问题的选择

C

2020 年
MCM / ICM
汇总表

团队控制编号 2010638

基于商品数据挖掘的销售策略推荐

总结

我们从以下四个方面进行数据分析和信息挖掘:评论之间的相关性、星级评分和有用性评分、产品品牌评分、预测产品的美誉度和星级评分对评论的影响，从而提出可靠的产品改进销售策略和建议。

首先，经过数据预处理后，我们通过 NLTK 工具对文本数据进行分词和初步情感分析，并将其量化为情感得分，范围从[-1,1]。我们采用数据可视化、描述性统计和相关性分析的方法，进一步构建多元逻辑回归模型，分析有用性评分与评论长度、评星和复合的关系。结果表明，有用性评分与评论长度呈倒“u 型”关系，与星级评分呈正“u 型”关系。

下一步，我们基于评分和评价模型进行分析。构建 LDA 分析模型，找到每个产品的主题特征，基于主题特征我们可以提出产品的改进建议。同时，我们从主题特征中总结出影响产品销售的五个指标，即质量、价格、外观、服务和尺寸。然后利用基于文本相似度的计算机搜索算法，计算每条评论的指标得分。此外，我们将评分与层次分析法相结合，确定各个指标的权重，构建一个加权的品牌评分体系。最后，我们通过系统聚类对所有产品品牌进行聚类，筛选出有潜力的优质品牌，推荐给阳光公司。

进一步，我们计算评论和星级的综合得分，并将其作为产品的美誉度，这有利于通过时间序列分析预测三个产品的未来美誉度。并描绘出三款产品具有季节性特征，在近期内可能会保持稳定的季节周期。但三款产品的信誉综合评分高峰时段存在差异，进一步分析说明，在信誉评分高峰时段，产品销量数字较大，阳光公司可根据此制定销售计划。

最后，我们分析了星评与评论的关系。通过建立分布式滞后模型，发现客户在当期的评论会受到其他客户的评分和评论的影响。同时，我们观察到情感评分和星级评分之间的时变同步性，很明显，包含积极词汇的评论会导致更高的星级评分，而包含消极词汇的评论会导致更低的星级评分。换句话说，星级评分和特定质量描述符之间存在很强的相关性。

关键词:相关性分析，多项逻辑回归，自然语言处理，
时间序列分析

内容

1.介绍

1.1 背景

随着互联网的普及和发展，线上销售逐渐取代线下销售，在销售行业占据主要地位。销售模式的转型，也将是商品公司面临的巨大挑战。分析挖掘市场信息和消费者反馈成为网络营销的关键。

亚马逊创建的在线市场为顾客提供了评价其购买商品的机会。顾客可以使用 1(低评分，低满意度)到 5(高评分，高满意度)的数量来表达他们对产品的满意度。此外，顾客还可以提交自己对产品的评价，表达自己的意见。其他客户可以利用这些评论获得一个 initial

在购买之前了解产品，并区分这些评论(称为“有用性评级”)是否有用。这些数据将反馈给公司，公司可以利用这些数据对市场进行深入分析，确定自己参与的时机，并根据客户的建议修改产品。

1.2 问题重述

阳光公司计划在网上市场推出并销售微波炉、电吹风和婴儿奶嘴。为了了解这三个商品市场，制定销售策略，需要对客户反馈数据进行分析。我们将根据给定的数据完成以下任务:

(1)为阳光公司制定销售策略。

识别潜在的重要设计特征，以提高产品的可取性。

为了完成上述两项任务，我们的具体工作如下:

- 分析星级评分、有益投票和评论之间的关系。
- 对评论和评分进行深度分析，区分产品的优缺点，从而为产品改进提出建议。

□根据对评论和评分的分析，建立产品评分体系，挑选优质品牌产品推荐给阳光公司进行销售。

- 建立美誉度评分体系，预测美誉度发展趋势。
- 分析当前评论是否受到之前星级评分和评论的影响。
- 分析基于文本的评论的特定质量描述符是否与评级水平强相关。

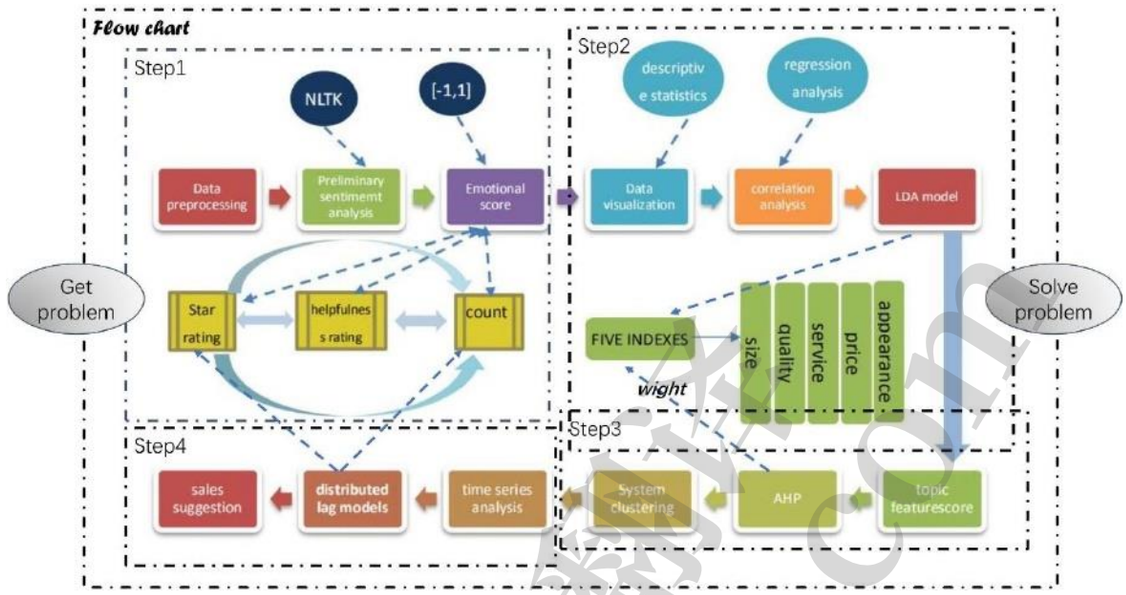


图 1:本文的流程图

2.假设和命名法

2.1 假设

我们在模型中做了几个假设。之后我们可能会放松其中的一些假设来优化我们的模型，使其更适用于复杂的现实环境。

- 所有评论都来自客户，没有自动评论。
- 在客户评分和评论中不存在蓄意抹黑产品的行为。
- 未经验证的订单在亚马逊上不出售。数据集中确认的销售额之和为 total sales。 经过 Amazon Vine 认证的会员审核，可信度高。

2.2 术语

符号	定义
党卫军	星级评分平方
数	每条评论的字数。
CS	单词数的平方
复合	情感得分
H_r	有用的投票占总投票的百分比。
wi	i的权重 th 指数
平方	六项指标(质量、规模、服务、价格、外观、星级)
老	产品信誉总分

3.模型 1:星级评分、有用性评分和评论分析

3.1 数据预处理

在进行数据分析之前，必须保证数据的可用性。如果依据的是不可靠的数据，任何衡量方法，无论其价值如何，都无法提供准确的评估。我们首先剔除包括产品名称、市场和产品类别在内的无用信息，在此基础上我们可以进行数据预处理。

改善数据集的条件，有四个步骤:数据分类、数据清洗、信息过滤、新属性设置和数据测量，如图所示。

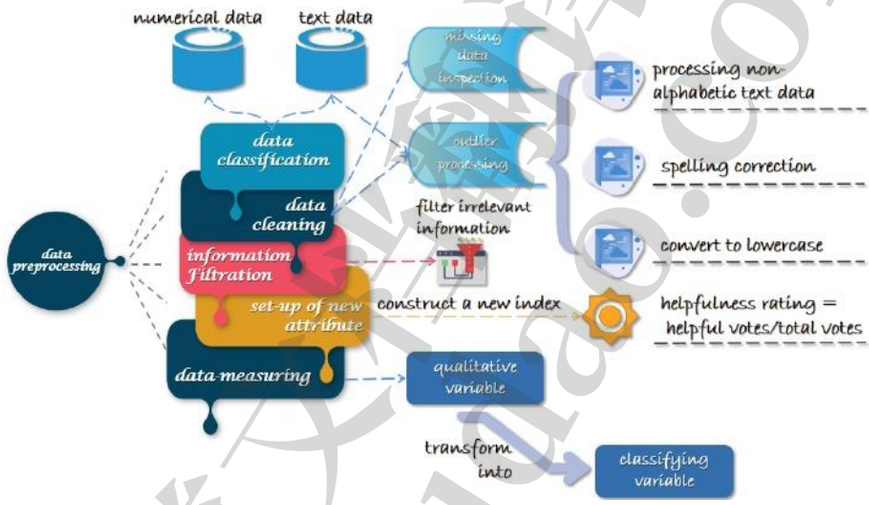


图 2:数据预处理流程图

步骤 1:首先，将数据分为数值型和文本型，分别进行处理。

第二步:在数据清洗阶段，最初我们使用 python 中的“ duplicated()方法”检查缺失值，而没有发现有缺失值的项。

第三步:异常值处理主要用在文本数据上。以安抚奶嘴为例，它包含了以下内容:

- 处理非字母文本数据。有一些非字母数据，比如:、、□、、They 电脑程序无法识别，而它们的情感倾向却可以被我们直观地识别出来。于是，我们把他们全部改造成“五星”，方便后来的计算机编程。
- 拼写校对。一般来说，英文文本中可能会有拼写错误，所以需要进行拼写检查。基于 python 程序中的“PyEnchant library”，对所有的复习文本进行检查和更正。
- 小写:计算机对“Five stars”和“Five stars”这两篇文本的反应是不一样的，而我们在数的时候期望它们是同一个单词。因此，我们将所有的文本数据统一为小写字母。

第四步:经过以上步骤后，我们的数据集已经比较容易接受了，但当我们再仔细看评论内容时，发现很多评论内容与产品无关，可以视为冗余信息，比如:

 “好清洁布”

 “我买了 hits basket 来储存杂志和报纸。 “这条毯子是我妈妈送给我的礼物，我从来没有摸过这么柔软的东西。”

以安抚奶嘴为例，需要剔除与安抚奶嘴无关的评论。我们根据文本相似度识别冗余信息，流程如下:

将商品评论视为词的集合，通过计算文本中每个词的数量，建立文本的特征向量。之后，利用向量间的余弦相似度计算文本间的相似度。如果某条评论与其他评论的平均相似度小于我们设定的阈值，则该评论被删除。

第五步:设置新属性。为了更好地分析评论的有用性，我们建立了一个名为有用性评级(helpfulness rating, Hr)的新属性，它是通过“有用的投票”除以“总投票”得到的

3.2 分词和情感分析

虽然前面提到的文本数据已经进行了初步处理，但来自评论的信息仍然需要评估。然而，评论数据是非结构化的，需要不同的机制来提取信息。情感分析或观点挖掘是计算的过程，以识别作者对一篇文本的态度是积极的、消极的还是中性的。对于评论数据，我们通过数据挖掘进行了初步的情感分析，以确定和量化每条评论的情感倾向，进一步的分析将在下面的问题中讨论。基于此，使用 NLTK(自然语言工具包)进行情感分析，NLTK 是一个领先的平台，能够通过 python 程序与人类语言年龄的数据进行工作。

使用 NLTK 库，在 Python 3 中构建了一个基本的情感分析模型。预处理 g 的操作是标记文本，对单词进行归一化，去除噪声。接下来，使用 NTLK 情感分析器构建模型，将文本与特定情感关联并量化。

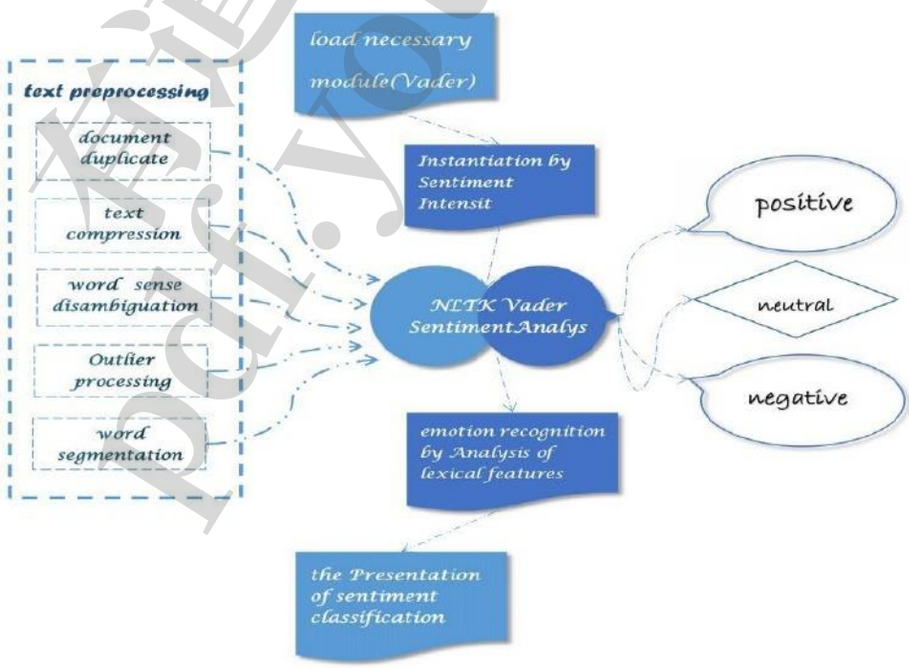


图 3:实现过程

3.3 数据描述与可视分析

下表描述了这三款产品的星评总、平均和方差。对有用票数的平均值和评论中的字数也进行了计算。下图描述了它们的分布情况。

表 1:数据统计

描述性的统计数据	总计	星级平均	星级方差	有用的票平均	数	数平均
奶嘴	18939	4.30456	1.19043	0.827182	4841870	256
微波炉	1615	3.44458	1.64524	5.62167	748130	463
吹风机	11470	4.11604	1.30033	2.17908	3268716	285

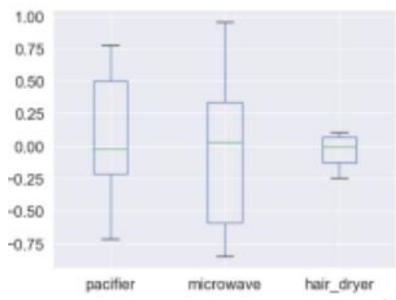


图 4:盒子图

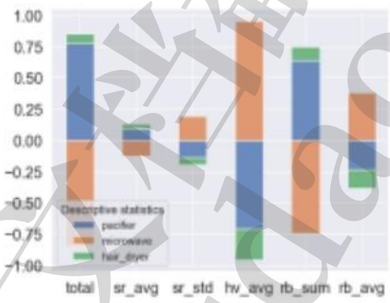


图 5:堆叠的 Barplot

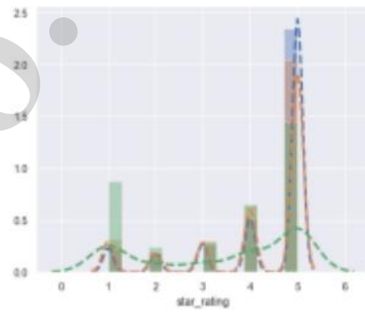


图 6:星级评级分布

将数据可视化，挖掘其中的内在规律，有助于建模。下图描述了有用性评分和星级评分之间的关系，

评论长度和情感得分。

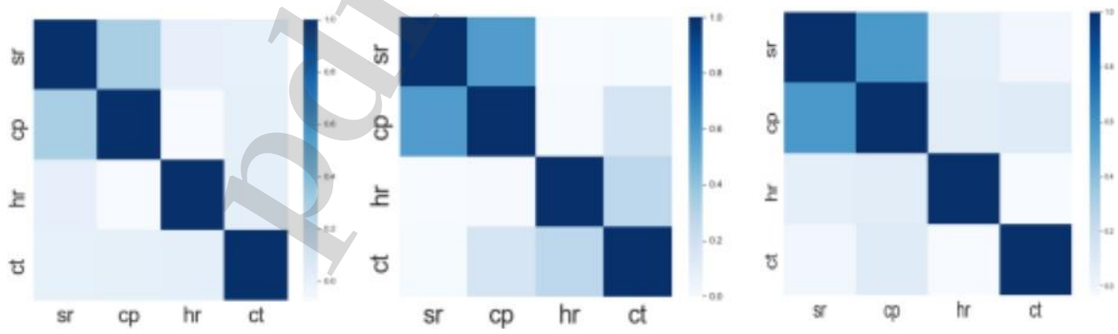


图 7:相关热图

注:sr 代表“星级”;Cp 代表“复合”;Ct 表示“计数”;Hr 代表“有用性评级”

图中颜色越深，说明两个指标之间的相关性越大。我们可以发现，情感评分和星级评分在这三张图中有很强的相关性。有用性评分和评论长度有轻微的相关性。

3.4 相关性分析

我们将相关分析的方法应用于有用投票的百分比(H$_r$)、星级评分、每条评论中的字数(计数)和情感得分(复合)。，用皮尔逊相关系数衡量如下:

$$R = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

$\underline{X}_i$代表星级、评论长度、情感得分等。Y_i 表示有用性评分，。*X, Y 表示这些指标的平均值*。r 为 H 之间的相关系数 $_r$ 以及星级、计数、复合等。

相关性分析的结果为

表 2:相关系数

指数	相关系数(微波烤箱)	相关系数(头发干燥机)	相关系数(奶嘴)
星级	-0.149 **	-0.138 **	-0.151 **
数	-0.071 **	0.356 **	0.332 **
复合	-0.024 **	0.057 **	-0.101 **

*注:**表示结果在 0.01 水平下显著(双侧检验)*

从上表我们可以看到，对于三个数据集，H 之间存在显著相关 $_r$ 以及星评、计数和复合(均在双侧检验中显著)，这与可视分析的结果一致。更具体地说，微波炉、电吹风、安抚奶嘴的星级和情绪得分与 H 呈负相关 $_r$。电吹风、安抚奶嘴计数与 H 呈正相关 $_r$，而微波炉计数与 H 呈负相关 $_r$。此外，微波炉、安抚奶嘴与 H$_r$，而电吹风的化合物与 H$_r$ 是正的。

3.5 多项 Logistic 回归模型

采用多项 logistic 回归模型进一步分析评星、评论长度和情感得分对有用性评分的影响。事实上，有用性评分是每个消费者投票(是或否)的累积结果。因此，有帮助的投票数量服从二项分布，logistic 模型适合对这类数据进行实证分析。

多元 logistic 回归可以确定解释变量 X 的作用和强度 $_n$ 在预测菌株 y 发生的概率时，设 X 为响应变量，P 为模型的响应概率，相应的回归模型如下:

$$\ln\left(\frac{p_1}{1-p_1}\right) = \alpha + \sum_{k=1}^{K}\beta_k x_{ki} \quad (2)$$

p_1 P (y)$_i$ 1 | x$_{1i}$, x$_{2i}$, x$_{ki}$)是在 x 的值的情况下 I 发生的可能性 $_{1i}$, x$_{2i}$, x$_{ki}$ 是 …

考虑到。p 也 $_1$ 是一个二项分布参数，表示评论收到有用投票的概率。x_{ki} 表示一系列自变量。β 是对应的估计系数，α 是截距，它反映了乘积层面的随机效应。

一个事件发生的概率是一个由解释变量 X 组成的非线性函数 $_i$。表达方式如下:

$$p=\frac{\exp(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)}{1+\exp(\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_n X_n)} \quad (3)$$

在这 3 个指标之上，我们引入了两个额外的变量:星评平方(SQ)和评论长度平方(CS)，以进一步探索星评和评论长度对评论有用性评分的影响。

同时，考虑到顾客是否为 vine 会员、购买是否被验证对评论真实性和有效性的影响，我们将这两个指标作为自变量。

自变量分为三个层次，分别代表直观评价(星级评分、星级评分方)、评论内容特征(评论长度、评论长度方、情感得分)和评论人 特征(vine、VP)，详细探究影响有用性评分的各种原因。

至于因变量，我们根据数值将其分为四类，

即:

$$\text{有用性评分}\ \begin{cases}1, 0.75<compound\le 1\\ 2、0.50 \text{ 化合物 } 0.75\\ 3, 0.25<compound\le 0.50\\ 4, 0\le compound\le 0.25\end{cases} \quad (4)$$

对自变量进行分块，分别做回归分析。将模型 1、2、3 中的自变量和 SPSS 得到的回归系数汇总为表。由于篇幅原因，将电吹风和安抚奶嘴的结果纳入附录。

	独立的变量	缩写	模型 1	模型 2	模型 3
直观的评价	星级	星级	0.213 **	0.469 **	0.423 **
	星级广场	党卫军	0.314 **	0.207 **	0.110 **
特征复习内容	回顾长度	数		-0.020 **	-0.019 **
	回顾长度广场	CS		0.002 **	0.002 **
	情感分数	复合		-0.146 **	-0.160 **
评论家的特征	葡萄树	葡萄树			0.033 **
	验证	验证			0.398 **
	购买	购买			

表 3:指数和回归系数注::p<0.05;* *: p < 0.01;+: p < 0.1;Ns:不显著*

由上表可知，各变量均通过显著性检验。微波炉、电吹风、奶嘴的星级回归系数均为正，星级平方项的回归系数也为正。说明星级评分与之间存在“u 型”关系

评论的有用性评分，这与 Mudambi 和 Schuff[8]的研究结论相反。我们的分析表明，评分较高或较低的评论者可能更愿意表达对产品的明确态度，从而提供更有价值的评论;而给予中等星级评价的点评者，则可能因为态度不鲜明而缺乏参考价值。

在模型 1 的基础上，在模型中加入评论长度、评论长度平方和情感得分。从表中的数据可以明显看出，评论长度项的回归系数为负，而评论长度的平方项中有一项为正，这表明评论长度和有用性评分呈倒 u 型。根据我们的分析，过短的评论往往在内容上受到限制，无法提供足够的有用信息。然而，过长的评论可能会提供很多有价值的信息，其他客户可能没有耐心花时间和精力阅读他们。因此，有用性评分高的评论应该具有适度的长度。

在模型 2 的基础上，在模型 3 中我们引入了其他与评论者相关的特征，如 Vine 和 Verified Purchase。可以看出，大部分客户并不是 vine 会员，很多购买在样本数据中也没有经过验证，因此很难对这两个指标进行分析并得出有价值的结论。尽管如此，它们还是让模型更完整、更适用。

4.模型 2-建立评分体系来确定产品定位

根据这三个数据集，每个产品都有上百个不同的品牌。如果阳光公司希望进入微波炉、电吹风、安抚奶嘴的市场，准确把握各个品牌的市场定位意义重大。不同品牌的市场定位主要是由客户的星级评价和评价来决定的。星评的处理比较简单，所以我们主要对客户评论的文本进行详细的处理。

4.1 LDA 模型

潜在狄利克雷分配(Latent Dirichlet Allocation)是 Blei 等人[2]在 2003 年提出的一种生成式主题模型。它又被称为三层贝叶斯概率模型，具有文档(d)、主题(z)和单词(W)三层结构，可以有效地对文本进行建模。基于 LDA 主题模型，我们能够挖掘数据集中潜在的主题，进而分析数据集的主要信息和相关特征词。

表 4:符号解释

,	Dirichlet 函数的先验参数
,	表示主题倍数的参数 在文档中的分布。

LDA 模型假设每一篇评论都是由各学科按一定比例随机混合的。而混合比例服从多重分布，记为:

Z | = *多项式*( ) (5)

每个题目都是由词汇表中的词按一定比例组成的，混合的比例也要服从多重分布。

$$W \mid Z, \phi = Multionmial(\phi) \tag{6}$$

在审查的条件下 d_j，生成单词 w 的概率 $_i$ 表示为:

$$P(w_j \mid d_j) = \sum_{i=1}^{K} P(w_i \mid z = s) \times P(z = s \mid d_j) \tag{7}$$

$P(w_i \mid z$ s)是单词 w 出现的概率 $_i$ 属于 s^{th} 的话题。$P(z$ $s \mid d_j)$表示 s^{th} 主题在评论中。

经过基于 LDA 的情感分析后，评论文本被聚集成三个主题，结果在每个主题下生成了 10 个最常用的单词及其对应的概率。表 5 展示了微波炉正面评价文本中潜在的话题，另一个描述了负面评价。考虑到文章篇幅有限，我们将电吹风和安抚奶嘴的结果放在附录中。

表 5:微波炉正面话题

主题 1	主题 2	主题 3	主题 1	主题 2	主题 3
好好好产品产品小井好大空间没有好喜欢好价格价格容易好使用使用好产品					
没有小尺寸喜欢喜欢喜欢					

表 6:微波炉负面话题

主题一	主题二	主题三	主题一	主题二	主题三
新没有喜欢的时间时间没有使用使用新的更多的服务老容易喜欢使用产品伟大的更多没有更多的服务模式小的小喜欢新的时间服务好好					

根据提取的三个潜在主题词进行正面和负面评价，我们可以确定客户对微波炉 5 个方面特性的态度。

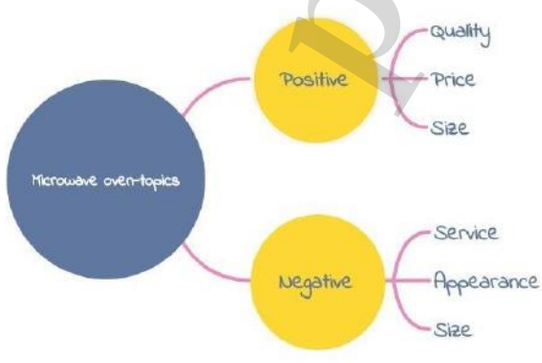


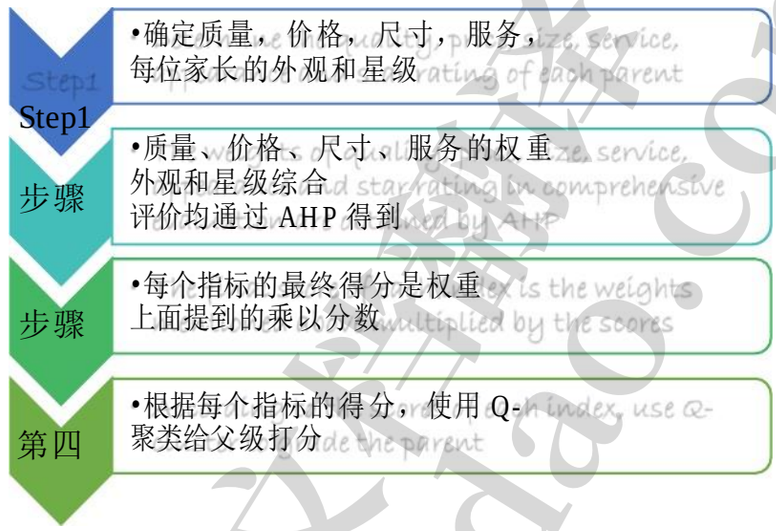
图 8:微波炉的指标

图 9:词云

这些评论中的高频词，也能反映出顾客在购买微波炉时关注的重点，那就是质量、价格、尺寸、服务和外观。这五个指标可以全面反映一个产品。我们对安抚奶嘴和吹风机做了同样的分析，那么发现这五个指标也可以作为考量是否值得购买的标准。

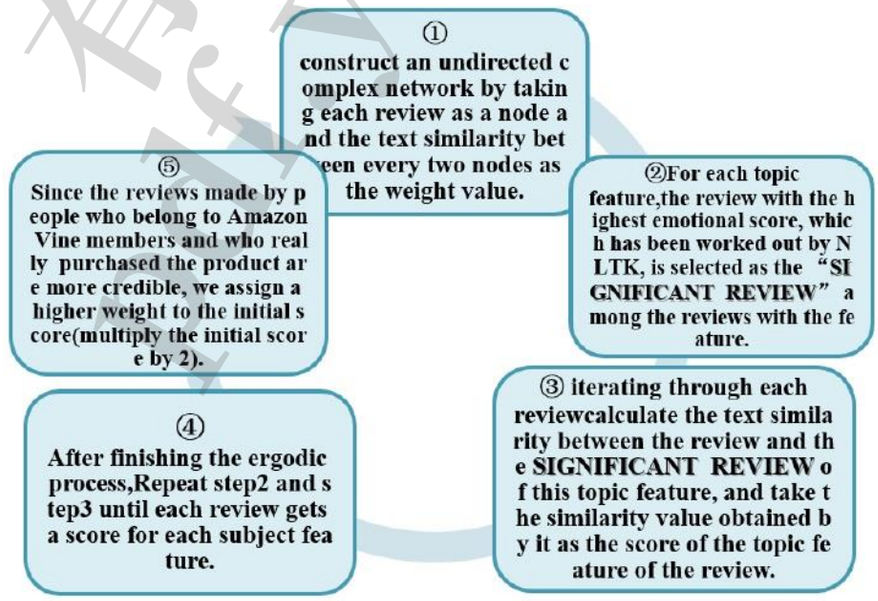
以质量、价格、尺寸、服务、外观和星级为评价指标，我们建立了一个评分体系，对这 3 个产品在当前市场上销售进行综合评价，从而获得每个品牌产品的准确市场定位，进而为阳光公司制定一个好的销售策略。

我们的品牌评分系统的步骤如下:



4.2 确定指标得分

在实现 LDA 算法后，引入复杂网络计算每条评论的重要度。我们使用遍历搜索算法来实现这个过程，通过这个过程我们可以得到所有商品的每个主题特征的加权平均得分。算法是这样的:



4.3 层次分析法(AHP)

4.3.1 层次分析法的介绍

层次分析法(AHP)可以深入分析复杂决策问题的本质、影响因素及其内在联系。它利用较少的量化信息，使决策的思维过程变得数学化，从而为复杂决策问题提供了一种简单的方法。AHP 可以将评价标准分解为不同的层次结构，然后通过求解判断矩阵的特征向量来计算每个因素的权重。

根据 LDA 模型，我们提取的指标是质量、外观、价格、大小和服务满意度。现在我们构建一级评价指标的判断矩阵:

$$A=\begin{bmatrix} w1 & w1 & W_1 & W_1 & W_1 & W1 \\ w & w & W_2 & W_3 & W_4 & W_5 \\ w & w & W_2 & W_2 & W_2 & W_2 \\ w & w & W_2 & W_3 & W_4 & W_5 \\ w & w & W_3 & W_3 & W_3 & W_3 \\ w & w & W_2 & W_3 & W_4 & W_5 \\ w & w & W_4 & W_4 & W_4 & W_4 \\ w & w & W_2 & W_3 & W_4 & W_5 \\ W_1 & W_2 & W_5 & W_5 & W_5 & W_5 \\ & & W_3 & W_4 & W_5 & W_7 \end{bmatrix} \quad (8)$$

算元素的乘积:

$$M_i=\Pi_{i=1}^{n}a_{ij}(i=1,2,3,4,5,6) \quad (9)$$

取 M 的 n 次根 $_i$:

$$W_i=\sqrt[n]{M_i} \quad (10)$$

计算特征向量:

$$W_i=\frac{W_i}{\sum} \quad (11)$$

$$\frac{1}{n}\sum\frac{(AW)_i}{W_i}$$

计算最大特征值:

$$\lambda_{max}= \quad (12)$$

其中 W =[W1 W2 W_3 W_4 W_5 $W_6]^T$

4.3.2 AHP 一致性检验

构建判断矩阵后，通过判断矩阵计算两层各元素的相对权重，进行一致性检验。不允许判断偏离一致性太大，因此需要对判断矩阵进行一致性检验。具体的检验步骤如下:

Step1:计算出的一致性指数 CI:

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{13}$$

Step2:从相关数据中检查检验判断矩阵一致性的标准

RI (n)

表 7:RI 与 n 的关系

n	1	2	3.	4	5	6	7	8	9
国际扶轮	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

注:如果一致性比小于 0.1，则 AHP 的判断矩阵具有满意的一致性 Step3:计算判断矩阵 CR 的随机一致性比:

$$CR = \frac{CI}{RI}^{1} \tag{14}$$

如果一致性比小于 0.1，则 AHP 的判断矩阵具有满意的一致性，即其一致性程度可接受。

4.3.3 AHP 的结果

经过基于 LDA 的情感分析，我们筛选出 5 个影响显著的指标，结合星级评分，有 6 个评价指标。每个指标的判断矩阵由专家打分法给出，以反映其相对临界程度。根据数据处理的结果，这些指标的判断矩阵已经通过了一致性检验，见附录。

表 8:指标判断矩阵

	质量	外观	价格	大小	服务	星评级
质量	1	2	1	2	2	1
外观	0.5	1	1	1	2	1
价格	1	1	1	2	1	1
大小	0.5	1	0.5	1	0.5	1
服务	0.5	0.5	1	2	1	0.5

表 9:评价指标及权重

	质量	外观	价格	大	服务	星评级
重量(微波炉)重量(电吹风)重量(奶嘴)	0.21	0.13	0.20	0.13	0.13	0.20
	0.22	0.14	0.21	0.10	0.13	0.20
	0.28	0.27	0.10	0.15	0.10	0.10

4.4 系统聚类分析

4.4.1 模型建立

聚类分析是一种逐步分类的方法。其主要思想是根据一定的相似性指标，对研究对象进行合理的合并和分类。用于解决样本的分类问题时称为系统聚类，用于解决变量的分类问题时称为 r 聚类。我们主要用聚类分析来解决品牌分类和评价的问题。根据不同品牌的观察指标(星级、质量、服务、外观、价格、大小)和系统聚类算法，计算品牌之间的相似度，将相似的品牌归为一类，将不同的品牌归为另一类。与小分类单元关系密切的，与大分类单元关系不密切的。

总而言之，系统聚类分析的结果就是形成一个大到小的分类谱系或聚类图。聚类图不仅可以直观地表示研究对象之间的相似关系和分类，还可以反映各类品牌，定量地表明相似程度，从而为综合评价提供良好的依据。

距离系数是系统聚类分析中常用的统计量。如果将 m 个变量上观察到的 n 个品牌视为 m 维空间中的 n 个点，则任意两个品牌之间的相似度点为 x_j 和 x_k 可以表示为 m 维空间中两点之间的距离，那么距离系数的定义呢

为:

$$d_{jk}=[\frac{1}{m}\sum_{i=1}^{m}(x_{ij}-x_{ik})^2]^{\frac{1}{2}} \tag{15}$$

相似系数是衡量品牌之间相似性的指标。每个品牌都被视为 m 维空间的一个向量，两个品牌之间的相似度 x_j 和 x_k 定义为两个向量夹角的余弦，即

$$\cos\theta_{jk}=\frac{\sum_{i=1}^{m}x_{ij}x_{ik}}{\sqrt{\quad\cdot}} \tag{16}$$

4.4.2 分类结果

我们使用 SPSS 对不同的父母进行聚类，并将其划分为 10 个等级。一流的产品是最好的，第十级的产品是最差的。在我们的销售策略中，我们会推荐阳光公司销售市场中反应较好的一流品牌。限于篇幅，文中只列出了每个产品的 5 个一流品牌。

表 10:部分分类结果

微波炉	862802057	423421857	692404913	464779766	423421857
吹风机	244516305	266176173	468944538	741916038	112413045
奶嘴	22060147	22189989	51496920	62352351	79207704

注:表中编号对应产品父级。

5.模型 3-时间序列分析

在这一节中，我们构建了一个时间序列分析模型来分析微波声誉、电吹风声誉和奶嘴声誉。然后，通过分析产品的评论和星级来预测产品声誉。

产品声誉由产品的星级评价和情感倾向组成。根据统计数据显示，消费者在购买商品时更倾向于关注评论，因此评论对产品声誉的影响比星级更大。正因为如此，我们给平均情感评分分配了 70%的权重，给平均星评分配了 30%的权重。二者之和，就是产品口碑的综合得分。

由于假设，产品信誉度得分会在时间序列上受到自相关的影响。因此，我们选择 p^{th} -阶自回归模型拟合曲线，即 AR(p)。

$$Y_n = a_1 Y_{n-1} + a_2 Y_{n-2} + \ldots + a_p Y_{n-p} + \mu_n \tag{17}$$

其中 Y 表示第 n 年的产品声誉得分。

一个$_1$ 一$_p$ 表示不同滞后阶数的影响系数。

n 表示均值为 0，方差为 2 的误差项。分布匹配白噪声过程 WN(0， 2)

首先，我们采集了三个产品信誉度得分的时间序列数据。然后我们求解信誉评分的自回归模型，计算信誉评分的时间序列自相关函数和偏自相关函数来识别顺序，并具体计算参数

p 。

各产品声誉综合得分的序列相关性刻画了一个明显的自回归结构 AR1，这意味着所涉及的函数不具有截断性质，而部分函数具有截断性质。因此，我们得出结论，p=1 和所有三个产品的声誉分数共享相同的模型设置。

$$N_t = \phi_1 N_{t-1} + \varepsilon_t \tag{18}$$

系数 1 和误差项 的均值方差与声誉不同

改变。通过计算，我们应用 MLE 来估计这三个乘积的系数。下面我们列出结果:

表 11:是不同产品的模型

产品	AR (1)		假定值	R^2
吹风机	0.125	0.046	0.007	0.901
奶嘴	0.054	0.043	0.021	0.930
微波炉	-0.075	0.050	0.000	0.952

很明显，系数检验的结果是显著的，同时整体拟合得到了一个高 r 平方的好结果。

之后我们进行过拟合和欠拟合测试，这有助于 p=1 的准确性。测试使用了信息准则的算法来同时考虑误差项和参数的复杂性。

$$AIC = \ln \sigma^2 + \frac{2p}{n} \tag{19}$$

$$BIC = \ln \sigma^2 + \frac{p}{n} \ln n \qquad (20)$$

其中 n 为样本大小。

检验表明，当 p=1 时，AIC 和 BIC 得到最小值。因此，我们认为 AR(1)的拟合是合理的。

构建 AR(1)模型来预测产品信誉度的综合得分。结果如下:

以电吹风为例，为了观察产品信誉度综合得分与时间变化趋势的关系，我们在 SPSS 中对时间序列进行了分析和预测。

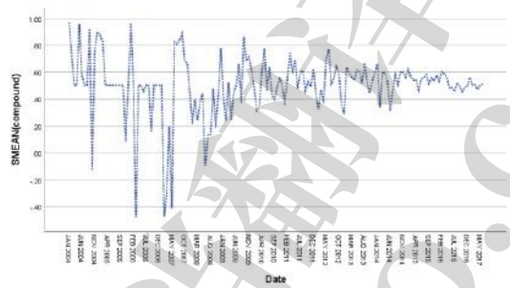


图 10:序列图

通过上面的序列图，我们发现时间序列的季节波动基本是恒定的，所以我们选择了加法模型来分解季节因素。去掉季节因素后，误差序列的值很小，因此长期趋势和循环变化序列(长期趋势+循环变化)和经过季节因子修正后的序列(长期趋势+循环变化+不规则变化，即误差差)基本可以重合。

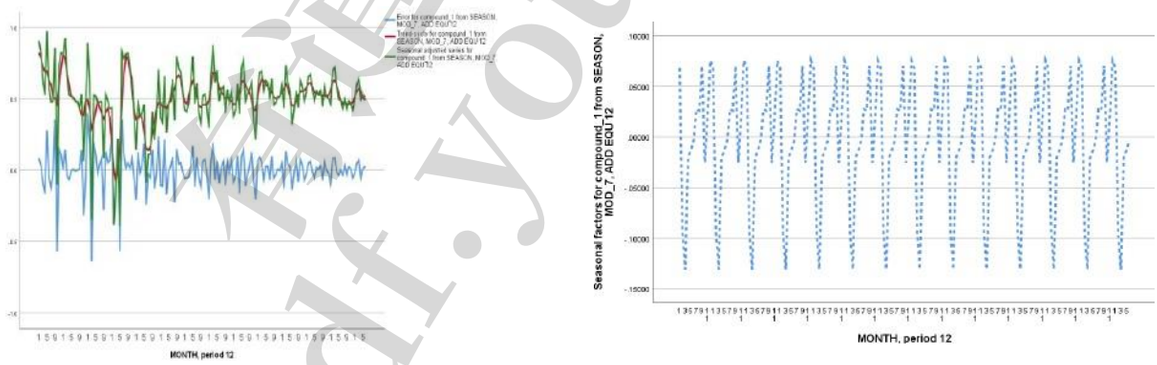


图 11:剔除季节变化趋势的序列图图 12:原始序列

通过分析季节趋势可以发现，趋势在每年的第一季度先下降，然后上升到第一个峰值。然后在大幅下降之后，有明显的上升趋势，在 8 个月左右达到第二个高峰，然后出现小幅下降。之后在 10 月左右达到峰值，然后下跌，直到明年年初。为了分析季节变化的原因，我们用时间图表介绍了年销售量。可以看到，美誉度和销量的时间点具有大致相同的季节趋势，说明随着销量的增加，客户更容易给出好评。所以我们得出结论，与实际情况是一致的。夏天，人们洗头更频繁。因此，他们倾向于在 6 月至 8 月购买吹风机，获得好评的概率更高。然而，当 10 月左右冬天来临，气温是

越冷，吹的时间越长，这也可能提高吹风机的需求和口碑。

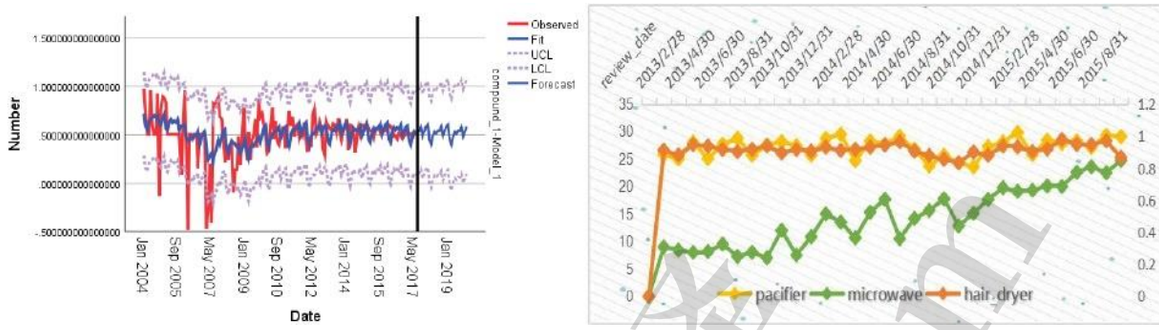


图 13:时间序列预测图14:三种产品的年销量

通过时间序列的预测，得到 2016-2019 年的数据。发现电吹风的口碑和需求与季节性高度相关，未来极有可能保持稳定的季节周期模式。微波炉和安抚奶嘴大致相同。微波炉的综合得分在 4、10 月份达到峰值，安抚奶嘴的综合得分在 11 月份达到峰值。

从上面的分析来看，认为电吹风、微波炉、安抚奶嘴的市场大致稳定。参考季节趋势，阳光公司有能力在产品美誉度得分上升时增加销售份额，或在得分即将暴跌前减少销售投入。

6.模型 4-分布滞后模型

6.1 分布滞后模型

在这一部分，我们应用分布式滞后模型来确定客户的评论是否会受到其他人的星级评分的影响。

分布滞后模型是基于被解释变量受到解释变量的影响，分布在不同时期解释变量的滞后值上。

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-s} \qquad (21)$$

S 为滞后长度。

该模型中的每个系数都反映了解释变量的每个滞后值对被解释变量的不同影响，也就是通常所说的乘数效应:

β_0:短期乘数，代表星评变化一个单位对评论情感得分的平均影响;

β_i:延迟乘数，表示在中变化一个单位的平均影响

前一时期星评对评论情感评分的影响;

$\sum_{i=0}^{s} \beta_i$:长期乘数，表示星评级变化的总影响

滞后效应。

6.2 阿尔蒙法

利用滞后为 3 个周期的二阶 Almon 多项式分布滞后模型，建立当前评论与之前的星级评分和评论之间的回归方程，从而判断是否存在之前的星级评分影响当前客户评论的现象。我们选取产品母体为 732252283 的电吹风、392768822 的安抚奶嘴和 423421857 的微波炉的数据进行分析。

利用 Almon 方法估计了如下的有限分布滞后模型。系数用二次多项式近似。

$$\begin{aligned} Y_t &= \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 X_{t-3} + u_t \\ \beta_0 &= \alpha_0 \\ \beta_1 &= \alpha_0 + \alpha_1 + \alpha_2 \\ \beta_2 &= \alpha_0 + 2\alpha_1 + 4\alpha_2 \\ \beta_3 &= \alpha_0 + 3\alpha_1 + 9\alpha_2 \end{aligned} \tag{22}$$

那么原模型就可以转化为

$$\begin{aligned} Y_t &= \alpha + \alpha_0 Z_{0t} + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \mu_t \\ Z_{0t} &= X_t + X_{t-1} + X_{t-2} + X_{t-3} \\ Z_{1t} &= X_{t-1} + 2X_{t-2} + 3X_{t-3} \\ Z_{2t} &= X_{t-1} + 4X_{t-2} + 9X_{t-3} \end{aligned} \tag{23}$$

通过回归可以得到分布滞后模型的最终估计公式。回归结果如下:

表 12:回归结果

	β0	β1	β2	β3	R^2
吹风机	1.13221* *	0.32379* *	-0.05354	0.00020	0.902197
奶嘴	0.79191* *	0.21743* *	-0.03039 (0)	0.04845	0.929296
微波	1.05871* *	0.17710* *	-0.13628	0.11858	0.835728

Note:: p<0.05; **: p<0.01; +: p<0.1; ns: not significant*

三款产品过往星级评分综合得分与当前评论情感得分之间的回归方程如下:

$$\begin{aligned} Y_t &= -0.342472 + 1.13221X_t + 0.32379X_{t-1} \\ Y_t &= -0.054537 + 0.79191X_t + 0.21743X_{t-1} \\ Y_t &= -0.226388 + 1.05871X_t + 0.17710X_{t-1} \end{aligned} \tag{24}$$

因此，可以解释为，客户的当前评论与过去的星级评级和评论之间存在相关性，即客户的当前评价会受到过去的评级和评价的影响。

6.3 星级评分和评论之间的相关性分析

根据时间序列模型，可以快速得到客户评论的拐点。通过分析这些拐点对应的评论中具体的质量描述符，以及

将它们与评论对应的星级进行对比，很容易知道这些词是否与星级密切相关。在上面的模型中，已经得到了评论和星级评分在每个时间段的平均得分。通过比较这两种变化随时间的同步性，我们可以得出结论:带有积极词汇的评论对应的星级评分更高，而带有消极词汇的评论对应的星级评分更低。(由于篇幅原因，我们只把微波炉的图片放在文字中，电吹风和奶嘴的图片放在附录中。)

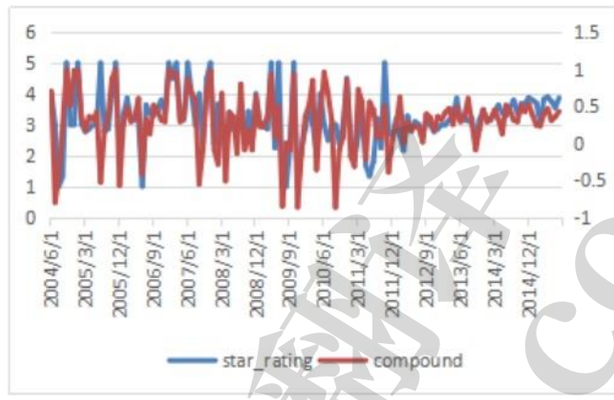


图 15:星 评和评论顺序图

但从图表中仍然可以看出，一些星评较高的评论情感得分较低，因此不排除存在高分、差评(或反之)和虚假评论。

7.敏感性分析

在时间序列分析模型中，我们给星评和评论分别赋予 0.3 和 0.7 权重，代表每个客户对其购买的产品的综合评价。基于评论重要性高于星级评分的原则，为了检验权重的合理性，我们改变星级评分的权重，分别在 0.4 和 0.2 的权重比下检验模型结果是否发生显著变化。结果如下所示:

表 13:星级评分权重敏感性分析

产品	*AR (1)*[a]	变化(单位:%)	*AR (1)*[b]	变化(单位:%)
吹风机	0.126	0.8%	0.127	20%
奶嘴	0.054	-0.9%	0.054	18%
微波炉	-0.075	0.3%	-0.074	-0.6%

注意:□□(1) :星级评分权重:0.4; （1）*:星级评分权重:0.2*

从表中可以看出，回归系数的偏差不超过 1.8%，说明权重的变化对模型的结果没有显著影响，所以我们设定的 0.3 和 0.7 的权重有一定的合理性。

8.模型评估

8.1 优势

- 应用数据可视化技术对原始数据进行解读，直观简洁地呈现结果。
- 通过对 LDA 主题模型的分析，可以给出产品的优缺点，对产品的设计有帮助。
- 我们的产品品牌评价体系适用性广泛，可以帮助销售公司掌握市场上的客户评价。

8.2 缺点

- 我们的产品评分系统采用 AHP 来确定各个指标的权重，有一定的主观性。
- 我们的时间序列模型使用的是月度数据，准确性不高，也没有考虑节假日等特殊情况，可能与实际结果有较大差距，无法有效预测产品的长期声誉。

参考

[1]Arreola Elsa Vazquez,Wilson Jeffrey R. Bayesian 多隶属多分类 logistic 回归模型对高校教师和专业学生成绩随机效应的影响[J]。《公共科学图书馆•综合》,2020 年,15(1)。

刘志强，刘志强，刘志强，等。基于深度学习的深度学习方法研究[J].中国生物医学工程学报，2016,35(6):1102 - 1102。机器学习研究，2003,3:2003。

曹娟，夏田，李金涛，一种自适应 LDA 模型选择的密度方法[J]。Neurocomputing 2009(72): 1775 - 1781。

[4]康纳斯,L。、穆丹比、s.m.、舒夫、D..是《评论》还是《评论人》?确定在线评论有用性前因的多方法方法[P]。系统科学(HICSS)， 2011 第 44 届夏威夷国际会议，2011。

[5]万长轩，彭云，肖克丽，刘西平，蒋腾蛟，刘德喜。面向产品属性和观点联合抽取的关联约束 LDA 模型[J]。信息科学,2020,519。

[6]张，Christy M. K.， Dimple R. Thadani。《电子口碑传播的有效性:文献分析》。《流血会议》，2010 年，第 18 页。

[7]Kim, Soo-Min, et al. "自动评估评审有用性。2006 年自然语言处理经验方法会议论文集，2006 年，第 423-430 页。

[8]Mudambi, Susan M.;、舒夫、大卫、张哲伟。星星为什么不对齐?网络评论内容与星评分析[P]。, 2014 年。

[9]冷艳、赵薇薇、林婵、孙成丽、王荣艳、袁琪、李登旺。基于 lda 的声学场景分类数据增强算法[J]。以知识为基础的系统,2020 年。

信

亲爱的阳光公司市场总监:

我们很荣幸地通知您，经过数据分析和建模后，我们为贵公司提出了产品改进和销售策略的建议。以下是根据我们的分析提出的一些建议。

1.产品改进建议

我们应用 LDA 分析模型，找到每个产品的主题特征，其中的负面主题词可以反映客户对市场上现有产品的不满。

通过对评论文本的分析发现，“成本”、“重”、“吸烟”、“噪音”等词出现在电吹风的负面评论列表的前列。将这些高频词一一分析后，我们发现，从客户的角度来看，电吹风存在一些弊端:

- 效率低下:吹干需要很长时间。
- 重量:太重，使用不方便。
- 质量差:使用产品时，有时会有很多噪音。

所以，根据上面的反馈，我们认为电吹风产品的改进，应该从提高电吹风的工作效率和质量入手。可以在生产中采用轻质材料，减轻电吹风的重量，也应该降低其工作时的噪音。比如，可以用轻量化的材料来减轻电吹风的重量，减少电吹风工作时的噪音。

至于微波炉，“服务”、“小”等词出现在差评榜的前列。根据反馈，我们认为这款微波炉可能存在以下缺陷:

- 尺寸:有人认为微波炉太小，也有人认为小的更方便使用。
- 服务:部分微波炉的售后服务不完善。

针对以上两点，我们认为微波炉产品可以设计成不同的尺寸，以满足不同客户的多样化需求。而且，提高服务水平和客户满意度也是重中之重。

“辣”是使用频率最高的安抚奶嘴否定词。有客户认为安抚奶嘴的保温功能差，这提醒我们在设计安抚奶嘴时要加强保温功能。另外，我们注意到，客户非常重视安抚奶嘴的“外观”，因此可爱的产品设计将有助于产品的销售。

2.销售策略

通过对星级评分和评论的深入分析，我们将从产品品牌、销售产品时间、客户评价心理三个方面提出销售建议和策略。

- 推荐待售产品品牌。

基于星级评分和评论的分析，我们总结出影响产品销售的五个指标，即质量、价格、外观、服务和大小，以及指标权重，然后以此为基础构建了品牌评分体系。通过对每个评分的系统聚类，对所有产品品牌进行聚类，最终选出优质的产品品牌。部分优质产品品牌的产品母表如下:

微波	862802057	423421857	692404913	464779766	423421857
电吹风	244516305	266176173	468944538	741916038	112413045
奶嘴	22060147	22189989	51496920	62352351	79207704

我们的品牌评分系统充分考虑了每个产品的特点，所以得出的结论非常可靠。相信推荐销售的优质品牌会受到更多顾客的青睐。

- 销售机会分析。

我们通过时间序列分析，预测三款产品未来的美誉度。结果表明，三款产品与季节性高度相关，未来季节周期格局很可能趋于稳定。但是，三种产品的信誉综合得分峰值出现在不同的时间点。

安抚奶嘴在 11 月左右达到峰值，吹风机在 8 月和 10 月达到峰值，微波炉在 4 月和 10 月达到峰值。美誉度的高峰会带来销量的增加，所以我们建议贵司根据不同产品的高峰时间调整销售结构。当产品美誉度即将达到峰值时，贵公司可以增加产品的销售投入，以获得更高的收益。

- 客户评价的心理分析。

我们分析了星级和评论之间的关系。通过分布式滞后模型的研究，发现客户在当期的评论会受到其他客户的评分和评论的影响。因此，我们建议贵司在销售产品后仍应关注客户对商品的评分，并尽量使自身产品的评分处于较高水平，以免对未来销售产生不良影响。综上所述，通过关注客户的喜好，并将其作为改进产品和服务的标准，贵公司可以逐步提高产品的销量和市场份额。

这些都是我们团队提供给贵公司的建议和策略。再次感谢您抽出时间阅读我们的建议。

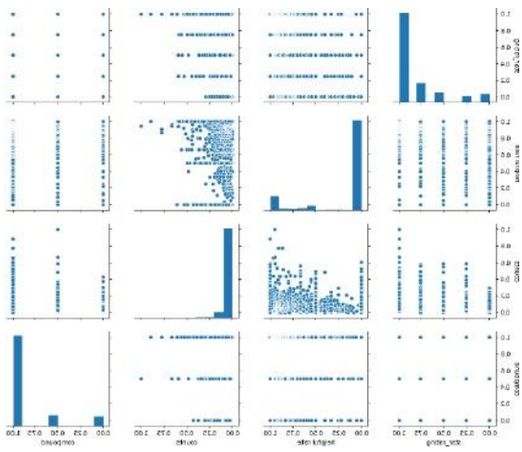
希望我们的模型和这些建议能对大家有所帮助!

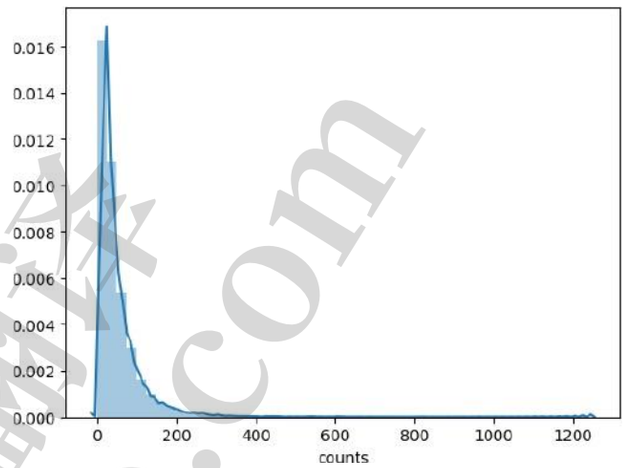
真诚地,

MCM 团队成员

附录

附录 1





单词数量的分布

相关分析

 AHP 一致性检验

一致性检验比率表

考虑到 tency 比率	曲 实施专业化	appeara 不错的	价格	大小	就是 ce	明星 评级
CR	0.0 0000	0.00857	0.00 688	0.00 0817	0.00 053	0.00 047

附录 2

分布滞后模型的回归结果

1.吹风机

变量	Coefficient
C	-0.342472
PDL01	0.323795
PDL02	-0.592878

0.215540

性病。

错误

0.06643

0

0.040057

2

0.040335

2

0.028611

2

	t 统计量	概
	-5.155354	0.0000
	7.980680	0.0000
	-14.69277	0.0000
	7.533150	0.0000

平方	0.902197	均值依赖 var S.D.依赖 var 赤池信息准则	0.648293 0.223659 -2.355086
调整后的回归 r 平方 se	0.896444 0.071974	施瓦兹的标准	-2.209099
和平方残差对数似然 f -统计量 Prob(f -统计量)	0.264190 68.76488 156.8187 0.000000	Hannan-Quinn 致命一击 。 Durbin-Watson 统计	-2.298632 1.711521

总人数的滞后分布		Coefficient	性病。错误	t-Statistic
。*	。*			
*		1.132210	0.055950	20.23567
。		0.323790	0.040570	7.980680
		-0.053534	0.040270	-1.329448
		0.000200	-0.055390	0.003580
	滞后和	1.402666	0.092430	15.17511

2.奶嘴

变量	调整后的回归 r 平方 se	Coefficient	0.609786
C		-0.054537	0.101461
PDL01		0.217428	
PDL02		-0.411152	
PDL03		0.163331	
平方		0.629296	

性病。

错误	概率。
0.109419	
0.052201	0.62011
0.052424	
0.036641	0.00001
均值依赖 var	S.D.依赖 var
赤池信息准则	0.00000
	0.00000
0.736799	
0.162422	t 统计量
-1.674949	-0.498424
	4.165216
	-7.842873
	4.457639

	0.586774	施瓦兹的标准	-1.536547
Sum squared residual Log likelihood F-statistic Prob(F-statistic)	55.08642 32.25389 0.000000	Hannan-Quinn 致命一击 。 Durbin-Watson 统计	-1.620717 1.591038

总量的滞后分布	Coefficient	性病。错误	t-Statistic
。*|	0.79191	0.08171	9.69200
。* |	0.21743	0.05220	4.16522
*。|	-0.03039	0.04982	-0.60999
.* |	0.04845	0.06081	0.79669

	滞后和	1.02739	0.14225 7.22253

3.微波炉

变量	Coefficient	性病。错误	t 统计量	概率
C	-0.226388	0.072006	-3.144010	0.00300
PDL01	0.177103	0.054822	3.230494	0.00231
PDL02	-0.597494	0.057525	-10.38663	0.00008
PDL03	0.284116	0.046045	6.170329	0.00000

平方	调整后	的回归 r 平方 se	和平方残差对数似然 f 统计量	0.521409 0.220240 -1.848816 -1.692883 -1.789889 1.492724

uinn 致命一击。

Durbin-Watson 统计

0.835728

均值依赖

0.824528

var

0.092257

S.D.依赖var

0.374502

赤池信息准则

48.37159

施瓦兹的标准

74.61626

Hannan-Q

概率(f 统计量) 0.000000

总滞后分布		Coefficient	性病。错误	t-Statistic
。*|				
.* |		1.05871	0.07634	13.8681
		0.17710	0.05482	3.23049
*。|		-0.1362		-2.4775
.* |		8	0.05500	9
		0.11858	0.07611	1.55802
	滞后和	1.2181		10.594
		2	0.11498	4

附录 3

部分系统聚类结果

1.

微波炉 10 个烤箱		9 集群	8 集群	7 集群	6 个集群	5 集群
1:109226352	1	1	1	1	1	1
2:147401377	2	2	2	2	2	2
3:149559260	3.	3.	3.	3.	3.	3.
4:155528792	2	2	2	2	2	2
5:166483932	4	4	4	4	4	2
6:168181302	4	4	4	4	4	2
7:215953885	2	2	2	2	2	2
8:242727854	5	5	4	4	4	2
9:295520151	4	4	4	4	4	2

10:305608994	1	1	1	1	1	1
11:309267414	1	1	1	1	1	1

12:311592014	6	6	5	5	5	4
13:313983847	7	7	6	6	6	5
14:379992322	4	4	4	4	4	2
15:392967251	2	2	2	2	2	2
16:423421857	1	1	1	1	1	1
17:454581724	7	7	6	6	6	5
18:459626087	1	1	1	1	1	1
19:464779766	4	4	4	4	4	2
20:486381187	5	5	4	4	4	2
21:494028413	2	2	2	2	2	2
22:494668275	5	5	4	4	4	2
23:522487135	6	6	5	5	5	4
24:523301568	1	1	1	1	1	1
25:539049610	2	2	2	2	2	2
26:542519500	5	5	4	4	4	2
27:542731946	2	2	2	2	2	2
28:544821753	2	2	2	2	2	2
29:550562680	8	8	7	7	5	4

2.

奶嘴	10 集群	9 集群	8 集群	7 集群	6 集群	5 集群
1:723849	1	1	1	1	1	1
2:	2	2	2	2	2	2

1006724					

3: 1398002	3.	3.	1	1	1	1
4: 1439995	3.	3.	1	1	1	1
5: 1448183	4	2	2	2	2	2
6: 1696639	3.	3.	1	1	1	1
7: 1892472	4	2	2	2	2	2
8: 2143250	5	4	3.	3.	3.	3.
9: 2332208	4	2	2	2	2	2
10: 2341622	4	2	2	2	2	2
11: 2775015	4	2	2	2	2	2
12: 3090006	6	5	4	4	3.	3.
13: 3146962	3.	3.	1	1	1	1
14: 3179934	5	4	3.	3.	3.	3.
15: 3729223	2	2	2	2	2	2
16: 3916839	4	2	2	2	2	2
17: 4145037	1	1	1	1	1	1
18: 4569674	4	2	2	2	2	2
19: 4649401	3.	3.	1	1	1	1
20: 4792175	2	2	2	2	2	2
21: 5180901	2	2	2	2	2	2
22: 5471085	2	2	2	2	2	2
23: 5645959	4	2	2	2	2	2
24:	4	2	2	2	2	2

25: 5749221	4	2	2	2	2	2
26 日:58486	4	2	2	2	2	2
27 日:59811	4	2	2	2	2	2
28 日:61561	2	2	2	2	2	2
29 日:67444	2	2	2	2	2	2
30: 6784496	2	2	2	2	2	2
31 日:70902	4	2	2	2	2	2

3.吹风机

	10 集群	9 集群	8 集群	7 集群	6 个集群
1: 423960	1	1	1	1	1
2: 4120409	2	2	2	1	1
3: 11468070	3.	3.	3.	2	2
4: 12536427	2	2	2	1	1
5: 14552349	2	2	2	1	1
6: 16483457	1	1	1	1	1
7: 16983648	2	2	2	1	1
8: 21033180	2	2	2	1	1
9: 21750700	2	2	2	1	1
10: 26711891	1	1	1	1	1
11: 30965255	3.	3.	3.	2	2
12: 44138644	4	4	4	3.	3.

13: 44703144	5	5	5	4	4

14: 45575190	1	1	1	1	1
15: 46450049	2	2	2	1	1
16: 46677591	2	2	2	1	1
17 岁:47684938	1	1	1	1	1
18 岁:50000317	2	2	2	1	1
19 日:54378879	2	2	2	1	1
20: 54987170	4	4	4	3.	3.
21 日:55445525	1	1	1	1	1
22 日:55520986	2	2	2	1	1
23 号:57056668	1	1	1	1	1
24: 61225676	4	4	4	3.	3.
25: 62808517	6	1	1	1	1
26 日:64142513	2	2	2	1	1
27 日:66014174	1	1	1	1	1
28 日:66259499	1	1	1	1	1
29 日:66279275	2	2	2	1	1
30: 68100320	1	1	1	1	1
31 日:68816102	2	2	2	1	1
32: 71698270	4	4	4	3.	3.
33: 74202592	7	6	6	5	5
34: 74735317	1	1	1	1	1

35 岁:77898021	2	2	2	1	1

36: 80193353	8	7	7	6	2
37: 84440271	1	1	1	1	1
38: 91277457	2	2	2	1	1
39 岁:98133587	1	1	1	1	1
40 岁:99665579	2	2	2	1	1
41:10734 1965	1	1	1	1	1
42:10819 1918	1	1	1	1	1
43:10910 6777	1	1	1	1	1
44:11093 5305	2	2	2	1	1
45:11241 3045	9	8	5	4	4
46:11526 4052	2	2	2	1	1

附录 4

部分月评和星评 1 分。吹风机

审查 _date	star_r 操作	电脑及相关知识一样
2002/3/31	3.36	0.9713
2002/4/30	5	0.7392
2002/5/31	3.829601	0.505325
2002/6/30	3.829601	0.505325
2002/7/31	5	0.9592
2002/8/31	3.5793	0.5792
2002/9/30	3.829601	0.505325

2002/10/31	3.829601	0.505325
2002/11/30	4	0.9261
2002/12/31	5	-0.1297
2003/1/31	4.5	0.7576
2003/2/28	4	0.899
2003/3/31	4	0.88425
重提/30	5	0.8294
2003/5/31	3.829601	0.505325
2003/6/30	3.829601	0.505325
2003/7/31	3.829601	0.505325
2003/8/31	3.829601	0.505325
2003/9/30	3.829601	0.505325
2003/10/31	3.829601	0.505325
2003/11/30	3.829601	0.505325
2003/12/31	3.829601	0.505325
2004/1/31	2.5	0.08485
2004/2/29	3.829601	0.505325

2004/3/31	5	0.95955
2004/4/30	3.829601	0.505325
2004/5/31	1	-0.4767
2004/6/30	3.829601	0.505325
2004/7/31	3.829601	0.505325

2004/8/31	2	0.4588
2004/9/30	3.829601	0.505325
2004/10/31	3.829601	0.505325
2004/11/30	4	0.15315
2004/12/31	3.829601	0.505325
2005/1/31	3.829601	0.505325
2005/2/28	3.829601	0.505325
2005/3/31	3.829601	0.505325
2005/4/30	1	-0.4714
2005/5/31	1	-0.2966
2005/6/30	5	0.2008
2005/7/31	5	-0.4066
2005/8/31	3.	0.8255
2005/9/30	4.333333	0.8087
2005/10/31	5	0.839467
2005/11/30	3.5	0.88995
2005/12/31	4.5	0.68685

2006/1/31	3.357143	0.667943
2006/2/28	2.875	0.453613
2006/3/31	3.309524	0.2146
2006/4/30	3.	0.396333
2006/5/31	1.666667	0.2426

2006/6 /	2	0.37806
2006/7 / 31	3.333333	0.447167

2.奶嘴

审查_date	star_r操作	电脑及相关知识一样
重提 / 30	2	0.5835
2003/5 / 31	4.4	0.60842
2003/6 / 30	4.333333	0.770685
2003/7 / 31	3.375	0.742837
2003/8 / 31	4.111111	0.487467
2003/9 / 30	4.210547	0.571278
2003/10/31	4.55	0.751675
2003/11/30	3.605442	0.523395
2003/12/31	1	0.92
2004/1 / 31	4.137209	0.697009
2004/2 / 29	4.196297	0.703307
2004/3 / 31	4.146052	0.45704
2004/4 / 30	4.5	0.6737
2004/5 / 31	2	0.9186
2004/6 / 30	2	0.8108
2004/7 / 31	4.499116	0.653345

2004/8/31	4.455441	0.646819
2004/9/30	4.361652	0.581429

2004/10/31	4.452883	0.646562
2004/11/30	4.379498	0.662509
2004/12/31	4.344347	0.62298
2005/1/31	4.333011	0.593006
2005/2/28	5	0.9696
2005/3/31	3.534091	0.492098
2005/4/30	3.883333	0.639112
2005/5/31	4.391507	0.50169
2005/6/30	4.296748	0.49209
2005/7/31	4.434554	0.525116
2005/8/31	4.332839	0.495799
2005/9/30	4.344876	0.519694
2005/10/31	5	0.8472
2005/11/30	5	0.9689
2005/12/31	4.541667	0.724842
2006/1/31	4.6375	0.759315
2006/2/28	4	0.9406

2006/3/31	5	0.9494
2006/4/30	3.534091	0.492098
2006/5/31	2.7	0.75511
2006/6/30	3.666667	0.21575
2006/7/31	4.267442	0.540129

2006/8/31	5	0.9691
2006/9/	4.25	5613-0.73
2006/10/31	5	0.95905
2006/11/30	3.49	0.73

3.微波炉

审查_date	star_r操作	电脑及相关知识一样
2004/6/30	4	0.705233
2004/7/31	3.	-0.7992
2004/8/31	1	-0.4215
2004/9/30	1.333333	0.241867
2004/10/31	5	0.9925
2004/11/30	3.	0.5063
2004/12/31	3.	0.9818
2005/1/31	5	0.9854
2005/2/28	3.	0.347794
2005/3/31	2.773333	0.169515
2005/4/30	2.865385	0.360717
2005/5/31	3.	0.320475
2005/6/30	3.	0.418019
2005/7/31	5	-0.5267

2005/8/31	2.773333	0.169515
2005/9/30	2.865385	0.360717

2005/10/31	4	0.874467
2005/11/30	5	0.9925
2005/12/31	2	-0.5707
2006/1/31	3.294118	0.253166
2006/2/28	3.857143	0.4814
2006/3/31	3.177083	0.290946
2006/4/30	3.15625	0.352928
2006/5/31	3.115385	0.602631
2006/6/30	1	-0.4215
2006/7/31	3.634921	0.319398
2006/8/31	3.384615	0.126777
2006/9/30	3.291667	0.510772
2006/10/31	3.533333	0.445201
2006/11/30	3.793333	0.313961
2006/12/31	3.378788	0.290235
2007/1/31	5	0.9868
2007/2/28	4.5	0.9294

2007/3/31	5	0.9459
2007/4/30	3.177083	0.290946
2007/5/31	3.15625	0.352928
2007/6/30	5	0.8613
2007/7/31	4	0.6808

2007/8/31	3.	0.5866
2007/9/30	4	-0.5529
2007/10/31	2.364583	-0.08993
2007/11/30	4.5	0.7006
2007/12/31	5	0.977667
2008/1/31	2.364583	-0.08993
2008/2/29	3.666667	-0.28567
2008/3/31	3.	0.6705
2008/4/30	2	-0.5106
2008/5/31	3.4	0.41524
2008/6/30	2.865385	0.360717
2008/7/31	2.666667	-0.13823
2008/8/31	3.666667	0.795633
2008/9/30	2.364583	-0.08993
2008/10/31	3.433333	0.172403
2008/11/30	2.364583	-0.08993
2008/12/31	4	0.6369

2009/1/31	3.1	0.22467
2009/2/28	3.	0.214
2009/3/31	2.865385	0.360717
2009/4/30	5	0.9257
2009/5/31	2.25	0.089067

2009/6/30	5	0.49495
2009/7/31	1	-0.8516
2009/8/31	1	0

附录 5

1.LDA 模型代码

导入 pandas 作为 pd

from snownlp 导入 snownlp

进口再保险

从 gensim 导入语料库，模型

从 nltk #。Tokenize 导入 word_tokenize

数据= pd.DataFrame(test2)

print(类型(数据))

#data_null = data.drop_duplicates()

事情就让它# data_null.to_csv (, C: /用户/联想/桌面/comments_null.csv')

#data_null_comments = data_null['contents']

事情就让它# data_null_comments.to_csv (, C: /用户/联想/桌面/ contents.txt',指数= False, enco 事情就让它叮=,utf-8')

#data_len = data_null_comments[data_null_comments.str.len()>4]

#打印(data_len)

#data_len.to_csv('contents.txt'，index=False,encoding='utf-8')

com = []

com =数据[0]。应用(λx: SnowNLP (x) .sentiments)

Data_post = data[coms>=0.01]

Data_neg = data[coms<0.01]

打印(data_post)

打印(data_neg)

事情就让它 data_post [0] .to_csv (, C: /用户/联想/桌面/ comments_positive.txt',事情就让它编码=,utf-8',头=没有)

事情就让它 data_neg [0] .to_csv (, C: /用户/联想/桌面/ comments_negative.txt',事情就让它编码=,utf-8-sig',头=没有)

以 open('C:/Users/lenovo/Desktop/comments_positive.txt'，encoding='utf-8&ap os;)为 fn1:

String_data1 = fn1.read()

模式= re.compile (u' \ t | \ n |\.|-|——|:|!|、|,|,|。|;|\)|\(|\?|"& 13 日,)

String_data1 = re.sub(pattern， ''，
String_data1)

打印(string_data1)

fp = open('C:/Users/lenovo/Desktop/comments_post.txt'， 'a'，
encoding ='utf8')

《外交政策》。写入(string_data1 +
'\n')

fp.close ()

以 open('C:/Users/lenovo/Desktop/ comments_negative .txt'，encoding='utf-8')作为 fn2:

String_data2 = fn2.read()

模式= re.compile (u' \ t | \ n |\.|-|——|;|!|、|,|,|。|;|\)|\(|\?|"& 13 日,)

String_data2 = re.sub(pattern， ''，
String_data2)

打印(string_data2)

fp = open('C:/Users/lenovo/Desktop/comments_neg.txt'， 'a'，
encoding ='utf8')

《外交政策》。写入(string_data2 +
'\n')

fp.close ()

data1 = pd.read_csv('C:/Users/lenovo/Desktop/comments_post.txt'， encoding='utf-8'， header=None)

data2 = pd.read_csv('C:/Users/lenovo/Desktop/ comments_post .txt'，
encoding='utf

事情就让它 8,头=
None)

#mycut = lambda s: '. join (word_tokenize ' (s))

#data1 = data1[0].apply(mycut)

#data2 = data2[0].apply(mycut)

#mycut = lambda s: '. join (word_tokenize ' (s))

Data1 = Data1 [0]

Data2 = Data2 [0]

事情就让它 data1.to_csv (, C: /用户/联想/桌面/ comments_post_cut.txt',指数= False,头= Fa 证交所,事情就让它编码=,utf-8')

事情就让它 data2.to_csv (, C: /用户/联想/桌面/ comments_neg_cut.txt',指数= False,头=歧视，事情就让它编码=,utf-8')

print (data2)

post = pd.read_csv('C:/Users/lenovo/Desktop/comments_post_cut.txt'，
encoding='utf-8'， header=None,error_bad_lines=False)

neg = pd.read_csv('C:/Users/lenovo/Desktop/comments_neg_cut.txt'，
encoding='utf-8'头= None, error_bad_lines = False)

stop = pd.read_csv('C:/Users/lenovo/Desktop/stoplist.txt'， encoding='utf-8'
事情就让它头= None, 9 =, tipdm'事情就让它引擎=,python')

Stop = [''， ''] + list(停止[0])

Post [1] = Post[0]。应用(lambda s: s.split(''))

Post [2] = Post[1]。Apply (lambda x: [i for i in x if i not in stop])

Neg [1] = Neg[0]。应用(lambda s: s.split(''))

```
Neg [2] = Neg[1]。应用(lambda x: [i表示 i在 x 中，如果 i 不
在 stop 中])

post_dict = corpus .dictionary (post[2])

Post_corpus = [post_dict.doc2bow(i) for i in post[2]]

Post_lda =模型。LdaModel(post_corpus, num_topics=4, id2word=post_dict)

For I in range(3):

    print
    (post_lda.print_topic(我))

事情就让它
print ()
```

neg_dict = corpus . dictionary (neg[2])

Neg_corpus = [neg_dict.doc2bow(i) for i in neg[2]]

Neg_lda =模型。LdaModel(neg_corpus, num_topics=4, id2word=neg_dict) for i in range(3):

print
(neg_lda.print_topic(我))

2.分词代码和 NLTK 模型

导入 pandas 作为 pd

导入 seaborn 作为
SNS

将 numpy 导入为 np

nltk。Tokenize 导入 word_tokenize,sent_tokenize

进口 matplotlib。Pyplot 作为
PLT

pacifier=pd.read_csv('C:/Users/lenovo/Desktop/pacifier.tsv'， sep='\t')
b = pacifier['review_body'].duplicated()

Pacifier ['marketplace'] = Pacifier ['marketplace'].str.lower() Pacifier ['product_category'].str.lower() Pacifier ['vine'] = Pacifier ['vine'].str.lower()

安抚奶嘴['verified_purchase'] =安抚奶嘴
['verified_purchase'].str.lower() text_a_count = []

For I in pacifier['review_body']:

如果
isinstance(我,str):

打印(我,类型
(i))

打印(len (i))

浮动 elif
isinstance(我):

打印(我,类型
(i))

I = str(I)

text_a_count.append (len (i))

plt。图(figsize = (8,6))

U =安抚奶嘴
['helpful_votes']

X = np.array(text_a_count)

Y = np.array(u)

p = pd.DataFrame({'review_body_length':x， 'helpful_votes':y})
#p = p.drop_duplicates(['helpful_votes'])

事情就让它事情就让它#颜色= [;r'; y', ' k', ' g', ' m'] # plt.scatter (p [' review_body_length&apos], [' helpful_votes&apos], c =颜色, 标志= & apo;事情就让它>,)

```
# plt.legend ()
# plt.show ()
V = pacifier.describe()
    导入 pandas 作为 pd
进口 nltk
导入 numpy 作为 np
nltk。Tokenize 导入 word_tokenize,sent_tokenize
事情就让它奶嘴= pd.read_excel (, C: /用户/联想/桌面
```

事情就让它/ pacifier_lower.xlsx' 9月=,\
t')

B = pacifier['review_body'].duplicated()

Test11 = pacifier.groupby('product_parent').sum()

Pacifier ['marketplace'] = Pacifier ['marketplace'].str.lower() Pacifier ['product_category'].str.lower() Pacifier ['vine'] = Pacifier ['vine'].str.lower()

安抚奶嘴['verified_purchase'] =安抚奶嘴 ['verified_purchase'].str.lower() test2 =安抚奶嘴['review_body'].tolist()

Text_a_count = []

Text_b_count = []

在 test2 中发送:

单词= word_tokenize(发送)

text_b_count.append(单词)

For word in words:

if word.isalpha() == False:

words.remove(单词)

text_a_count.append (len(单词)

Counts1 = {}

For word in words:

Counts1[单词]= Counts1 .get(单词，0)+ 1

Items = list(counts1.items())

物品。排序(键=lambda x:x[1]，反向=True)

For I in range(len(items)):

单词，count1 = items[i]

打印(“{0:< 10}{1:> 5}”。格式(词、 count1))

from collections import Counter

导入 pandas 作为 pd

导入 numpy 作为 np

C ={"counts": text_a_count，

“分裂”:text_b_count}

data = pd.DataFrame (c)

w =计数器(text_a_count)

打印(w)

W1 = set(text_a_count)

打印(text_a_count)

W2 = w.most_common(len(w))

M = []

For I in range(305):

m.append(列表
(w2[我]))

Df = pd。DataFrame(m，列数=['number'， 'count'])

事情就让它事情就让它颜色= [;r',, y', ' k', ' g', ' m'] plt.scatter (df ' number&apos, df [' count&apos], c =颜色、标记事情就让它=,事情就让它>,)plt.legend ()

plt.show ()