Summary

With the popularization of Internet technology, online shopping has gradually become an important way for people to consume. In Amazon's online marketplace, customers can rate and review after purchase, and these ratings and reviews help other customers make purchasing decisions.

First, we screened the indicators in the three data sets provided to obtain seven important indicators related to product review analysis. For the review indicators, we used a classifier based on Naive Bayesian theory to perform sentiment analysis on the review indicators. In order to achieve the purpose of quantifying indicators and considering the importance of review indicators, we further extracted the information in the reviews, and based on the reviews and its relationship with time, we obtained the review length, review density, and total review indicators.

Secondly, we use the analytic hierarchy process to establish an SRR evaluation model, and analyze the comprehensive score of a single product based on 9 related indicators such as star ratings and reviews. In the SRR evaluation model, these basic indicators are combined to obtain three effective indicators that reflect product quality: enthusiasm for comment, credibility of comment, and product popularity. Combined with time, a comprehensive evaluation level of the product is obtained. Using the SRR model, we can score a product.

Then, in order to reflect whether the rating of star ratings can lead to more reviews, we conducted an analysis of variance on the rating star index and the comment density index, and concluded that there is a significant relationship between the rating star rating and the comment density. Different star ratings have an impact on the number of reviews. Considering that triggering is a long process, we analyze the correlation between the amount of individual reviews and the total of reviews next month. It turns out that 1-star reviews can trigger more reviews, while 5-star reviews can't.

Then, we first extracted high-frequency words as features and calculated the TF-IDF of 12 words, of which 8 words were related to quality description and 4 were unrelated words. Then, a correlation analysis is performed between their TF-IDF and star ratings, and the results indicate that there is only a correlation between these specific descriptors and star indicators.

Finally, we have performed some analysis on the sensitivity and advantages and disadvantages of the model.

Keywords: sentiment analysis, analytic hierarchy process, analysis of variance, TF-IDF

1. **Introduction**
   1. 问题背景

阳光公司计划在网络上推出和销售三种产品：微波炉、婴儿奶嘴和吹风机。网络市场是一个竞争力极大的地方，且网络市场信息繁杂，如何从中提取出有用的信息且制定有效的模式处理信息，是公司改进产品、制定发展计划的重中之重。

亚马逊是一个有名的网络市场，顾客可以在产品销售网店内对产品打星，最低为1，最高为5，他们也可以在评论中发表他们对该产品的看法。

评论是以文字形式储存信息的，如何让计算机从文字中提取出信息并量化成可处理的数据是这篇论文要解决的问题，除此之外，我们还要从所给的数据中提取出各项指标，并建立相关模型帮助阳光公司在竞争中占据有利位置。

* 1. 文献回顾

1.3术语和定义

1.2我们的工作

为了从顾客的反馈中明确产品的质量，我们建立了一个名为SRR的模型，该模型能够直接或间接地分析顾客的反馈，并由此得出产品的质量。在该模型中，我们量化了评论可信度、评论积极性、产品知名度的重要指标，用这三个指标来确定单条评论中顾客的反馈，然后并结合时间，分配权重，并得出所有评论中顾客的反馈，进而得出产品的质量。在第（2）节中，我们陈述了SRR模型的基本假设。在第（4）节中，我们给出了模型中使用的每个指标的详细解释和计算。第（6）节提供了对SRR模型的全面分析。

为了明确顾客对产品的评星等级能否引发其他客户的评价，我们对单个产品额外计算了月评论总量，并进行单因素相关性分析。为了明确某些特定的单词与特定的星级指标是否有关联，我们提取出12个频率较高的单词，并与时间进行相关性分析。

**2.假设**

1. 一个产品的评论数越多，它的关注度越大

一个产品的评论数往往与它的销量正相关，评论越多，代表越多人关注过它，同时在网络市场，销量多的产品往往被置于商品列表的前列，更容易获得人们关注。

1. 时间越早，评论的价值越低

商家在运营过程中，会不断更新销售策略和生产策略，越早的评论对产品的描述差异越大，对顾客而言参考价值越低

3、没有购买且没有试用产品的顾客的评论可信度较低

亚马逊对没有购买商品的用户也提供了评论功能，没有购买也没有试用用产品的顾客多为商家寻找的刷单用户或者来自竞争对手的恶意用户，他们的评论没有实际的价值。

4、最有价值的单词应该是在特定文档中出现频率最高的单词，同时在所有文档中出现频率最低的单词。

**3数据初步分析**

3.1数据预处理

本文采用多元统计处理方法，由于数据来源于真实生活，存在一定误差，且数据间量纲不同，需要对数据进行预处理。

1. 无关变量剔除

表格中一共给给出15个相关指标，观察数据不难发现。所有数据来源于美国；product\_id与评论涉及的产品唯一对应，不同product\_id可能分属同一个product\_parent，产品分类时仅使用product\_parent指标就足够；product\_title是产品描述；product\_category为消费者类型；这些指标对于基于星级和评论的分析用处不大，因而在实际数据使用过程中，我们将marketplace、customer\_id、review\_id、product\_id、product\_title指标去除。

（2）量纲处理

考虑到本文的个别指标数据变化范围非常大，直接处理计算量大，且不好分析指标和因素间的相关关系。为了方便起见，本文将所有数据进行相对化处理，即

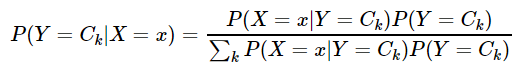


其中i表示试验样本号，j表示指标编号，xjmax表示同一编号指标的最大值。为了方便起见，以后无特殊说明均省略‘\*’符号，用xij表示相对化后的数据。相对化后的数据并不影响指标间的相关关系，但处理起来更加方便和直观。

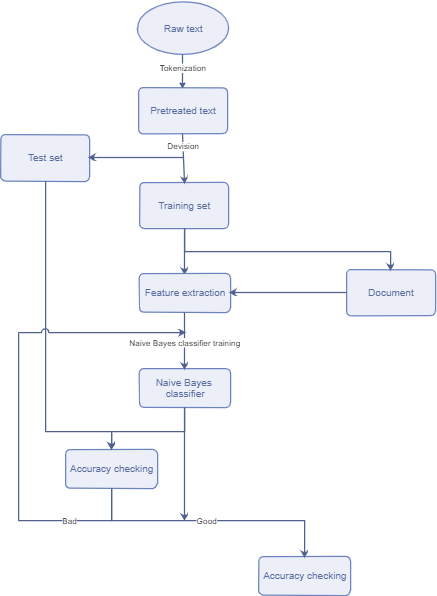
3.2 Quantitative processing of comments

3.2.1. Establishment of comment text indicators

When a product is purchased, reviews from other customers can often play an effective guiding role. Considering that reviews are unstructured data, it is difficult to directly apply them to our model. In order to effectively extract the information that a comment can contain, we split the text information of the comment into length and emotion. The length can reflect the reliability of a review. The longer the length, the more authentic the emotional expression of customers in general. Emotions indicate the probability of this review being a positive review, and represent the degree of customer dislike of the product.

The length is obtained by counting the number of words. For emotions, we try to transform them into structured data using sentiment analysis. First, we need to do preprocessing such as word segmentation, face reduction, and then select the characteristics that are most relevant to the sentiment of the comment. Textblob helped me achieve this step automatically. Then we need to use a good classification algorithm. After a series of tests and comparisons, we finally chose the Naive Bayes classifier. If the features are independent, then its performance will be very good. 

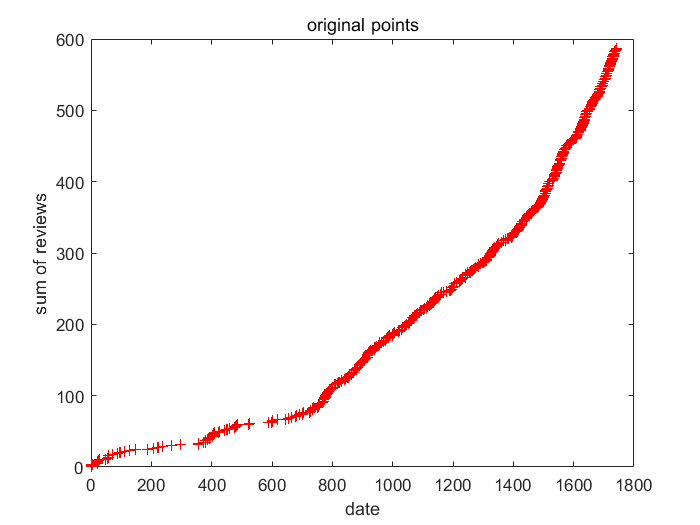
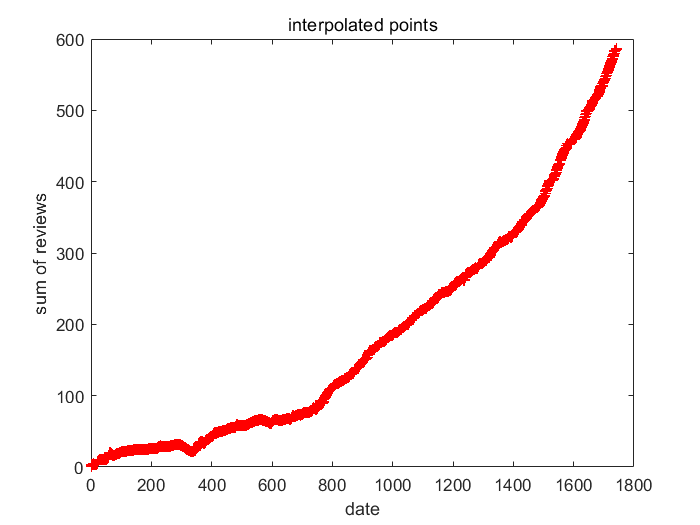
The basic principle is as follows.



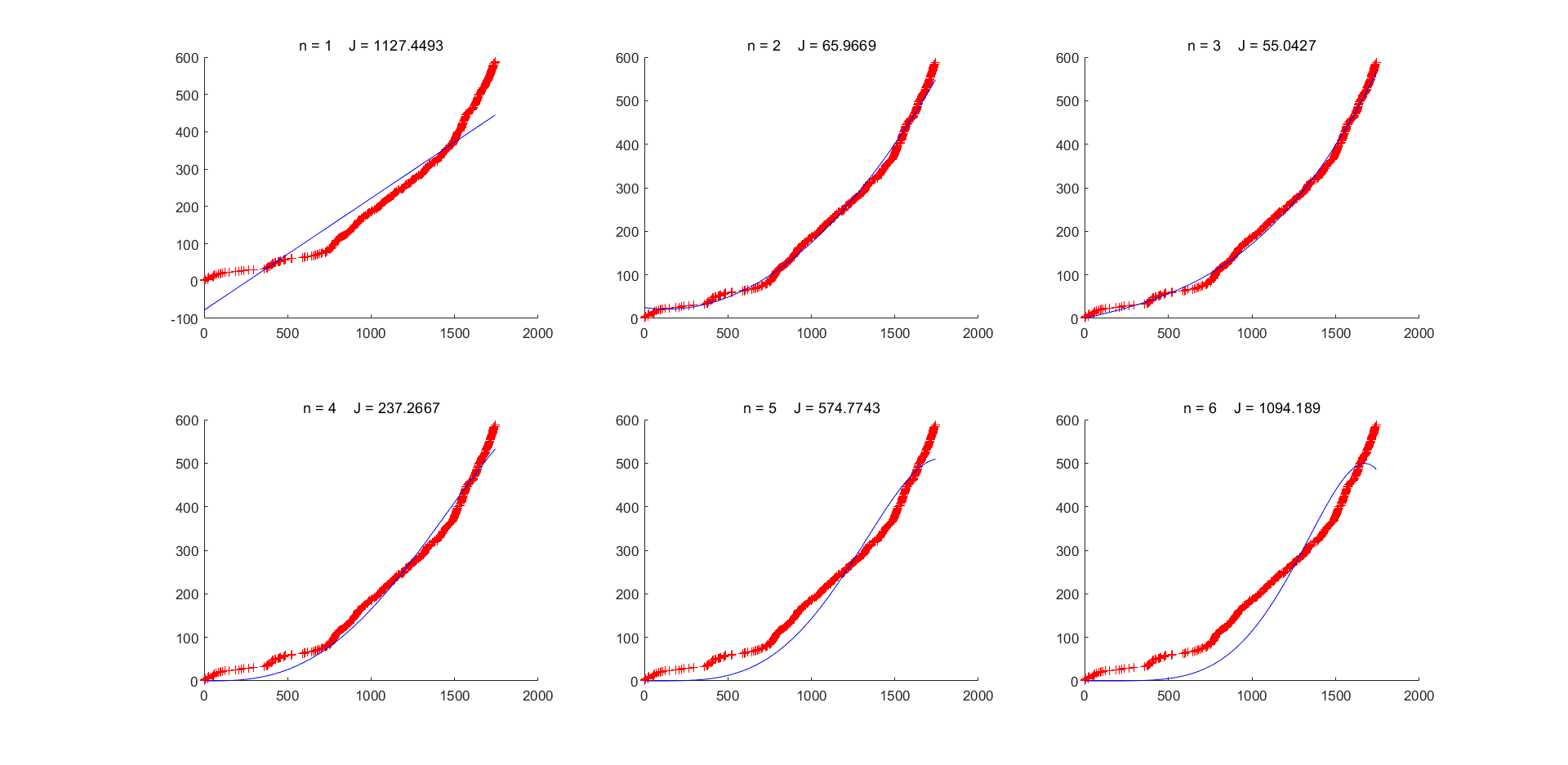
3.2.2 Comment information indicators established

For reviews, its value is not only in its textual information, but also in its quantity. For each review, we get the total number of reviews on that day and the review density.

We want to get everyday increasing rates of the total number of reviews. Hence, we plan to fit curve to the everyday total number of reviews and then derive function of the curve with respect to time. Considering that the distribution of the points is uneven, we adopt cubic spline to interpolate some points between original points in order to get more evenly distributed points, since the uneven distribution effects the weight of residual error when we fit the curve.



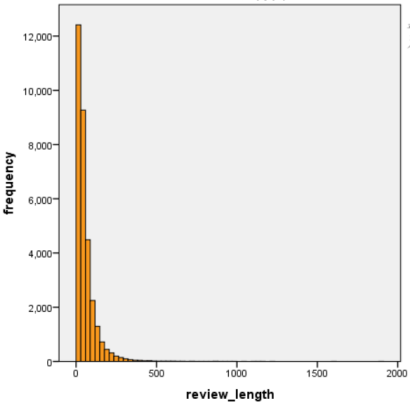
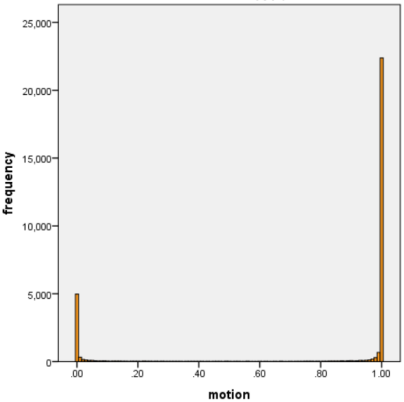
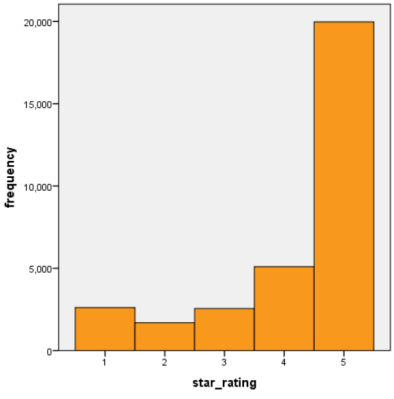
At last we plan to use linear regression to fit the curve. Then we choose the curve with the smallest J(θ).



3.3 Correlation analysis between variables

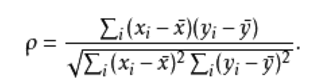
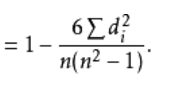
Set star\_rating to X1, sentiment to X2, review\_length to X3, helpful\_vates to X4, total\_votes to X5, density of review is X6.

In order to study the correlation between variables, we perform a normal distribution test on each variable. The histogram of the frequency distribution of some variables is shown in Figure (k)



(Figure k)

As can be seen from the figure, the distribution of the variables does not obey the normal distribution. Since the overall population of the variables is unknown, non-parametric correlation analysis is performed on the variables.



Xi and Yi are the ranks of the X and Y variables respectively. di2 is the square of the difference between the corresponding ranks of xi and yi , n is the number observation.

Spearman correlation matrix between variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | X5 | X6 |
| X1 | 1.000 | 0.487 | -0.206 | -0.216 | -0.257 | 0.108 |
| X2 |  | 1.000 | -0.625 | -0.372 | -0.391 | 0.194 |
| X3 |  |  | 1.000 | 0.357 | 0.356 | -0.251 |
| X4 |  |  |  | 1.000 | 0.896 | -0.252 |
| X5 |  |  |  |  | 1.000 | -0.245 |
| X6 |  |  |  |  |  | 1.000 |

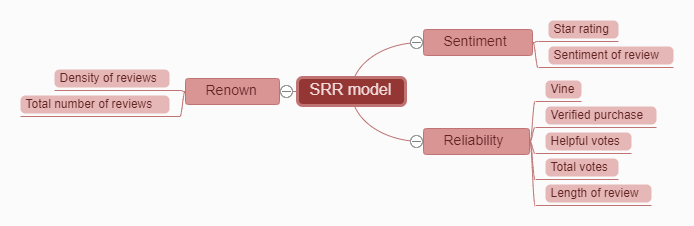
(Table 1)

From the correlation coefficient matrix (Table 1), we can see that the correlation between Y3 and Y4 is strong, and the other indicators have weak correlations. The correlation coefficient between star index and sentiment index is 0.487, the correlation coefficient between sentiment index and comment length is -0.625, and the correlation coefficient between star index and comment length index is -0.206. This is approximately in line with our experience. In actual life, customers who rate online shopping tend to rate five stars when there is no problem with the product, but sentences with low emotional scores will appear in the reviews. At the same time, the length of customer reviews of purchased problem products will increase to express their dissatisfaction.

**4.Sentiment- Reliability-Renown Model**

4.1 Model description

We use 4 indicators to evaluate the quality of a product: review reliability, review sentiment, product popularity, and time. In the modeling process, we will quantify the above four indicators and integrate them to obtain a product quality assessment score.



4.2 Comment reliability

We use 5 indicators to measure the reliability of reviews.

Whether it is a trial user: If it is a trial user, it means that the customer has used the product personally, but because it is used for a shorter time than the purchaser, the contribution to the credibility of the review is correspondingly low. This is 1 if you are a trial user, otherwise 0.

Whether it is a purchaser: If it is a purchaser, it means that the customer has used the product personally and has used it for a long time and knows the product better, so the contribution to the credibility of the review is correspondingly higher. This item is 1 if it is a buyer, otherwise it is 0.

Useful votes: The useful votes represent the affirmation of others to the comment. The more votes, the more referenced the comment and the higher the credibility.

Total votes: The total votes include useful votes and useless votes, that is, the sum of other people's positive and negative comments on the comment.

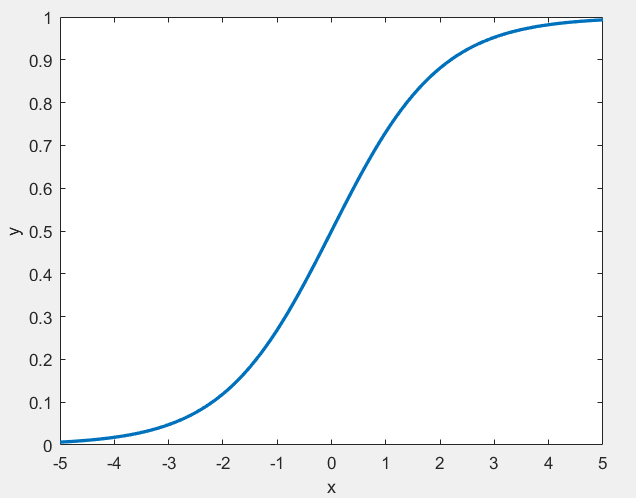
Comment length: Generally, the longer a comment is, the greater the amount of information it contains and the higher its credibility.

The mathematical expression of the comment credibility index in the SRR model equation is:

写式子

Where a12 is given in Section (5.5).

The sigmoid function in the formula is an X formula, and its image is as follow.



Since the sigmoid function is a monotonically increasing function, it converges to 1 at x towards positive infinity, and converges to 0 at x towards negative infinity, which is helpful for weight decay over time. Because the difference between the number of useful votes and the number of useless votes, and the length of the review are independent of whether they are trial or not, they should be multiplied in the formula. If the difference is larger, the comment length is longer, and the credibility of the review should increase, and we don’t want them to have a large impact on the growth of review credibility, so we use the sigmoid function to reflect this The effect of poor and increased comment length on the credibility of a comment.

4.3 Comment on emotions

We use 2 indicators to measure the positivity of reviews.

Star rating: Customers rate the product. If the star rating is higher, it means that the customer is more satisfied with it, which means that the review is more motivated.

Emotions: The emotions revealed from the semantics of the text reviews directly indicate the customer's satisfaction with the product. We extracted the emotions from the text reviews in section (3.2.1) and quantified them. The minimum is 0 and the maximum is 1. The closer to 0, the less satisfied, and the closer to 1, the more satisfied.

The mathematical expression of the comment motivation index in the SRR model equation is as follow.

写式子

The b12 is given in (5.5).

4.4 Product awareness

We use the density of reviews and the total number of reviews to measure product awareness at that point in time.

Review density: The review density represents the number of new reviews per unit time and directly reflects the popularity of the product at this point in time.

The total number of reviews: represents the total number of reviews before that point in time, and can also reflect the popularity of the product at this point in time, but it is less important than the density of reviews.

The mathematical expression of the product popularity index in the SRR model equation is as follow.

写式子

The c12 will be given in (5.5).

4.5 time weight

Because earlier comments have a larger deviation from the current product situation, and fewer people may pay attention to earlier comments, the earlier comments have a smaller impact on the SRR model score. We use the sigmoid function to assign time weights. The earlier the time, the lower the weight should be, and the closer the time is to the present, the higher the weight should be.

4.6 Calculation

The score of the SRR model measures the quality of a product. Higher scores indicate better quality, while lower scores indicate worse quality. The formula for calculating the SRR score is

写式子

The score of the SRR model measures the quality of a product. Higher scores indicate better quality, while lower scores indicate worse quality. The formula for calculating the SRR score is:

Write formula

The formula () shows the relationship between the SRR score and review reliability, review sentiment, product popularity, and time. Comment sentiment, comment reliability, and product popularity are independent of each other, and comment reliability affects comment sentiment, and the earlier the comment, the lower the reference value, so we use time to assign weight to each comment, the earlier the time, the more weight low. The score of each review is multiplied by the weight of the corresponding time and summed to get the score of the final product to measure the quality of the product.

The coefficients in the reliability of reviews, sentiment of reviews, and popularity of products are determined by the analytic hierarchy process (AHP).

The results are shown in Table ().

画表

1. 问题解答：

5.1问题一：

在数据的初步分析中，我们对数据进行筛选，删除了部分无关的指标，然后对评论信息进行处理，得到四个新的指标，情感得分，评论长度，评论密度以及评论总量。再对三个产品按product\_parent分类，其中hair\_dryer产品有473类，microwave产品有55类，pacifier产品有5432类。e而后对各个指标进行非参数相关性分析。

5.2问题二：

a.我们使用XXX模型来确定一个产品的质量。因为评论过少的产品分析误差较大，所以我们将评论数低于50的产品给筛选掉。对剩下的数据进行分析，我们对每条评论进行评分，再根据求出每天评论的平均得分，然后我们用sigmoid函数给把时间权重分配给每天的评论，最后再求和得到XXX得分。XXX得分小于0.3属于质量差的产品，XXX得分大于0.3属于质量好的产品。

b.我们选择product\_parent为123456的产品来分析，用该产品当天的XXX得分来代表它当天的声誉。此产品收到的评论数为587，评论数足够大，模型分析误差小。因为时间较早的数据对得分的偏差影响较大，所以我们将前25天的数据去掉。我们算出每天的得分后，每隔10天取一次得分，并作出得分与时间的曲线图，由图可知，该产品的声誉总体随时间减少。

c.该题需要我们找到预示产品的成功或失败的方法，我们用XXX模型来解决此问题。图X中为某产品的得分与时间的曲线图，可以看到该曲线图中有明显的拐点，该拐点说明产品得分从上升转为下降，可代表由成功向失败发展的起点，这是一个潜在失败点。所以我们可以把由下降转为上升的拐点用来预示产品失败，相同的，我们可以把由上升转为下降的拐点用来预示产品的成功。

5.3 problem d

5.3.1方差分析

为了探究特定星级是否能引发更多评论，我们首先需要确定星级对评论数量间是否存在显著性影响。这里我们采用单因子方差分析。

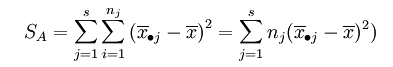
在上例中，因素A（即星级）有s（=5）个水平A1,A2,...,A5,在每一个水平Aj=(j=1,2,...,s)下进行了n次独立试验。这些结果是一个[随机变量](https://wiki.mbalib.com/wiki/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)。表中的数据可以看成来自s个不同总体（每个水平对应一个总体）的[样本值](https://wiki.mbalib.com/wiki/%E6%A0%B7%E6%9C%AC%E5%80%BC)，将各个总体的均值依次记为µ1,µ2,...,µs,则需检验假设

IMG_256

IMG_257不全相等

检验的统计量为

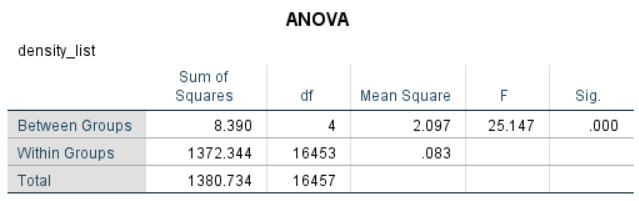
IMG_256

其中IMG_256，

在[显著性水平](https://wiki.mbalib.com/wiki/%E6%98%BE%E8%91%97%E6%80%A7%E6%B0%B4%E5%B9%B3)α下，本检验问题的拒绝域为

IMG_256

我们使用spss工具，得到的结果如图



在p<0.05情况下，拒绝原假设，即说明评星等级的不同试验水平间具有显著性差异。

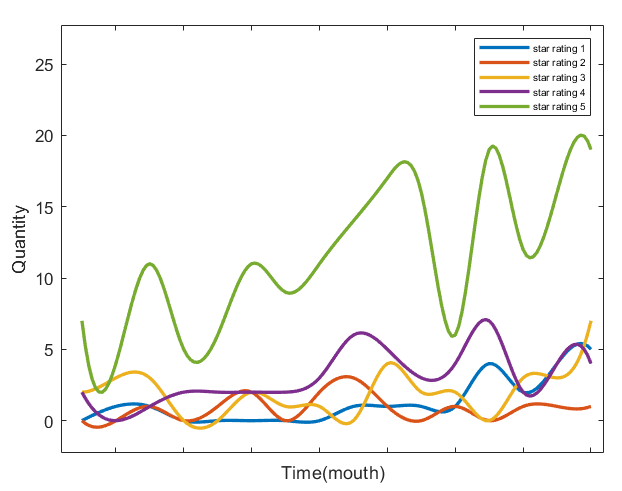
5.3.2相关性检验

为了进一步探究星级指标与评论数量之间的关系，考虑到引发评论需要一定时间，我们针对挑选了评论数最多，product\_parent为“246038397”的产品进行分析，该产品共有734条评论。我们根据原始数据计算出

Star\_quantityij:某产品第i星级指标在j月份的评论总数 i={1，2，3，4，5}

Star\_allj:某产品在j月份的下个月的所有星级评论总数

可得对应图像如下图



分别求出每个Star\_quantity变量与Star\_all变量之间的相关系数，可得

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Star\_quantity | 1 | 2 | 3 | 4 | 5 |
| R | 0.520 | 0.350 | 0.154 | 0.499 | 0.290 |

由相关系数的矩阵我们可以得出，一星和四星评论更能引发顾客的评论，而3星和5星评论则对引发评论的相关性不大。

5.4 problem e

Considering this question, we are required to select specific quality descriptors of text-based reviews to analyze the association between these words and rating levels. Hence, the first job we have to do is to select specific words. And then we plan to regard these words as features each of which denotes the frequency of the word in a sentence. In the last step, we plan to analyze the correlation between if-idf of every word and rating levels.

流程图：words selection -> correlation frequency

5.4.1Selecting quality descriptors

1.Tokenization

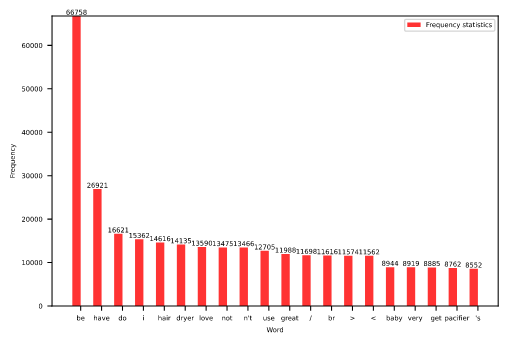
Because we want to analyze the correlation between frequency of every select quality descriptor we choose and rating levels, we plan to split the sentences into words and to count the number of each words in order to select those most frequent words that can serve as quality descriptors. Therefore, we use words tokenization that based on the space between every word to select sentences.

2. Lemmatization

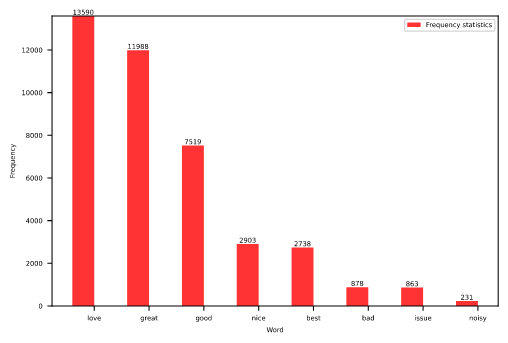
Besides, considering some words have different tenses or parts of speech, different tenses or parts of speech of the word needs to be served as the same word in order to better calculate every words’ frequency. Therefore, we need to group together them into their lemma based on its intended meaning. NLTK provides tagging and lemmatization methods that are able to precisely determining the lemma of words based on their potential meaning.

3.Counting and selection

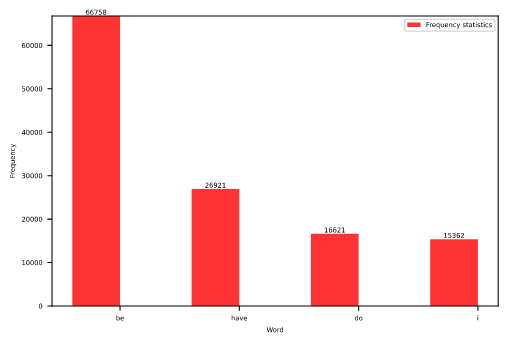
We exclude function words which generally cannot be quality descriptors. As a result, we get the word list and plot the most common 20 words’ frequency as follow.



These are the most frequent words. However, not all of these can serve as quality descriptors or even valid words such as ‘<’. We choose words that can act as quality descriptors as follow. And there shows the frequency as well.



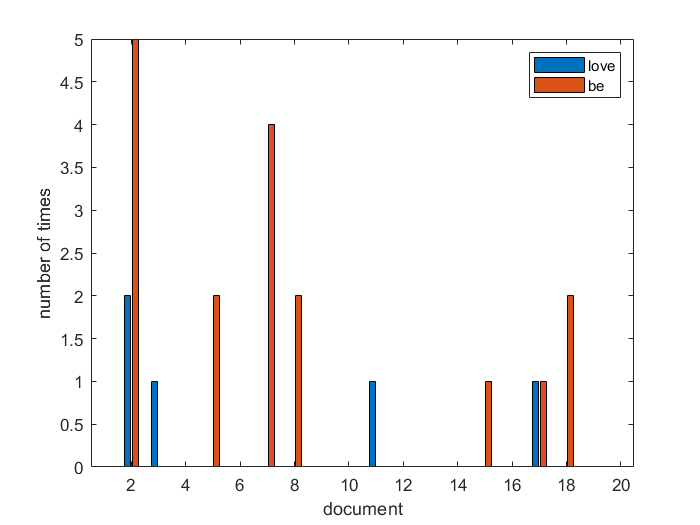
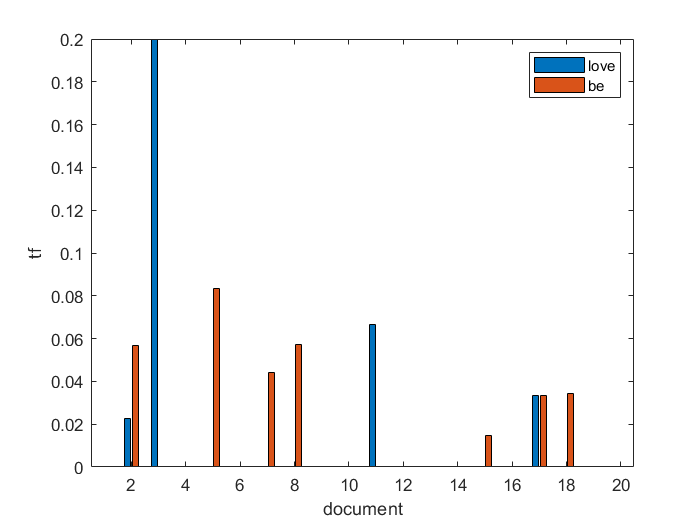
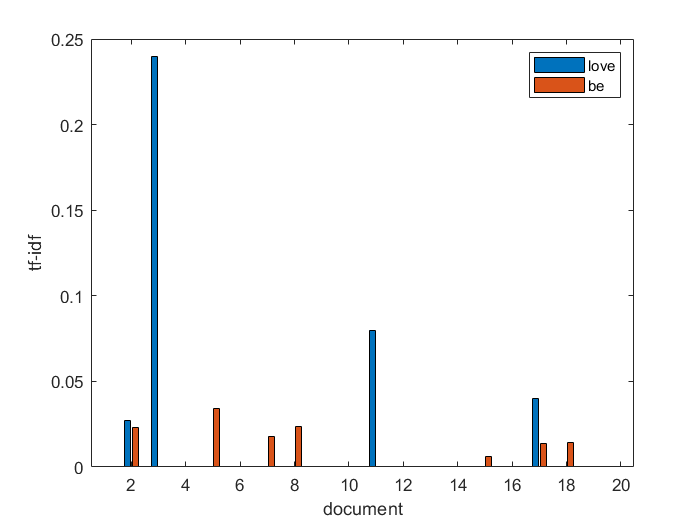
In order to directly recognize the possible correlation, we select some neutral words as control group.



5.4.2.Correlation analysis

1,Tf-idf

Tf-idf stands for term frequency–inverse document frequency. It bases on an assumption that the most valuable words should be those words that are most frequent in the specific document and the least frequent in all the documents at the same time. Since the proportion of every word is not even as shown above, we choose to use Tf-idf of those quality descriptors as features.



2. Correlation analysis

Measured by the linear correlation coefficient between two indicators, that is, the R statistic is calculated.

1 means positive linear correlation, 0 means negative linear correlation, and the two random variables near 0 are basically uncorrelated. The result is as follow.

love great good nice best bad issue noisy be

0.168 0.127 0.042 0.034 0.077 -0.075 -0.01 -0.014 -0.060

have do i

-0.014 -0.157 -0.031

We find that all of these words’ correlation coefficient are not high enough. However, some of these quality descriptors’ correlation coefficient are higher than others’. Then we find that these words are not necessary to a positive or negative review. Therefore, we combine some of these words and calculate the correlation coefficient again. The result is as follow.

love , great, good, nice, best bad, issue, noisy

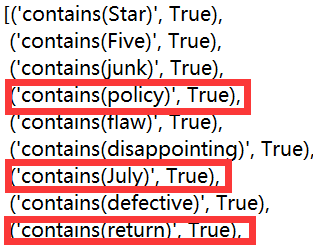
0.215 -0.113

The result is better. Hence, we draw the conclusion that they are associated with star ranking.

6. Model analysis

Our model can assess the quality of products, so that Sunshine can improve products and develop development plans based on customer feedback in a timely manner. In addition, we can use this model to identify outstanding products and learn their excellence.

There are also many deficiencies in our model. First, our SRR model is based on text reviews and star ratings. To extract emotions from text reviews and quantify this step, we used naive Bayesian classification and overly simple feature extraction. On the one hand, Naive Bayes classification is based on the assumption that features are independent of each other, and feature words are generally not independent of each other. On the other hand, feature extraction is not good enough, and some non-quality descriptor words are extracted, such as



Moreover, from the graph (k), it can be seen that the positive rating is significantly more than the negative rating. If our classifier results are all positive, the accuracy will be very high at this time. It is not appropriate to use the accuracy to judge our algorithm. Then for our SRR model, due to the strong subjectivity of our model, the score has a relatively large error compared with the quality of the product in reality.