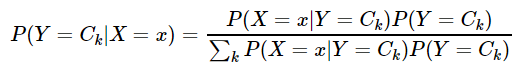
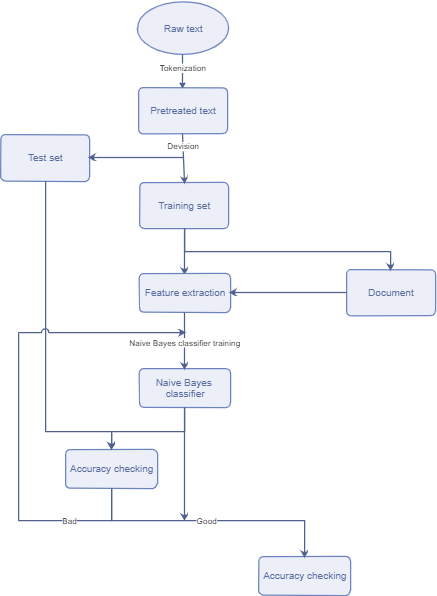
3.2 Quantitative processing of comments

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. 

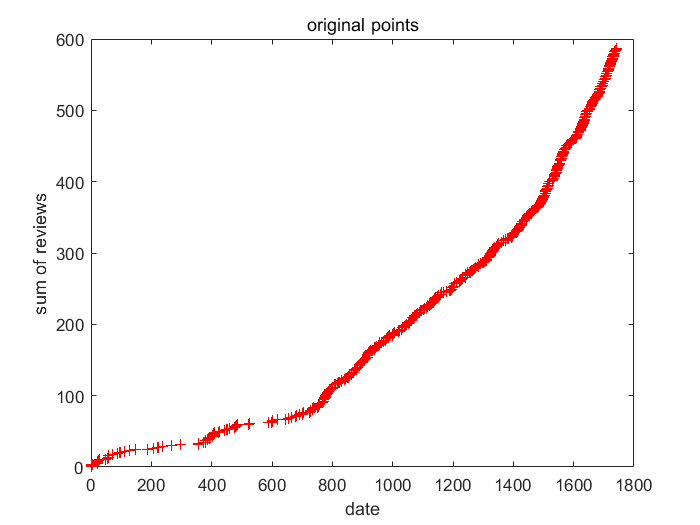
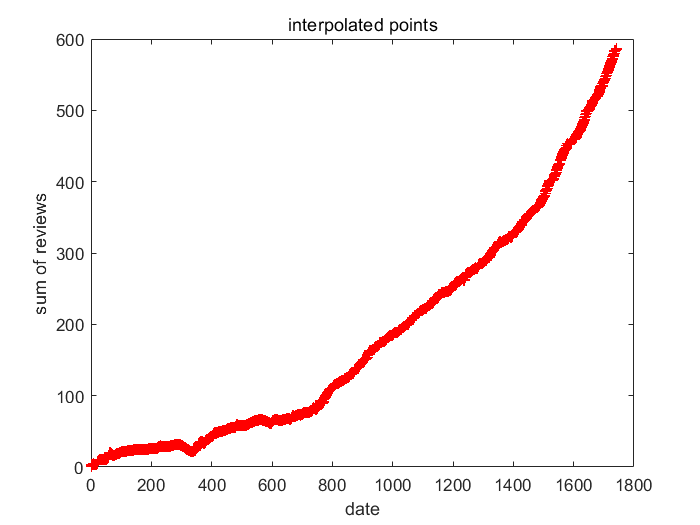
The basic principle is as follows.



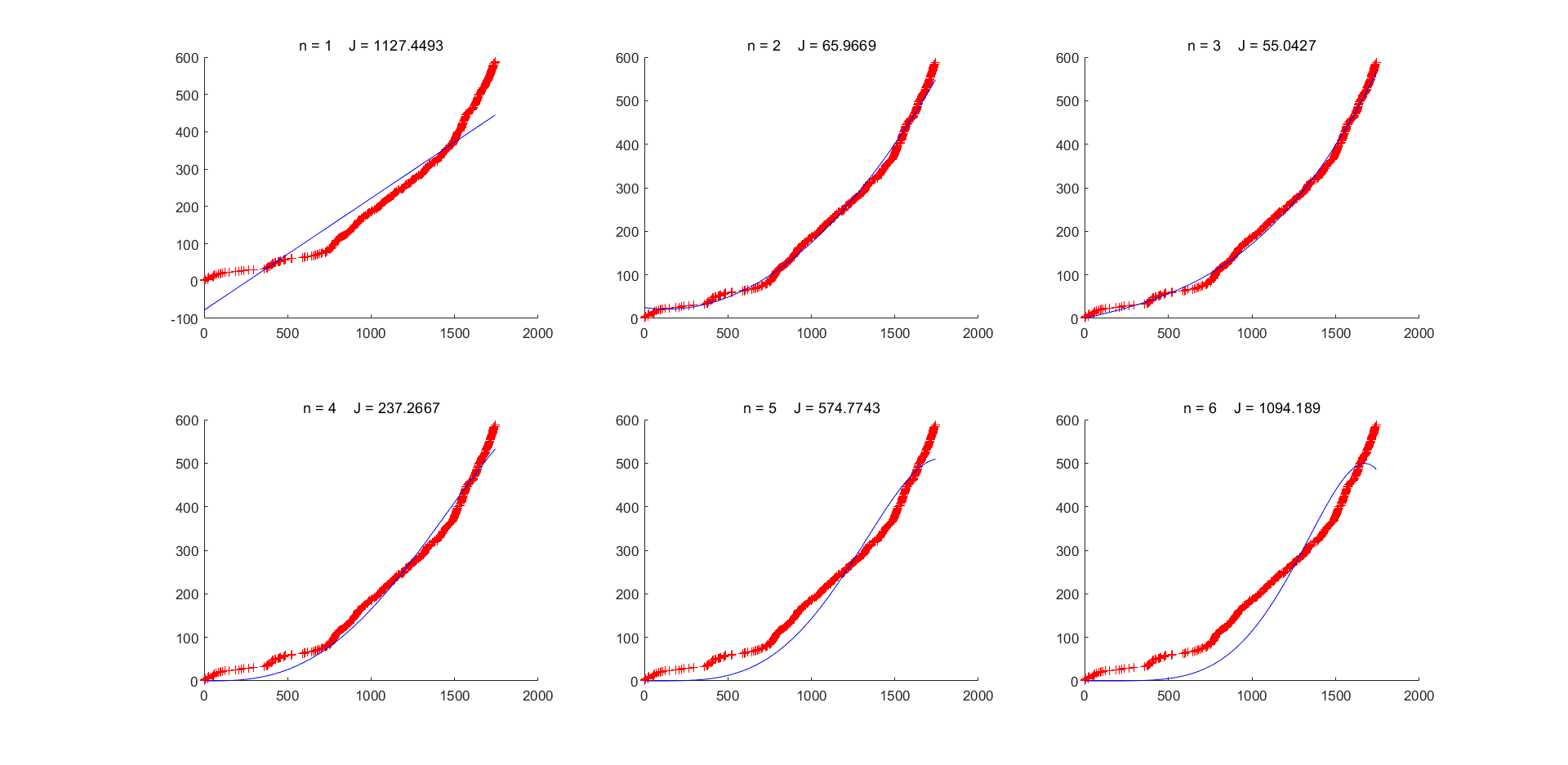
3.2.2 Comment information indicators established

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.



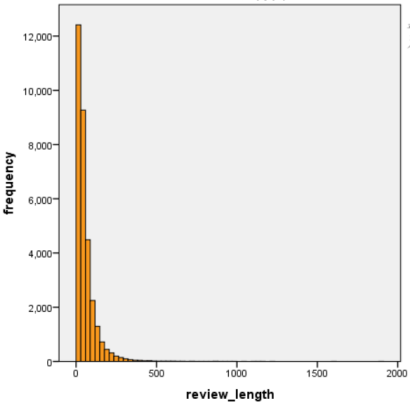
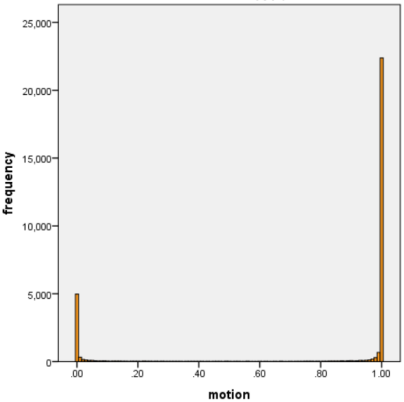
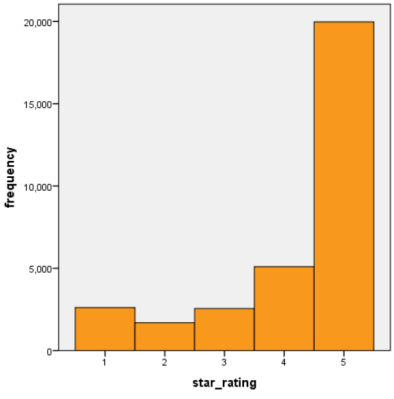
At last we plan to use linear regression to fit the curve. Then we choose the curve with the smallest J(θ).



3.3 Correlation analysis between variables

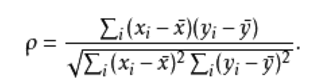
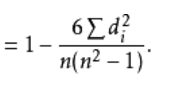
Set star\_rating to X1, sentiment to X2, review\_length to X3, helpful\_vates to X4, total\_votes to X5, density of review is X6.

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.