旗	128笔	500.00	636	//detail.tmall.com/item.htm?id=541168153718&skuId=4611686559595541622&areaId=430100&user_id=1658137737&cat_id=50041298&is_b=1&rn=
视酷眼镜专	69笔	368.00	3449	//detail.tmall.com/item.htm?id=533892553122&skuId=3767817839110&areaId=430100&user_id=2215001090&cat_id=50041298&is_b=1&rn=1f4a2c2
视无界眼镜	67笔	178.00	598	//detail.tmall.com/item.htm?id=562674569246&skuId=3716965927179&areaId=430100&user_id=555580240&cat_id=50041298&is_b=1&rn=1f4a2c26
亿超眼镜官	55笔	138.00	371	//detail.tmall.com/item.htm?id=573932559598&skuId=3915578303128&areaId=430100&user_id=190884141&cat_id=50041298&is_b=1&rn=1f4a2c26
JINS睛姿旗	231笔	300.00	1929	//detail.tmall.com/item.htm?id=541176887738&skuId=4611686559604275642&areaId=430100&user_id=1658137737&cat_id=50041298&is_b=1&rn=
哆哪旗舰店	331笔	168.00	1408	//detail.tmall.com/item.htm?id=541129513123&skuId=3247267200675&areaId=430100&user_id=3002057416&cat_id=50041298&is_b=1&rn=1f4a2c2
术米旗舰店	54笔	150.00	1094	//detail.tmall.com/item.htm?id=549305293400&skuId=3503053362542&areaId=430100&user_id=3044438141&cat_id=50041298&is_b=1&rn=1f4a2c2
conant眼镜	53笔	275.00	126	//detail.tmall.com/item.htm?id=43232170225&skuId=3737951660934&areaId=430100&user_id=664484078&cat_id=50041298&is_b=1&rn=1f4a2c261
视酷眼镜专	1笔	1080.00	23	//detail.tmall.com/item.htm?id=534137860169&skuId=3185623923195&areaId=430100&user_id=2215001090&cat_id=50041298&is_b=1&rn=1f4a2c2
inmix眼镜	142笔	98.00	2358	//detail.tmall.com/item.htm?id=21194036289&skuId=3126546650829&areaId=430100&user_id=506921075&cat_id=50041298&is_b=1&rn=1f4a2c261
aojo眼镜旗	45笔	299.00	22	//detail.tmall.com/item.htm?id=574961048988&skuId=3939865683184&areaId=430100&user_id=3000600339&cat_id=50041298&is_b=1&rn=1f4a2c2
目匠旗舰店	5笔	399.00	302	//detail.tmall.com/item.htm?id=36107415820&skuId=3721892701555&areaId=430100&user_id=1666692227&cat_id=50041298&is_b=1&rn=1f4a2c26
inmix眼镜	77笔	198.00	1398	//detail.tmall.com/item.htm?id=401427791358skuId=3126442855966&areaId=430100&user_id=506921075&cat_id=50041298&is_b=1&rn=1f4a2c261
曼丝旗舰店	66笔	500.00	91	//detail.tmall.com/item.htm?id=576267213645&skuId=3908352017375&areaId=430100&user_id=3110027453&cat_id=50041298&is_b=1&rn=1f4a2c2
汉眼镜旗舰	74笔	150.00	1328	//detail.tmall.com/item.htm?id=549132678346&skuId=4611686567560066250&areaId=430100&user_id=2234849392&cat_id=50041298&is_b=1&rn=
派丽蒙官方	53笔	199.00	110	//detail.tmall.com/item.htm?id=537102919084&skuId=3206601003254&areaId=430100&user_id=928190478&cat_id=50041298&is_b=1&rn=1f4a2c26
大视野眼镜	18笔	188.00	146	//detail.tmall.com/item.htm?id=550374807907&skuId=3354110881205&areaId=430100&user_id=828813357&cat_id=50041298&is_b=1&rn=1f4a2c26

爬取到的商品评论数据



3. 数据可视化分析

1. 销售眼镜度数规格分析(matplotlib + json + re)

从用户购买的情况分析近视度数的分布情况

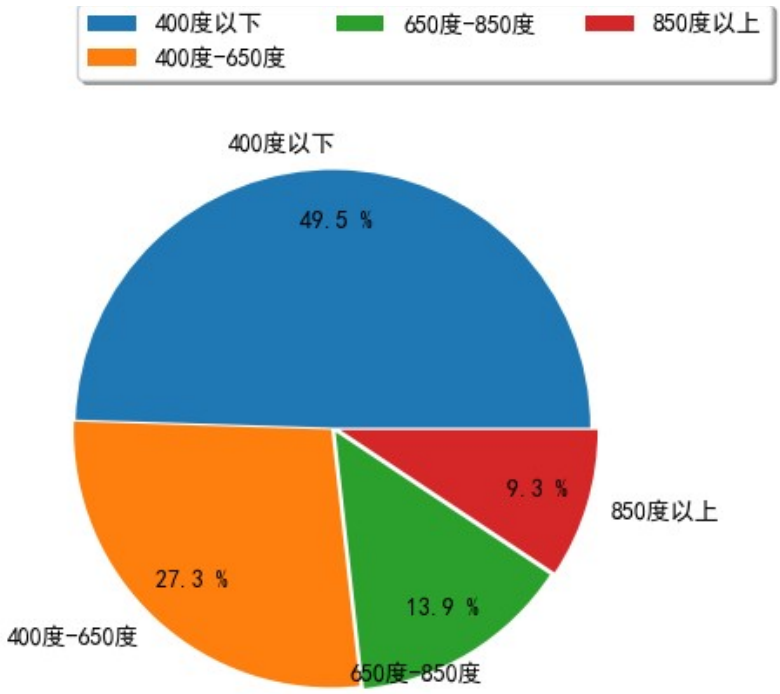
- 通过json在线解析，分析json内容
- 读取excel中的json数据并解析

有的数据json格式有错误，需要捕获异常跳过

```
try:
    dict_data = json.loads(json_data, strict=False)
    for json_item in dict_data['rateDetail']['rateList']:
        if json_item['appendComment'] is not None:
            appand_comment_count = appand_comment_count + 1
            f.write(json_item['appendComment']['content'] + "\n")
        comment_time = json_item['rateDate'].split()[0]  # 评论日期
        dushu = json_item['auctionSku'][json_item['auctionSku'].index
('镜片适合度数') + 7:]
        dushu = re.findall(r"\d+\.?\d*", dushu)  # 提取度数
        dushus.append(dushu)
        comment_time_list.append(comment_time)
        count = count + 1
except Exception:
    continue
```

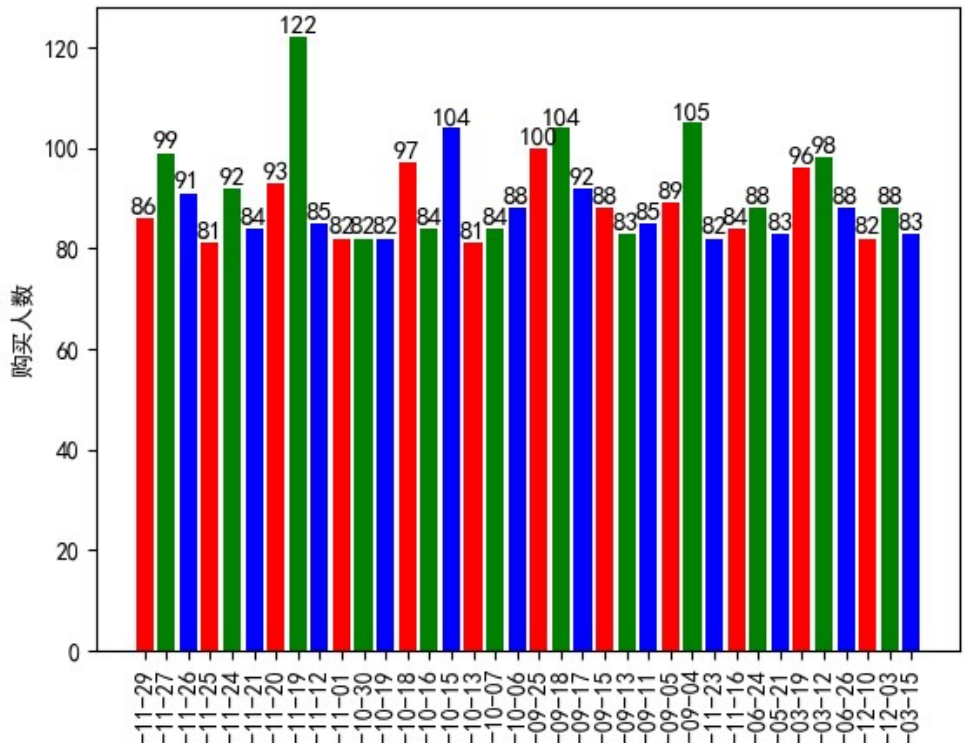
dushus 保存的是所有眼镜片销售的度数规格

comment_time_list 保存的是所有评论的日期



1. 评论日期分析

分析网购的规律



分析可以发现销量最高一天为11月19号，这个比较符合实际规律，双11销量比较高（考虑8天的快递时间）。

1. 追加评论情感分析(snowNLP)

从追评的角度分析网购眼镜片的可靠性

```
## Features

* 中文分词（[Character-Based Generative Model](http://aclweb.org/anthology//Y/Y09/Y09-2047.pdf)）
* 词性标注（[TnT](http://aclweb.org/anthology//A/A00/A00-1031.pdf) 3-gram 隐马）
* 情感分析（现在训练数据主要是买卖东西时的评价，所以对其他的一些可能效果不是很好，待解决）
* 文本分类（Naive Bayes）
* 转换成拼音（Trie树实现的最大匹配）
* 繁体转简体（Trie树实现的最大匹配）
* 提取文本关键词（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf)算法）
* 提取文本摘要（[TextRank](http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Mihalcea.pdf)算法）
* tf，idf
* Tokenization（分割成句子）
* 文本相似（[BM25](http://en.wikipedia.org/wiki/Okapi_BM25)）
* 支持python3（感谢[erning](https://github.com/erning)）
```

如果有追加评论的话，那么aooendComment字段不为None
在json字符串解析时顺便将追评保存到AppendComment.txt

```
if json_item['appendComment'] is not None:
    appand_comment_count = appand_comment_count + 1
    f.write(json_item['appendComment']['content'] + "\n")
```

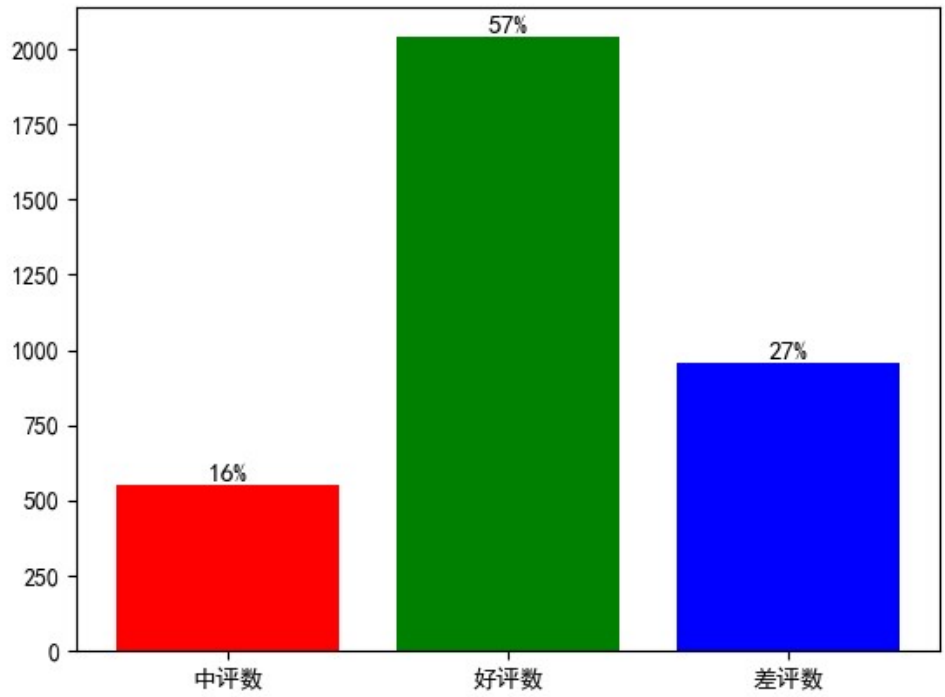
AppendComment.txt

代码：

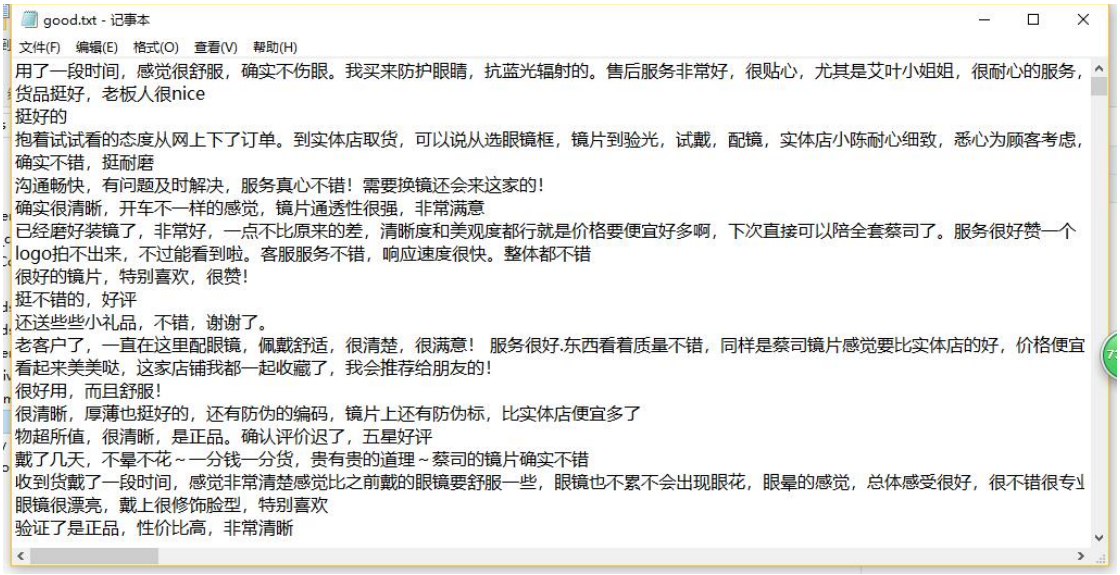
```
from snownlp import SnowNLP
import matplotlib.pyplot as plt

file = "appendComment.txt"  # 评论数据文件
bad = "bad.txt"  # 差评保存文件
good = "good.txt"
comment_dict = {}
z = 0  # 总数
with open(file, "r", encoding="gbk", errors='ignore') as text:
    bad = open(bad, "w", encoding="utf-8")
    good = open(good, 'w', encoding='utf-8')
    for comment in text:
        z += 1
        s = SnowNLP(comment)  # 文本分析
        s = s.sentiments  # 情感系数
        if s >= 0.66:
            good.write(comment)
            comment_dict['好评数'] = comment_dict.get('好评数', 0) + 1
        elif 0.66 > s > 0.33:
            comment_dict['中评数'] = comment_dict.get('中评数', 0) + 1
        else:
            bad.write(comment)  # 写入差评数
            comment_dict['差评数'] = comment_dict.get('差评数', 0) + 1
    bad.close()
    good.close()
for a, b in comment_dict.items():
    pctb = b/z*100
    # ha 文字指定在柱体中间， va指定文字位置 fontsize指定文字体大小
    plt.text(a, b + 0.05, '%.0f%%' % pctb, ha='center', va='bottom',
fontsize=11)
# 设置X轴Y轴数据，两者都可以是list或者tuple
x_axis = tuple(comment_dict.keys())
y_axis = tuple(comment_dict.values())
plt.bar(x_axis, y_axis, color='rgb')  # 如果不指定color，所有的柱体都会
是一个颜色

plt.savefig("cat_spider_img\\append_comment_analyze.png")  # 保存为图
片
plt.show()
print(comment_dict)
```

最后得到评论总数 37309 追评数 3638

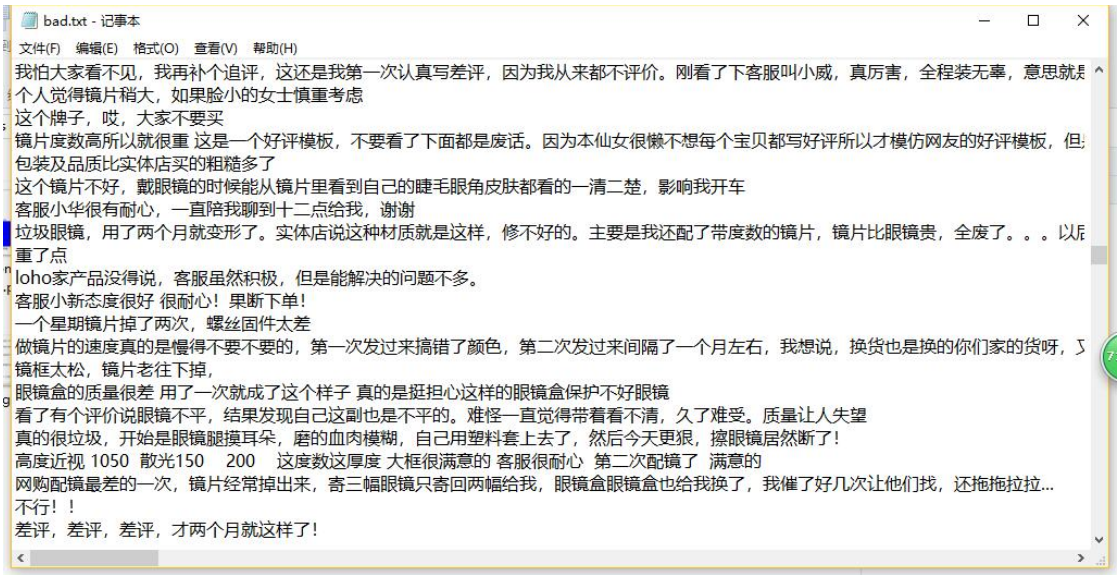
情感分析结果


57%
27%
16%
2000
1750
1500
1250
1000
750
500
250
0
中评数
好评数
差评数

好评结果


good.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
用了一段时间，感觉很舒服，确实不伤眼。我买来防护眼睛，抗蓝光辐射的。售后服务非常好，很贴心，尤其是艾叶小姐姐，很耐心的服务，
货品挺好，老板人很nice
挺好的
抱着试试看的态度从网上下了订单。到实体店取货，可以说从选眼镜框，镜片到验光，试戴，配镜，实体店小陈耐心细致，悉心为顾客考虑，
确实不错，挺耐磨
沟通畅快，有问题及时解决，服务真心不错！需要换镜还会来这家的！
确实很清晰，开车不一样的感觉，镜片通透性很强，非常满意
已经磨好装镜了，非常好，一点不比原来的差，清晰度和美观度都行就是价格要便宜好多啊，下次直接可以陪全套蔡司了。服务很好赞一个
logo拍不出来，不过能看到啦。客服服务不错，响应速度很快。整体都不错
很好的镜片，特别喜欢，很赞！
挺不错的，好评
还送些些小礼品，不错，谢谢了。
老客户了，一直在这里配眼镜，佩戴舒适，很清楚，很满意！ 服务很好.东西看着质量不错，同样是蔡司镜片感觉要比实体店的好，价格便宜
看起来美美哒，这家店铺我都一起收藏了，我会推荐给朋友的！
很好用，而且舒服！
很清晰，厚薄也挺好的，还有防伪的编码，镜片上还有防伪标，比实体店便宜多了
物超所值，很清晰，是正品。确认评价迟了，五星好评
戴了几天，不晕不花～一分钱一分货，贵有贵的道理～蔡司的镜片确实不错
收到货戴了一段时间，感觉非常清楚感觉比之前戴的眼镜要舒服一些，眼镜也不累不会出现眼花，眼晕的感觉，总体感受很好，很不错很专业
眼镜很漂亮，戴上很修饰脸型，特别喜欢
验证了是正品，性价比高，非常清晰

差评结果


bad.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
我怕大家看不见，我再补个追评，这还是我第一次认真写差评，因为我从来都不评价。刚看了下客服叫小威，真厉害，全程装无辜，意思就是
个人觉得镜片稍大，如果脸小的女士慎重考虑
这个牌子，哎，大家不要买
镜片度数高所以就很重 这是一个好评模板，不要看了下面都是废话。因为本仙女很懒不想每个宝贝都写好评所以才模仿网友的好评模板，但是
包装及品质比实体店买的粗糙多了
这个镜片不好，戴眼镜的时候能从镜片里看到自己的睫毛眼角皮肤都看的一清二楚，影响我开车
客服小华很有耐心，一直陪我聊到十二点给我，谢谢
垃圾眼镜，用了两个月就变形了。实体店说这种材质就是这样，修不好的。主要是我还配了带度数的镜片，镜片比眼镜贵，全废了。。。以后
重了点
loho家产品没得说，客服虽然积极，但是能解决的问题不多。
客服小新态度很好 很耐心！果断下单！
一个星期镜片掉了两次，螺丝固件太差
做镜片的速度真的是慢得不要不要的，第一次发过来搞错了颜色，第二次发过来间隔了一个月左右，我想说，换货也是换的你们家的货呀，又
镜框太松，镜片老往下掉，
眼镜盒的质量很差 用了一次就成了这个样子 真的是挺担心这样的眼镜盒保护不好眼镜
看了有个评价说眼镜不平，结果发现自己这副也是不平的。难怪一直觉得带着看不清，久了难受。质量让人失望
真的很垃圾，开始是眼镜腿摸耳朵，磨的血肉模糊，自己用塑料套上去了，然后今天更狠，擦眼镜居然断了！
高度近视 1050 散光150 200 这度数这厚度 大框很满意的 客服很耐心 第二次配镜了 满意的
网购配镜最差的一次，镜片经常掉出来，寄三幅眼镜只寄回两幅给我，眼镜盒眼镜盒也给我换了，我催了好几次让他们找，还拖拖拉拉...
不行！！
差评，差评，差评，才两个月就这样了！