李子柒 的辣酱真的那么好吃？用Python带你多维度分析

数据森麟 2020-01-04

以下文章来源于数据不吹牛 ， 作者小Z

数据不吹牛

有趣+干货的数据分析宝藏

作者：小Z

来源：数据不吹牛

最近收到不少留言，除了夸小Z脑洞奇清的，问最多的竟然是:

“我是偏运营/业务分析的，复杂的分析算法我不会，还有什么方法能够对评价类数据做更深入分析吗？”

刚开始我会昧着良心回复“词云啊！”

然鹅总是会被DISS“词云太简单了吧，有点low！”

我狰狞一笑，虽然这个问题有点像“我长的不帅也没钱，有没有什么办法能够追到白富美”。但从数据分析的角度来看，仍不失为一个好问题。

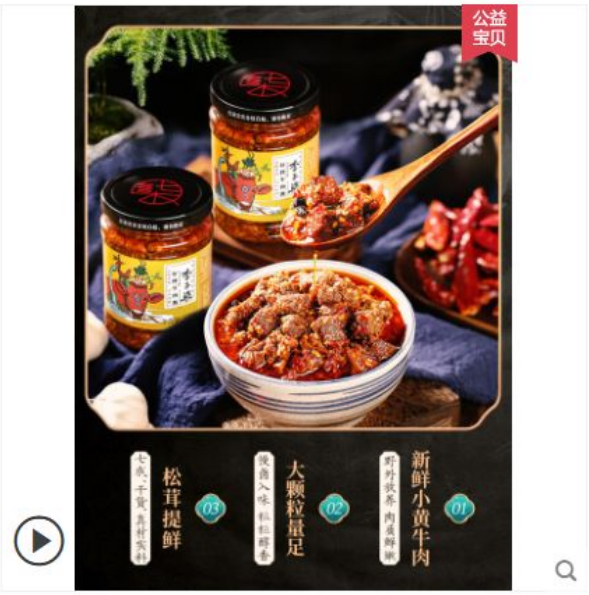
好在哪里呢？在回答之前，先瞧一个数据分析常见思维误区：

一些同学总是认为，分析不出来有价值的结果，是因为有两只拦路虎，一是数据量和维度不够，二是因为自己不会复杂的分析模型和算法。然后，**也没有再去思考，如何基于现状更进一步地分析问题**。

数据维度和算法的价值当然不言而喻，但总是把分析不出结果和价值的锅甩出去，这种**归错于外**的思维非常危险，它营造了一种**“分析不出结果，我也没办法”**的心安理得。

而“我不会高阶分析工具和方法，但基于现状，**去思考或者询问有没有更好的分析方式**”，虽然这种思维也有槽点，但本身算是一种在现阶段尝试去解决问题的思路。

So，我们循着后一种思路，以李子柒在天猫上卖的一款**辣椒酱评价为例**：



看看**基于现有的“单薄”数据维度，怎样让分析再向前迈进一步。**

明确目标

鲁迅曾经没说过：“明确分析目标，你的分析已经成功了一大半”。

做深入分析之前，面对这一堆评价数据，我们要明确，究竟想通过分析来解决什么问题？只有明确分析目标，才能把发散的思维聚焦起来。

为了给大家一个明确的分析锚点，假设我们是这款辣椒酱的产品负责人，要**基于评价，更好的获悉消费者对产品的看法**，从而为后续产品优化提供思路。

所以，我们的分析目标是**“基于评价反馈，量化消费者感知，指导优化产品”**。

注：这里给到的一个假设目标并不完美，主要是抛砖引玉，大家可以从不同的维度来提出目标假设，尝试不同分析方向。

是不是有那么一丢丢分析思路了？别急，目标还需要继续拆解。

2

拆解目标

这些年来，最有价值的一个字，便是“拆”了：

在数据分析中也是同理。

我们在上一步已经确定了“基于评价优化产品”的目标，但这只是一个笼统模糊的目标。要让目标真正可落地，“拆”是必不可少的一步。

“拆”的艺术大体可以分为两步，**第一步是换位思考。**

评价来源于客户，客户对产品有哪些方面的感知呢？我们可以闭上眼睛，幻想自己购买了这款辣椒酱。

接着进入**第二步，基于换位的逻辑拆解**，这里可以按照**模拟购物流程的逻辑来拆解：**

首先，李子柒本身有非常强的**IP光环**，大家在选购时或多或少是慕名而来。所以，在购买决策时，到底有多大比例是冲着李子柒来的？

Next，在没收到货前，影响体验的肯定是**物流**，付款到收货用了几天？派送员态度怎么样，送货上门了吗？

收到货后，使用之前，体感最强的则是**包装**。外包装有没有破损？有没有变形？产品包装是精致还是粗糙？

接下来是**产品体验**，拿辣椒酱来说，日期是否新鲜？牛肉用户是否喜欢？到底好不好吃？

吃完之后，我们建立起了对产品的立体感知——**性价比**。我花钱买这个产品到底值不值？这个价位是贵了还是便宜？实惠不实惠？

品牌、物流、包装、产品（日期、口味）和性价比五大天王锋芒初现，我们下一步需要量化消费者对于每个方面的感知。

Python实现

对于评价的拆解和量化，这里介绍一种简单粗暴的方式，**按标点把整条评论拆分成零散的模块，再设置一系列预置词来遍历。**

注：再次强调我们这篇内容的主题是“如何基于最基础的技术，做进一步的分析，这里假设我们只会最基础的python语法和pandas。

有同学会问“为什么不用分词”！此问可谓正中我怀。不过，我把这个问题当作开放式思考题留给大家——如果用分词，如何实现同样的效果，以及有什么优缺点？

言归正传，我们先看看实战爬取的评论数据，一共1794条：

```
df = pd.read_excel('李子柒辣椒酱评论.xlsx')
print('一共 %d 条评论数据' % len(df))
df.head()
```

一共 1794 条评论数据

	买家	初评内容
0	蓝**e	诚信的卖家，暴力快递碎了一瓶马上补发了，吃完继续，祝老板生意兴隆！
1	自**1	客服态度非常好（我主要是看重售后服务）、物流也棒棒的、李子柒酱超级好吃😋、有了李子柒牛肉酱、...
2	浓**浆	爱了，超符合口味，太喜欢了 李子柒我爱你
3	刘**芸	味道很不错！
4	林**3	好吃，不辣，稍贵

把每条评论按照标点拆分成短句，为了省事，用了简单的正则拆分：

```
result = []
for i in df['初评内容']:
    #根据标点符号来拆分内容
    result.append(re.split('[:: ,， .。!！~·`\;；  ……、]',i))
result[:5]
```

```
[['诚信的卖家', '暴力快递碎了一瓶马上补发了', '吃完继续', '祝老板生意兴隆', ''],
 ['客服态度非常好（我主要是看重售后服务）',
  '物流也棒棒的',
  '李子柒酱超级好吃😋',
  '有了李子柒牛肉酱',
  '麻麻再也不用担心我吃不饱啦',
  ''],
 ['爱了', '超符合口味', '太喜欢了', '李子柒我爱你'],
 ['味道很不错', ''],
 ['好吃', '不辣', '稍贵']]
```

我们发现，就算是比较长段的评论，也只是涉及到**品牌、物流、包装、产品和性价比**的部分方面，所以，我们依次去遍历匹配，看短句中有没有相关的内容，没有就跳过，有的话再判断具体情绪。

以物流为例，当短句中出现“物流”、“快递”、“配送”、“取货”等关键词，大体可以判定这个短句和物流相关。

接着，再在短句中寻找代表情绪的词汇，正面的像“快”、“不错”、“棒”、“满意”、“迅速”；负面的“慢”、“龟速”、“暴力”、“差”等。

在我们预设词的基础上进行两次遍历匹配，大体可以判断这句话是不是和物流相关，以及客户对物流的看法是正面还是负面：

```
for word in result[0]:
    #先判断是不是物流相关的
    if '物流' in word or '快递' in word or '配送' in word or '取货' in word:
        #再判断是正面还是负面情感
        if '好' in word or '不错' in word or '棒' in word or '满意' in word or '迅速' in word:
            print('物流正面倾向')
        if '慢' in word or '龟速' in word or '暴力' in word or '差' in word:
            print('物流负面倾向')
```

物流负面倾向

为方便理解，用了灰常丑陋的语法来一对一实现判断。包装、产品和性价比等其他模块的判断，也是沿用上述逻辑，只是在预设词上有所差异，部分代码如下：

```
def judge_comment(df,result):

  judges = pd.DataFrame(np.zeros(13 * len(df)).reshape(len(df),13),
            columns = ['品牌','物流正面','物流负面','包装正面','包装负面','原料正面',
                  '原料负面','口感正面','口感负面','日期正面','日期负面',
                  '性价比正面','性价比负面'])

  for i in range(len(result)):
    words = result[i]
    for word in words:
      #李子柒的产品具有强IP属性，基本都是正面评价，这里不统计情绪，只统计提及次数
      if '李子柒' in word or '子柒' in word or '小柒' in word or '李子七' in word\
        or '小七' in word:
          judges.iloc[i]['品牌'] = 1

      #先判断是不是物流相关的
      if '物流' in word or '快递' in word or '配送' in word or '取货' in word:
        #再判断是正面还是负面情感
        if '好' in word or '不错' in word or '棒' in word or '满意' in word or '迅速' in word:
          judges.iloc[i]['物流正面'] = 1
        elif '慢' in word or '龟速' in word or '暴力' in word or '差' in word:
          judges.iloc[i]['物流负面'] = 1

      #判断是否包装相关
      if '包装' in word or '盒子' in word or '袋子' in word or '外观' in word:
        if '高端' in word or '大气' in word or '还行' in word or '完整' in word or '好' in word or\
          '严实' in word or '紧' in word:
          judges.iloc[i]['包装正面'] = 1
        elif  '破' in word or '破损' in word or '瘪' in word or '简陋' in word:
          judges.iloc[i]['包装负面'] = 1

      #产品
      #产品原料是牛肉为主，且评价大多会提到牛肉，因此我们把这个单独拎出来分析
      if '肉' in word:
        if '大' in word or '多' in word or '足' in word or '香' in word or '才' in word:
          judges.iloc[i]['原料正面'] = 1
        elif '小' in word or '少' in word or '没' in word:
          judges.iloc[i]['原料负面'] = 1

      #口感的情绪
      if '口味' in word or '味道' in word or '口感' in word or '吃起来' in word:
        if '不错' in word or '好' in word or '棒' in word or '鲜' in word or\
          '可以' in word or '喜欢' in word or '符合' in word:
          judges.iloc[i]['口感正面'] = 1
        elif '不好' in word or '不行' in word or '不鲜' in word or\
          '太烂' in word:
          judges.iloc[i]['口感负面'] = 1

      #口感方面，有些是不需要出现前置词，消费者直接评价好吃难吃的，例如：
```

```
        if '难吃' in word or '不好吃' in word:
            judges.iloc[i]['口感负面'] = 1
        elif '好吃' in word or '香' in word:
            judges.iloc[i]['口感正面'] = 1

        #日期是不是新鲜
        if '日期' in word or '时间' in word or '保质期' in word:
            if '新鲜' in word:
                judges.iloc[i]['日期正面'] = 1
            elif '久' in word or '长' in word:
                judges.iloc[i]['日期负面'] = 1
        elif '过期' in word:
            judges.iloc[i]['日期负面'] = 1

        #性价比
        if '划算' in word or '便宜' in word or '赚了' in word or '囤货' in word or '超值' in word or \
            '太值' in word or '物美价廉' in word or '实惠' in word or '性价比高' in word or '不贵' in word:
            judges.iloc[i]['性价比正面'] = 1
        elif  '贵' in word or '不值' in word or '亏了' in word or '不划算' in word or '不便宜' in word:
            judges.iloc[i]['性价比负面'] = 1

    final_result = pd.concat([df,judges],axis = 1)

    return final_result
```

运行一下，结果毕现：

```
judges.head(3)
```

	买家	初评内容	品牌	物流正面	物流负面	包装正面	包装负面	原料正面	原料负面	口感正面	口感负面	日期正面	日期负面	性价比正面	性价比负面
0	蓝**e	诚信的卖家，暴力快递碎了一瓶马上补发了，吃完继续，祝老板生意兴隆！	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	自**1	客服态度非常好（我主要是看重售后服务）、物流也棒棒的、李子柒酱超级好吃😋、有了李子柒牛肉酱、...	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
2	浓**浆	爱了，超符合口味，太喜欢了 李子柒我爱你	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0

第一条评价，很明显的说快递暴力，对应“物流负面”计了一分。

第二条评价，全面夸赞，提到了品牌，和正面的物流、口感信息。

第三条评价，粉丝表白，先说品牌，再夸口感。

看起来还不赖，下面我们对结果数据展开分析。

04

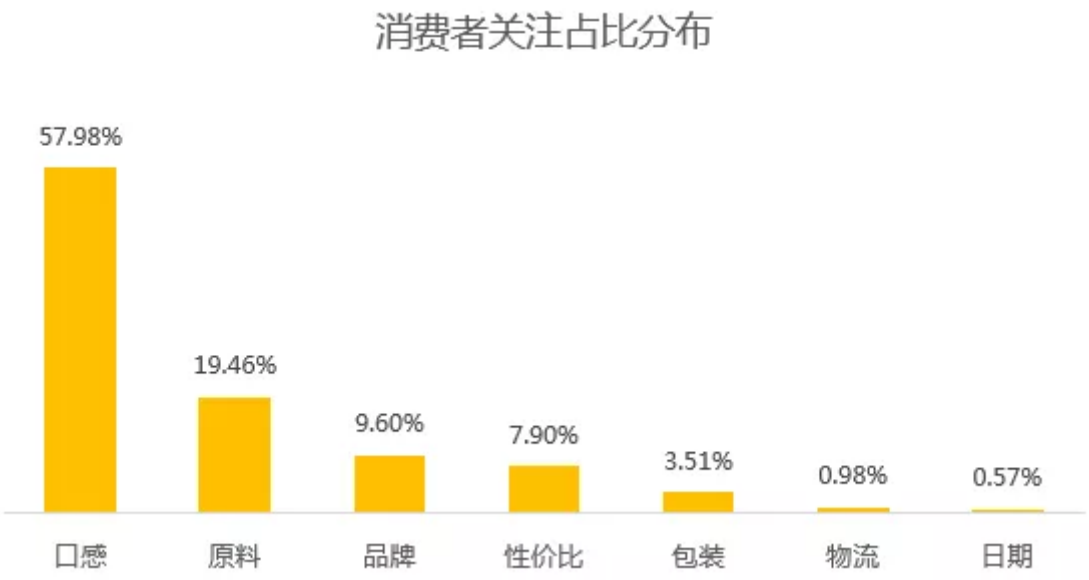
结果分析

我们先对结果做个汇总：

```
rank = judges.iloc[:,2:].sum().reset_index().sort_values(0,ascending = False)
rank.columns = ['分类','提及次数']
rank['占比'] = rank['提及次数'] / rank['提及次数'].sum()
rank['高级分类'] = rank['分类'].str[:-2]
rank
```

	分类	提及次数	占比	高级分类
7	口感正面	1117.0	0.576665	口感
5	原料正面	324.0	0.167269	原料
0	品牌	186.0	0.096025	
11	性价比正面	95.0	0.049045	性价比
3	包装正面	68.0	0.035106	包装
12	性价比负面	58.0	0.029943	性价比
6	原料负面	53.0	0.027362	原料
1	物流正面	10.0	0.005163	物流
2	物流负面	9.0	0.004646	物流
9	日期正面	7.0	0.003614	日期
8	口感负面	6.0	0.003098	口感
10	日期负面	4.0	0.002065	日期
4	包装负面	0.0	0.000000	包装

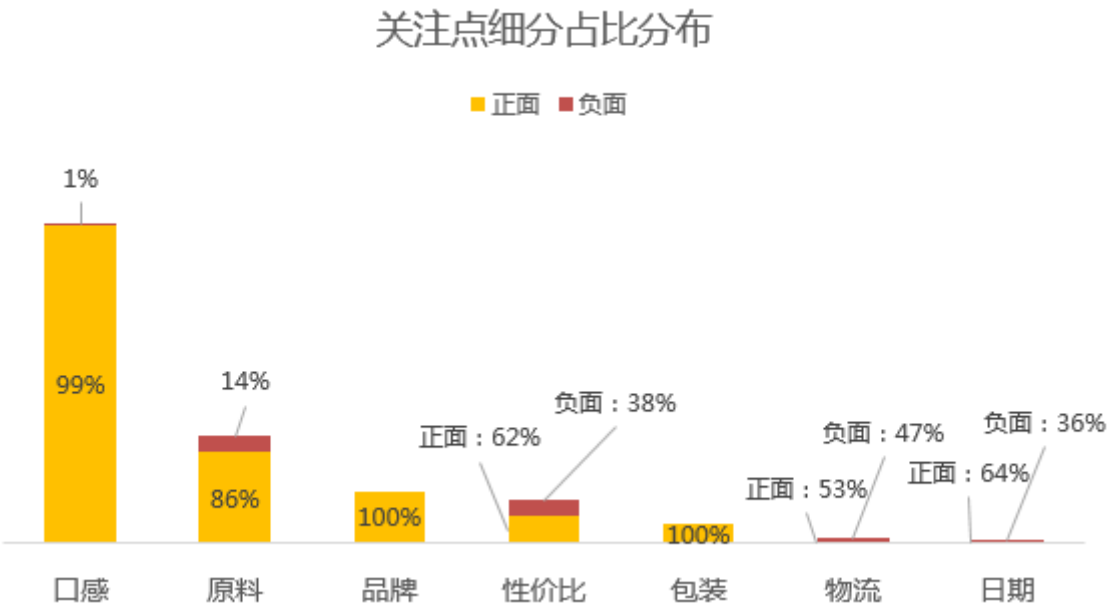
一共爬了1794条评论，评论中有提及到我们关注点的有1937次（之所以用次，是因为一条评论中可能涉及到多个方面）。粗略一瞥，口感和原料占比较高，画个图更细致的看看。



看来，**辣椒酱的口感（好不好吃）是客户最最最关注的点**，没有之一，占比高达57.98%，领先其他类别N个身位。

慢随其后的，是**原料、品牌、性价比和包装**，而物流和日期则鲜有提及，消费者貌似不太关注，或者说目前基本满足要求。

那不同类别正负面评价占比是怎么样的呢？



整体来看，主流评论以好评为主，其中口感、品牌（这个地方其实没有细分）、包装以正面评价占绝对主导。

原料和性价比，负面评价占比分别是14%和38%，而物流和日期由于本身占比太少，参考性不强。

作为一个分析师，我们从原料、性价比负面评价占比中看到了深挖的机会。

原料负面评价是单纯的在吐槽原材料吗？

```
source = judges.loc[judges['原料负面'] == 1,]
source_sum = source.sum().reset_index().loc[2:,:].sort_values(0,ascending = False)
source_sum.columns = ['分类','提及次数']
source_sum.loc[source_sum['提及次数'] > 0,:]
```

	分类	提及次数
8	原料负面	53
9	口感正面	24
7	原料正面	8
2	品牌	3
5	包装正面	2
13	性价比正面	2
14	性价比负面	2

初步筛选之后，发现事情并没有那么简单。

原料负面评价共出现了53次，但里面有24次给了口感正面的评价，甚至还有8次原料正面评价！罗生门吗？

```
source.loc[source['原料正面'] == 1,:]
```

	买家	初评内容	品牌	物流正面	物流负面	包装正面	包装负面	原料正面	原料负面
129	小**苏	味道挺不错的很喜欢，就是里面的牛肉很少，倒出来翻都只有两颗大的牛肉	0.0	0.0	0.0	0.0	0.0	1.0	1.0

这8次即正面又负面的原料评价，其实是揭了我们在预置词方面的不严谨，前面判断牛肉相关的短句，“小”就是负面，“大”就是正面，有些绝对。

而判断准确的原料差评中，虽然有一半说味道不错，但还是不留情面的吐槽了**牛肉粒之小，之少**，甚至还有因此觉得被骗。

如何让牛肉粒在体感上获取更多的好评，是**应该在产品传播层做期望控制的宣导？还是在产品层增加牛肉的“肉感”？**需要结合具体业务进一步探究。

性价比呢？

```
price = judges.loc[judges['性价比负面'] == 1,]
price_sum = price.sum().reset_index().loc[2:,:].sort_values(0,ascending = False)
price_sum.columns = ['分类','提及次数']
price_sum.loc[price_sum['提及次数'] > 0,:]
```

	分类	提及次数
14	性价比负面	58
9	口感正面	44
7	原料正面	8
13	性价比正面	7
2	品牌	3
8	原料负面	2
4	物流负面	1
5	包装正面	1

性价比相关负面评价共58次，负面情绪占性价比相关的38%。这些**负面评价消费者大多数认为价格偏贵，不划算**，还有一部分提到了通过直播渠道购买价格相对便宜，但日常价格难以接受。

坦白讲，这款辣酱的价格在线上确实属于高端价位，而价格体系是一个比较复杂的场景，这里暂不展开分析。

但是对于这部分认为性价比不符预期的客户，**是应该因此反推产品价格，还是把他们打上“价格敏感的标签”**，等大促活动唤醒收割，这是两条可以考虑并推进的道路。

物流和日期提及太少，不具备参考性，但为了不那么虎头蛇尾，我们还是顺手看一眼物流负面评价：

```
judges.loc[judges['物流负面'] == 1,].iloc[:,:2]
```

	买家	初评内容
0	蓝**e	诚信的卖家，暴力快递碎了一瓶马上补发了，吃完继续，祝老板生意兴隆！
314	原**像	一般，除了贵，物流也不是一般差，极差！！！！
315	辽**李	物流太慢了
319	恋**你	双十一买的，价格很划算，还没吃，不知道味道怎么样，物流挺慢的，很久才收到
450	全**7	物流极慢！，二十多天才收到货，无语了！┌极差，爱理不理
632	妖**吖	快递实在是太慢啦
790	厌**8	双十一最慢的一个快递
1693	胡**5	超级好评，味道服务都是非常好的。一共寄了三次来，前面两次因为快递太暴力了，各种漏油，瓶子都摔...
1700	一**那	牛肉粒很大，不错，味道也很好，货真价实的好，就是邮政快递感觉有点慢，等不及的想品尝这个酱，总...

果然，物流是一项必备需求，基本满足预期的话消费者并不会主动提及，没达预期则大概率会雷霆震怒。而物流暴力、速度太慢是两个永恒的槽点。

至此，我们基于看起来简单的评价数据，用简单浅白的方式，做了细致的拆分，并通过拆分更进一步的量化和分析，向深渊，哦不，向深入迈进了那么一丢丢。

总结

文中涉及到的代码，主要是抛砖引玉，大家还可以结合实际，做更精细的梳理和判断。在整个分析过程中，**去思考如何更深入的分析，如何明确分析方向，如何通过换位思考和流程拆解，把大目标拆成可以分析的小目标，并最终落地**，则需要在实践中反复磨练，与君共勉！

—END—

长按二维码关注我们

数据森麟公众号的交流群已经建立，许多小伙伴已经加入其中，感谢大家的支持。大家可以在群里交流关于数据分析&数据挖掘的相关内容，还没有加入的小伙伴可以扫描下方管理员二维码，进群前一定要关注公众号奥，关注后让管理员帮忙拉进群，期待大家的加入。

管理员二维码：

猜你喜欢

- 笑死人不偿命的知乎沙雕问题排行榜
- 我用Python纪念了那些被烂片收割的智商税！
- 互联网大佬学历&背景大揭秘，看看是你的老乡还是校友
- 上万条数据撕开微博热搜的真相！
- 你相信逛B站也能学编程吗？

喜欢此内容的人还喜欢

99%的Python用户都不知道的f-string隐秘技巧

Python大数据分析