Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2001159

A Product Evaluation System Promoting Sales Based on Data Insight

Summary

After a series of data processing and modeling, our team finally gets a reputation-based sales plan for microwaves, pacifiers and hair dryers to help Sunshine Company improve its market competitiveness.

First, we do the data-preprocessing, which includes data redundancy processing, word reduction, and index quantification. At the same time, for the convenience of research, we quantify two specific indexes: verified-purchase and vine. Also, in order to achieve the rankings from the text-based reviews, we use **TF-IDF** (Term Frequency–Inverse Document Frequency) and **LSA** (Latent Semantic Analysis) methods to deeply mine the customer satisfaction information contained in them and successfully classify the customer reviews.

Second, we build the Product Reputation Analysis Model based on the two dimensions of customer satisfaction and review credibility. When measuring customer satisfaction level, we examine the star ratings and review ratings, and finally determine the CSI (Customer Satisfaction Index). When measuring the credibility of customer reviews, we use **AHP**(Analytic Hierarchy Process) and **TOPSIS**(Technique for Order Preference by Similarity to an Ideal Solution) methods to weigh the various indicators on customers and reviews, which finally yield the RCI (Review Credibility Index). On this basis, we establish the relationship between the CSI and RCI, and respectively obtain the brand's PRI (Product Reputation Index) of the microwave, pacifier, and hair dryer. Based on the results of this part of the research, we are able to propose to the marketing director of the most attractive features of the product and recommend the most popular product brands.

Third, on the basis of the PRI, in order to explore the relationship between the reputation of different products and changes in time, we combined the **ARIMA** model in **Time Series** to fully analyze and reasonably predict the mean of different products. The accuracy of prediction was improved by establishing a regression model. By analyzing and solving long-term trends and cyclical trends, we successfully established a traditional time series forecasting model with appropriate fitting degrees.

Fourth, Spearman's rank correlation coefficient is also applied to measure the impact of specific star ratings on customer reviews before and after certain time node by month. At the same time, we find that although there is a certain correlation between product ratings and reviews, the connection is not so close.

Finally, we change the numeral value of the parameter CSI and PRI to examine the sensitivity of our Product Reputation Analysis Model. The result shows that our model is robust.

Keywords: Customer Satisfaction;Review Credibility;Product Reputation;TF-IDF;Time Series

Contents

1 Introduction

1.1 Background

In recent years, with the development of the economic globalization, Internet technology has developed rapidly. Thanks to the continuous development and improvement of the Internet, the market size of the world's online retail market has expanded rapidly.This has led to more and more e-commerce platforms pouring into the market.

Internet Retailer's 2019 U.S. Top 500 Report [1], which ranks North America's leading online retailers by web sales, shows that the retailers ranked Nos. 401-500 this year grew their collective web revenue by 24.3% in 2018 over 2017, faster than the 20.0% growth of Amazon, and well above the 14.1%.year-over-year e-commerce growth in North America.

Compared with the traditional offline market, a significant advantage of the online market supported by the Internet is the platform's evaluation system. Based on user ratings and reviews, users can quickly learn about product information so they can make smarter purchasing decisions. The rating and review system greatly reduces the information asymmetry between customers and sellers, and is an important way to improve user product satisfaction.

Take Amazon as an example, its star ratings and reviews provide customers with an opportunity to rate and review purchases, thus help other customers to make wiser purchasing decisions.

Rank	Retailer	Country	Merchant Type	Merchandise Category
1	Amazon.com	U.S.	Web Only	Mass Merchant
2	JD.com	China	Web Only	Mass Merchant
3	Suning Commerce Group	China	Retail Chain	Mass Merchant
4	Apple	U.S.	Consumer Brand Manufacturer	Consumer Electronics
5	Walmart	U.S.	Retail Chain	Mass Merchant
6	Dell Technologies	U.S.	Consumer Brand Manufacturer	Consumer Electronics
7	Vipshop Holdings	China	Web Only	Mass Merchant
8	Otto Group	Germany	Catalog/Call Center	Mass Merchant
9	Gome Electrical Appliances	China	Retail Chain	Mass Merchant
10	Macy's	U.S.	Retail Chain	Apparel/Accessories

Figure 1: Global E-commerce rankings

1.2 Problem Restatement

As discussed above, product reviews have important reference value for other users' purchase decisions. At the same time, it is of great significance to formulate the online sales strategy of the company's market, because the user's evaluation of the product is closely related to the user satisfaction of the product. Our goal is to provide a reference solution for the marketing director of Sunshine Company, so that the company's online sales of products are more competitive in the market. At the same time, based on user ratings and reviews, the design of the three products was improved and optimized to better meet market needs.

To achieve the above goals, we summarize the following issues that need to be addressed in the paper.

Task1 According to the provided product evaluation data, text mining is performed on the text-based user reviews to determine the criteria for judging the text-based reviews.

Task2 Based on user attributes and review attributes, establish a scientific user satisfaction model, and determine the product features that attract the most attention from users.

Task3 Establish a model of product satisfaction over time, determine the relationship between the reputation of a particular product and time, and predict future development trends.

Task4 Analyze the relevance of product reviews before and after time nodes.

Task5 Analyze the correlation between product ratings and reviews.

2 Symbols and definitions

Symbols	Definitions
PRI	Product Reputation Index
CSI	Customer Satisfaction Index
RCI	Review Credibility Index
X_t	Time series
s_r	Star rating
r_r	Review rating
i_v	Index Vector
t_i	The i^{th} word
d_j	The j^{th} review
$n_{i,j}$	Number of occurrences of t_i in d_j
D	The number of reviews in corpus
ϕ, θ	Auto regression coefficient

3 Assumptions

Due to the limited other data given on the products, which cannot accurately characterize the user's buying preferences, we make some basic assumptions to support our model. These simplistic assumptions will be used in our paper and can be complemented by more reliable data.

- **The overall persuasiveness of product reviews is positively related to the number of reviews.**
  In the data pre-processing, in order to make the data more representative, we deleted some of the product data with few comments and did not include these products in our discussion.
- **We do not consider the situation where users who purchase a certain product do not write reviews.**
  Due to the limited data, we can only study the satisfaction of the users who write reviews, and do not consider the product satisfaction of users who purchase but do not rate.

- **We do not take into account the vicious competition of merchants and the manual operation of the Amazon website.**
  In real life, there are cases where some bad merchants maliciously give their competitors a low score. We do not consider these actions that violate market morals. At the same time, we don't consider modifying or editing user reviews on Amazon.com.

4 Data preprocessing

The object of data mining is a large variety of data collected from the real world. Due to the diversity, uncertainty, and complexity of real production and real life, as well as scientific research, the original data we collected is relatively large and scattered. They do not meet the specifications and standards required by mining algorithms for knowledge acquisition research. Therefore, we need to preprocess the data to meet the needs of subsequent data mining.

Product Category	Number of review data
Microwave	1615
Hair-dryer	11470
Pacifier	18937

Table 1: Number of original data

4.1 Data redundancy processing

By observing the data given, we find that in the same category (microwave oven, hair dryer, baby pacifier), the number of user reviews for certain products is too small, and a limited number of user reviews are not objective and authentic, which is not conducive to the sales formed by research Strategy. Based on this assumption, we clean the data for the number of user reviews for different types of products. Firstly delete the user evaluation data with less comments, and then use SPSS to use K-S normality test method to clean the data.

K-S is used as the normality test to determine whether the data sequence meets the normal distribution by comparing the data series with the standard normal distribution. By comparing and detecting the progressive significance $\boldsymbol{sig}$ value, the $\boldsymbol{sig} > \mathbf{0.05}$, which means that the data distribution you want to test is not significantly different from the normal distribution, that is, the data belongs to the normal distribution. Through the above steps, we have eliminated some product data of microwave, hair dryers, and baby pacifiers to make the data more in line with the normal distribution. Take the microwave as an example, the data is shown in the figure below.

After our observation of the data, we found that the data values of marketplace and product-category of the three types of products are consistent internally, so for the convenience of processing, delete the above two attributes in the table.

4.2 Word Reduction Method

Although the product data lacking objectivity has been deleted, the review data at this time cannot be used directly, such as "good, good, good"or"Very nice, very nice". These sentences have a lot of repeated corpora, network homologies and abbreviated words and require Word Reduction Method to remove words. Using **Word Reduction Method** to remove words can deal with continuous repeated parts in the corpus.

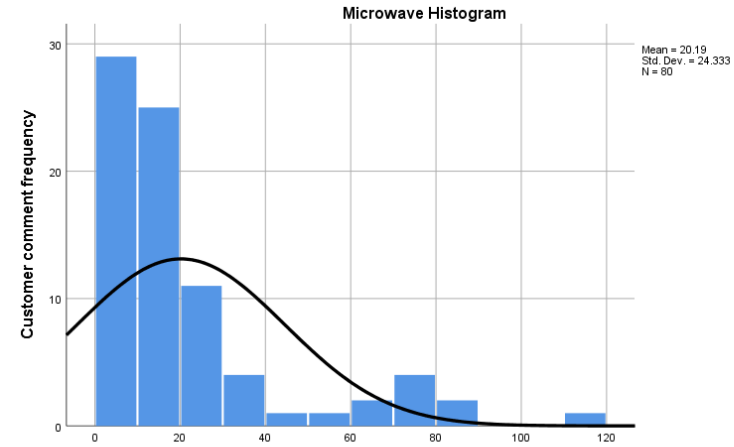


Figure 2: Microwave histogram with normal curve

After observing customer reviews, we find that continuous repeated parts are mostly the beginning or end of the sentence. Therefore, to determine whether there are continuous repeated parts in the review, we can establish two international character list. The first list and the second list are placed in turn, and then the relevant placement judgment and reduction rules are established, and they are placed in the first or second list or the reduction judgment is triggered and duplicate parts are removed.

4.3 Index quantification

Because the values of the two indicators of verified-purchase and vine in the data table can only be yes or no, for the convenience of calculating the RCI, we quantify the results into two variables, 0 and 1.If the result is yes, the value is 1; if the result is no, the value is 0.

4.4 Final data results

Product Category	Number of review data
Microwave	1517
Hair-dryer	1115
Pacifier	14004

Table 2: Number of preprocessed data

In summary, we successfully preprocessed the data to be used next, which is of great significance to the establishment of the analysis model afterwards.

5 Text mining based on TF-IDF and LSA

5.1 Quantifying customer reviews

A major difficulty of Task1 is how to convert the text information of user comments to facilitate the analysis and prediction of the following three kinds of products.

To estimate emotional tendency lying behind each review,we use $TF\text{-}IDF$. and LSA to quantify customer reviews. This chapter is devoted to the extraction of customer review

keywords and weighted processing, and finally, according to the weight of the score, the quantification of the reviews of three products.

5.1.1 Step 1: Extraction of review keywords

We use Python to build a rich and complete word list, and then assess the matching degree of each word in each document, a process similar to tagging, to achieve keyword extraction.We collected the key words reflecting the emotional tendency and product characteristics of the user reviews of three kinds of products.See the first appendix for code.

5.1.2 Step 2: Weighting process of review keywords

The $TF\text{-}IDF.$ technique is used to assess the importance of words to specific text of a document or corpus.The more frequently a word appears in a document, the more important it is.At the same time, the importance of words is also related to the frequency of words in corpus. The higher the frequency of words in corpus, the less important they are. [2] For customer review, we rate the reviews based on the keywords extracted in the previous step.

TF in $TF\text{-}IDF$refers to Term Frequency, which is the number of appearances of a given term in the file of customer reviews.Word frequency is usually normalized by dividing word frequency by the total number of words in the review to prevent it being biased towards longer reviews.The formula is as follows:

$$TF_w = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

During the process,some of the words with high frequency may not contribute to the subject while low-frequency words turn out contributory to the subject while low-frequency words turn out contributory to the subject.To solve this problem,we combine IDF,namely Inverse Document Frequency,to help weighting the words.

In this process, the fewer reviews specific word appears in, the better differentiating capacity it holds.The IDF of certain words can be calculated by dividing total number of reviews by the number of reviews they appear in,then taking the logarithm of the result.Considering the circumstances when the dividend is zero, here is our final formula

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1}$$

After that, $TF\text{-}IDF.$ can be calculated using

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i$$

5.1.3 Step 3: Rating customer reviews

Implicit Semantic analysis (LSA) can map reviews from a sparse n-dimensional space to a low-dimensional (k-dimensional) vector space, which is called Latent Semantic Space.

Based on the keyword extraction in the previous step, we excluded the sequential order of words appearing in each review, selected words with emotional tendency and high word frequency.Then we expressed the correlation of each review with the words using the value

of $TF\text{-}IDF$. Next,we vectorized the review before constructing word-review Matrix X and conducting singular value decomposition of the Matrix, which is demonstrated in fomula:

$$X = WSR^T$$

among which W is the eigenvectors of matrix XX^T, R is the eigenvectors of matrix X^TX rows of WS represents Coordinates of words, and the column vector of RS represents Coordinates of reviews.

Then, perform dimension reduction on the SVD-decomposed matrix: select **k** largest singular values and multiply them by W and R to get a new **k**-order approximation of the X matrix. After the original word-review matrix is subjected to dimension reduction processing, the ambiguous part corresponding to the original word vector will be added to the words that are similar to its semantics, while the remaining part will reduce the corresponding ambiguous components and use the latent semantic space between the reduced matrix.

After all this processing,the reviews with large similarity are classified into one category.All reviews are divided into five categories, and quantified using number: 1,2,3,4,5, where a higher score indicates greater customer satisfaction.

The results of our review quantification is shown below.

Figure 3: Hair dryer score chart based on ratings and reviews

5.2 Collecting emphasized product characteristics

Based on the word frequency statistics of the customer reviews we obtained in the previous section, we separately counted the attribute words most frequently mentioned in the reviews of microwave, pacifiers, and hairdryers to describe certain characteristics.To make it clearer, we made a word cloud map. Here are the results for hairdryer.

Figure 4: Word cloud of hair dryer reviews

6 Product Popularity Analysis Model

Overall, we measure the market popularity of the three products through review credibility and user satisfaction. The following is the process of measuring the above two dimensions.

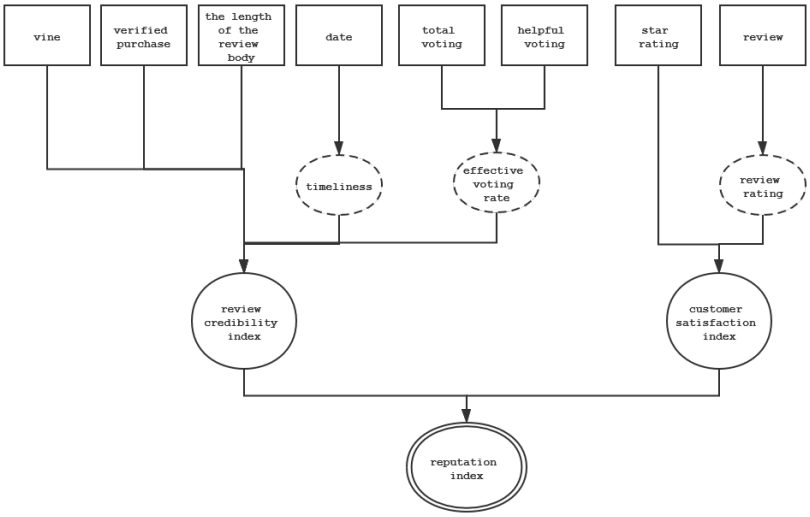


Figure 5: Model building illustration

1) RCI Determination

We make full use of the information provided in the data set, including whether the customer is Amazon Vine member, whether it has been proven to have purchased this product, the effective vote rate of reviews and ratings, the exact length of the reviews, and the date of customer reviews. For these factors, we use two comprehensive evaluation methods: Analytic Hierarchy Process (AHP) and Technology for Order Preference by Similarity to an Ideal Solution (TOPSIS), to clarify the impact of these factors on the credibility of user reviews.So that we can obtain customer satisfaction that is more in line with the actual situation.

- **Amazon Vine member**
  Customers recognized by Amazon as vine members can make more authentic and insightful reviews than other customers, and Amazon gives them more privileges

- **Verified purchase**

Customers who have purchased this product often have more experience and better understanding of product performance.

- **The length of reviews**
  Customers who are serious about using the product and who are deeply impressed tend to have a stronger expression when writing reviews. In order to fully express their ideas, more review length is required.

- **Effective vote rate**
  Visitors on the Amazon website will read reviews made by other customers in order to understand product features or share their feelings before buying. Reviews with higher voting rates will be more universal.

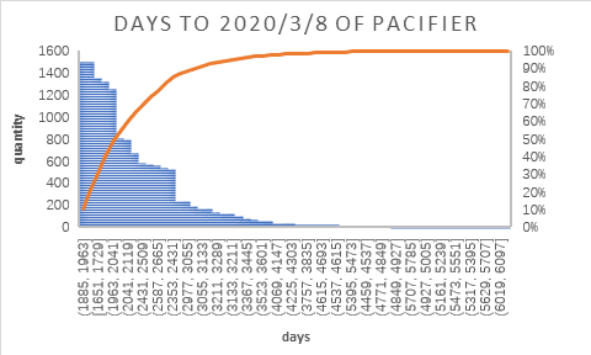
- **Timeliness of customer reviews**
  Online sellers tend to continuously improve and upgrade their products, and customers have different requirements for products as the social living environment changes, so the reviews that are relatively late are more informative.

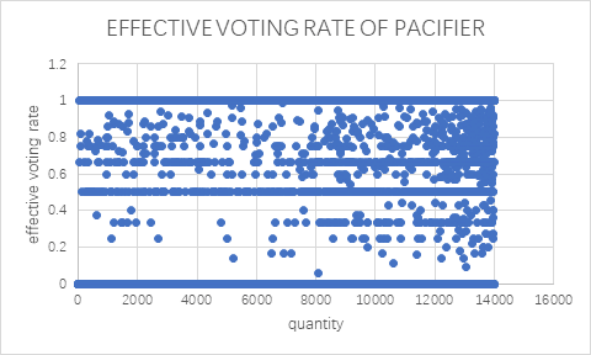
6.1 Index calculation

Because the above-mentioned quantified indicators are too complicated and have large gaps, which can easily lead to extreme credibility, we will use **TOPSIS** (Technique for Order Preference by Similarity to An Ideal Solution) to re-evaluate the ranking.

- TOPSIS is a method for ranking by detecting the distance between the evaluation object and the best and worst solutions. It is best if the evaluation object is closest to the optimal solution while being the farthest from the worst solution; otherwise it is not optimal. The value of each index of the optimal solution reached the optimal value of each evaluation index. The values of each index of the worst solution reached the worst value of each evaluation index.

Due to the large span of some indicators, we have approximated some of them.The following charts are all derived from the pacifier dataset.





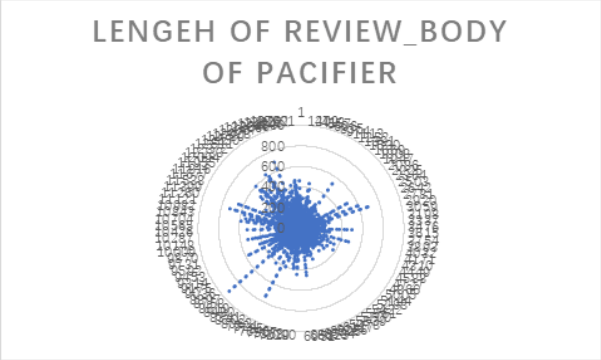


Figure 6: Pacifier review credibility data

Because effective vote rate and the length of views are all positive indexes.We can get the $index_1$ according to the equation

$$index_1 = \frac{x - \min}{\max - \min}$$

. Because the timeliness of customer views is a negative index.We can get the $index_2$ according to the equation

$$index_2 = \frac{\max - x}{\max - \min}$$

Since the vine and the verified-purchase is a 0-1 variable, we combined Amazon big data analysis, expert advice, and our own knowledge of existing dataset to determine the weights of the indexes above.

$$index_3 = \begin{cases} 0.6 \ (vine) \\ 0.4 \ (not \ vine) \end{cases}$$

$$index_4 = \begin{cases} 0.6 \ (verified \ purchase) \\ 0.4 \ (unverfied \ purchase) \end{cases}$$

6.2 Weight determination

Due to the difficulty of quantifying the importance of the indexes, for the weight relationship between the above factors, we use the method of AHP (Analytic Hierarchy Process) to determine the weights.

- Analytic Hierarchy Process decomposes the problem into different constituent factors according to the nature of the problem and the overall goal to be achieved, and aggregates and combines the factors according to different levels according to the interrelationship and membership relationship of the factors to form a multi-level analysis structure model Therefore, the problem is finally reduced to the determination of the relatively important weights of the lowest level (programs, measures, etc. for decision-making) relative to the highest level (overall goals) or the ranking of the relative priorities.

Step1: Establish a hierarchical structure model(see it in figure 5)

Step2: Construct all judgment matrices in each level

We compare each index pair by pair to get judgment matrixes:

$$A = \begin{bmatrix} 1 & 3 & \frac{1}{5} & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{3} & 1 & \frac{1}{2} & \frac{1}{5} & \frac{1}{5} \\ 5 & 2 & 1 & \frac{1}{2} & \frac{1}{2} \\ 2 & 5 & 2 & 1 & \frac{1}{2} \\ 3 & 5 & 2 & 2 & 1 \end{bmatrix}$$

Among them, the row (column) vector is: vine, verified-purchase, length of customer views, timeliness, and effective vote rate.

Step3:Hierarchical single ranking and consistency check

Consistency index:$CI = 0.10329$;

Consistency ratio:$CR = 0.09222$;

Consistency check result: Pass.

Eigenvalue $\lambda = 5.4131$

Get weight vector $w = [0.107190.0606590.218460.256750.35695]$

Step3:Hierarchical single ranking and consistency check

We multiply the index vector and the weight vector to obtain the customer evaluation credibility:

$$PCI = i\_v * w$$

2) CSI Determination Using the evaluation willingness value and customer scoring value obtained by text clustering, we considered the mismatch and malicious badness of rating evaluations such as "five-star bad reviews" and "one-star positive reviews" through customer psychological analysis, expert suggestions and group discussion results Rating factors to determine the following weights:

$$CSI = 0.5 * s\_r + 0.5 * r\_r$$

If the original index value is directly used for analysis, the role of the indicator with a higher value in the comprehensive analysis will be highlighted, and the role of the indicator with a lower value will be relatively weakened. Therefore, in order to ensure the reliability of the results, we standardize customer satisfaction with Z-score:

$$CSI' = \frac{csi - \overline{csi}}{\sigma}$$

The processed data conforms to the standard normal distribution, that is, the mean is 0 and the standard deviation is 1.

In this way, we can establish the relationship between **CSI** and **CRI**.Get **PRI** as a measure of product popularity.

$$PRI = CSI * RCI$$

subsectionProduct reputation rankings Based on the classification and summary of product brands, considering the impact of the number of customer evaluations, that is, fewer reviews cannot objectively reflect the popularity of the product, we sorted the PRI to obtain the following table:

microwave product_tittle	reputation	the number of reviews
whirlpool wmc20005yw countertop microwave, 0.5 cu. ft., white	78	1.811
whirlpool stainless look countertop microwave, 0.5 cu. feet, wmc20005yd	76	1.796
pem31dmww%2d profile spacemaker i%2dcountertop microwave oven %2d white	13	1.783
samsung mc11h6033ct countertop convection microwave with 1.1 cu. ft. capacity, slim fry technology, grilling element, ceramic enamel interior, drop down door, and eco mode in stainless steel	28	1.746
jx7227sfss - deluxe built-in trim kit for 2.2 microwave ovens/ compatible with peb7226sf models/ stainless steel finish	11	1.741
whirlpool wmh31017aw microwave	19	1.739
sharp 1.1-cubic-foot 850-watt over-the-range convection microwaves	31	1.738
sharp rmotda252wrzz microwave turntable motor	19	1.730

pacifier product_tittle	reputation	the number of reviews
oh! baby® 2015, gift basket, special newborn set high quality plush bear and blanket, exclusive pacifier box, burp cloth, mittens, socks, body suit and legging, including gift card	2.541	11
#1/16 serta sheep baby with pacifier	2.532	13
mam air orthodontic pacifier, boy, 6+ months, 2-count	2.520	50
pacifier - wubbanub infant plush pacifier - turtle	2.516	63
mam love & affection orthodontic pacifier, i love mommy, boy, 0-6 months, 2-count	2.506	24
philips avent soothie pacifier, 0-3 months, pink/purple - 6 pack	2.503	11
pullypalz - the interactive pacifier toy	2.502	41
mam monsters orthodontic pacifier, boy, 6+ months, 2-count	2.494	23

hair_dryer product_tittle	reputation	the number of reviews
turbo power turbo 1500 professional hair dryer	2.669	22
remington d3190a damage control ceramic hair dryer, ionic dryer, hair dryer, purple	2.631	23
conair 1875 watt turbo hair dryer and styler, 30.4 ounce	2.630	30
panasonic hair dryer nano care, vivid pink, 1 ounce	2.620	12
remington compact ionic travel hair dryer, (colors vary) d5000	2.610	21
panasonic nano-e nano care hair dryer eh-na95	ac100v 50-60hz (japan model)	2.607	49
mhd professional salon grade 1875w low noise ionic ceramic ac infrared heat hair dryer plus one concentrator and one diffuser black color	2.594	22
hello kitty 1875 watt hair dryer	2.592	15

Figure 7: PRI rankings of three kind of products

7 Time series analysis

In order to explore the relationship between the reputation of different products and changes in time, we combined the ARIMA model in the time series to fully analyze and reasonably predict the mean of different products.

7.1 Model establishment

Let the time series be $\{X_t : t = 0, \pm1, \pm2, \dots, \}$. Usually, we divide X_t into three different items,which is shown below

$$X_t = T_t + S_t + R_t, t = 1, 2, 3, ...$$

where T is the long-term trend term, S is the periodic term, R is the random term.

In this process, the time series discussed below are all wide stationary time series, that is, the following conditions are met:

- $EX_t^2 \leq \infty, t = 0, \pm1, \pm2, ...$
- Mean function is constant, $\mu_t = EX_t = c, t = 0, \pm1, \pm2, ...$
- The self-covariance function $\gamma_{t,s}$ depends only on the time interval (t-s), that is: $\gamma_{t,s} = \gamma_{s,t}, \quad t, s = 0, \pm1, \pm2, ...$

We first perform Auto Regression Moving Average, namely ARMA (p, q) analysis, in which p is an autoregressive order number, which is obtained by the AR (Auto Regression) process and can be expressed as

$$X_t - \phi_1 X_t - \cdots - \phi_p X_t - p = e_t, \quad t = 0, \pm1, \pm2, \dots$$

q is the moving average order number, which is obtained from the MA (Moving Average) process and can be expressed as

$$X_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}, \quad t = 0, \pm1, \pm2, \dots$$

Where $\theta_1, \theta_2, \dots, \theta_q (\theta_q \neq 0)$ are the final regression coefficients of the AR process.

The AR process and the MA process are combined to obtain a sequence that satisfies the following difference equation

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = e_t + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q}$$

with conditions below

$$\left.\begin{aligned} \Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0, |z| \leq 1 \\ \Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0, |z| \leq 1 \\ \Phi(z) \quad and \quad \Theta(z) \ \ coprime \end{aligned}\right\}$$

Next, considering the usual non-stationary factors on the existing stationary model, we will process the non-stationary data into stationary data through difference calculation, and then use the stationary model. This is Autoregressive Integrated Moving Average(ARIMA) model.

7.2 Solution and results

To begin with, we select an index, the microwave's RI (reputation index) as the dependent variable for time series analysis. We separated the seasonal factors in the periodic term to eliminate the possible influence of seasons that may come in the way of moving average calculation and gave the before and after seasonal adjustment comparison.

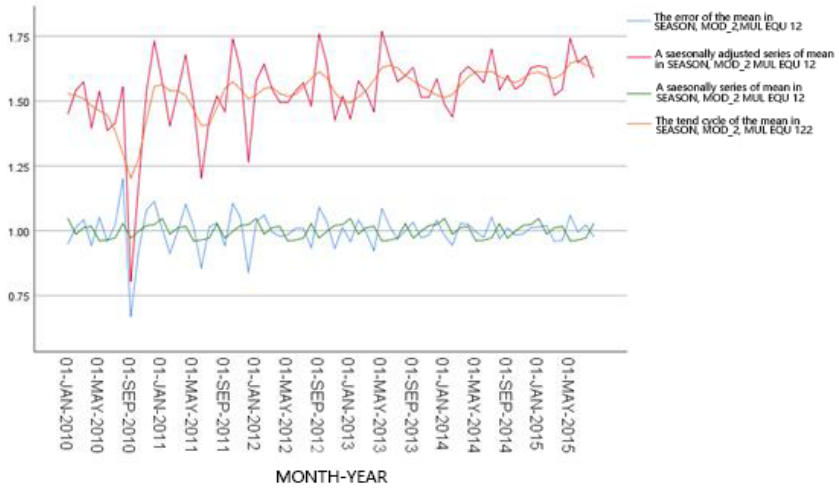


Figure 8: Before and after seasonal adjustment comparison

By performing Autocorrelation Function and Partial Autocorrelation Function on existing sequence, we find the autocorrelation value basically does not exceeds the confidence intervals after the 2nd order hysteresis. So we have $\boldsymbol{p = 2}$. Using the same method, we can obtain that $\boldsymbol{q = 2}$.

Next, we perform the data fitting and Data fitting was performed, and the fitting result of the model was obtained. The fitting degree with the observed values is shown below.

Model fit

Fit Statistic	Percentile 50	75	90	95
Stationary R-squared	.823	.823	.823	.823
R-squared	.277	.277	.277	.277
RMSE	.129	.129	.129	.129
MAPE	6.324	6.324	6.324	6.324
MaxAPE	71.230	71.230	71.230	71.230
MAE	.088	.088	.088	.088
MaxAE	.557	.557	.557	.557
Normalized BIC	-3.905	-3.905	-3.905	-3.905

	Ljung-Box Q(18)		
Stationary R-squared	Statistics	DF	significance;
.823	28.316	15	.54

Figure 9: Model fit and model accuracy

We can see that **Stationary R quare** is relatively large and sig exceeds 0.05 significantly And according to the data before 2015, the prediction for 2016 is as follows

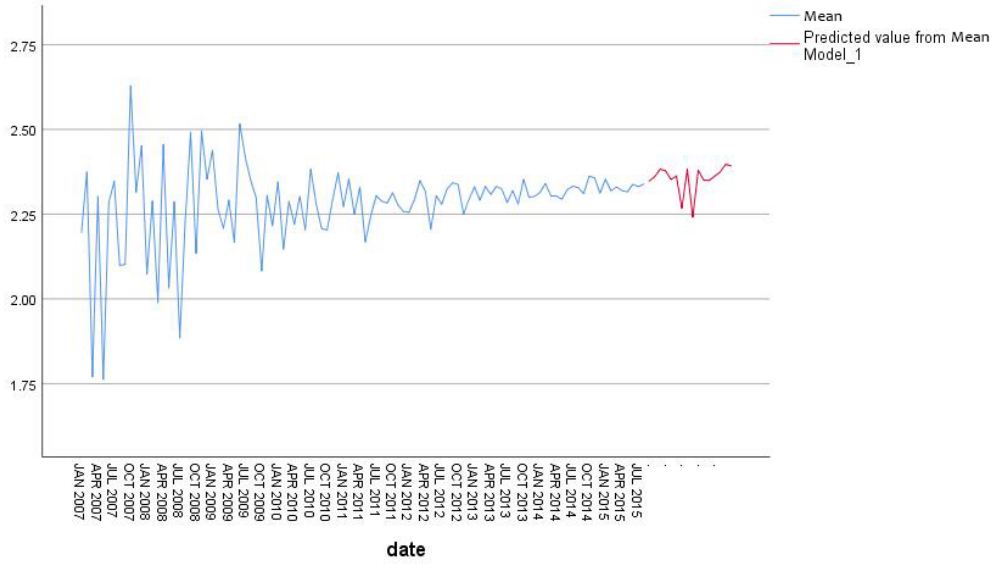


Figure 10: PRI prediction of microwave

After a time series analysis of the reputation of all the microwave ovens, we can see the changes in the reputation of the pacifiers in the entire market. Similarly, we introduced the RI of pacifiers and hair dryers and respectively graph them.

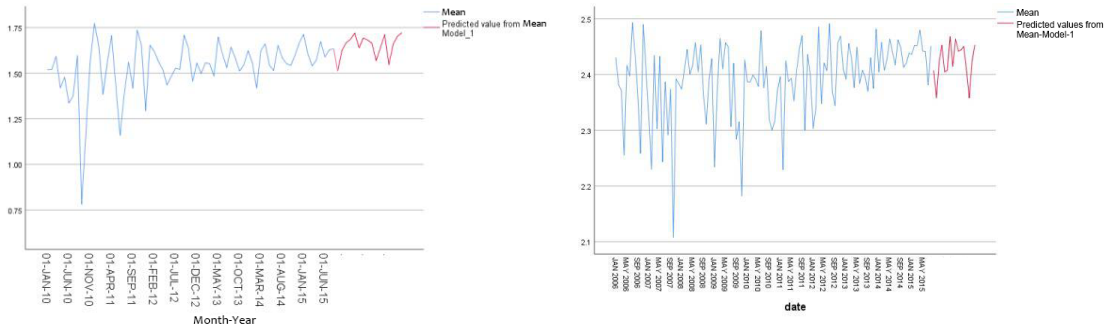


Figure 11: PRI prediction of pacifier and hairdryer

We can apply the above method to specific product models, and we can see the regularity of reputation changes.

7.3 Conclusion

We used the ARIMA time series model, and added differential calculations in the AR and MA processes to improve the accuracy of the prediction. By analyzing and solving long-term trends and cyclical trends, we successfully established the traditional time with appropriate fit. The time series model finally obtained the mean prediction curves of these three types of products until December 2016.

The same method can be used to predict the trend of each specific product's reputation over time for a specific model of a kind of product. Due to the limited number of reviews of a specific product and its uneven time span, we do not perform time series analysis on a single product in this paper.

8 Spearman Rank Correlation

8.1 Introduction to Spearman Rank Correlation

Spearman rank correlation is used to estimate the correlation between two variables. It requires that the observations of the two variables are paired and graded assessment data, or graded data transformed from observations of continuous variables, without having to consider the overall distribution of the two variables and the size of the sample capacity.

Based on the customer ratings and reviews for each type of product we obtained in the previous section, suppose the two variables are X and Y, and the number of elements is both n. The i^{th} $(1 \leq i \leq n)$ values of the two variables are expressed as X_i and Y_i, respectively. After that, we sort X and Y at the same time to get two new sets x and y respectively after ranking the elements, where the element x_i is the rank of X_i in X, and y_i is the rank of Y_i in Y. Then, we subtract the corresponding elements in the sets x and y to obtain a ranking difference set d, where $d_i = x_i - y_i$ $(1 \leq i \leq n)$.

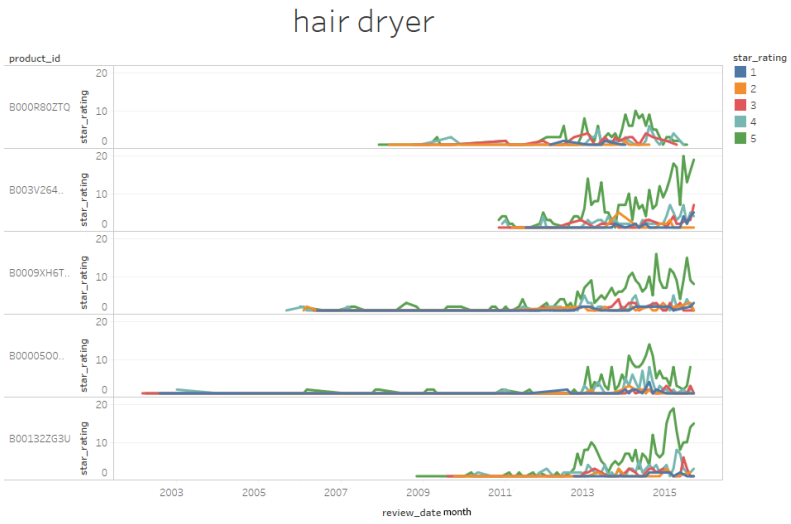
The Spearman rank correlation coefficient between the random variables X and Y is defined as

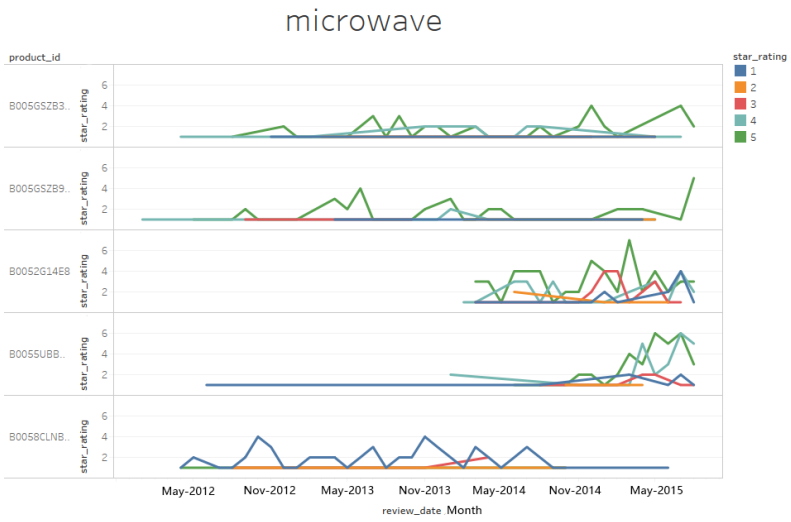
$$\rho = 1 - \frac{6\sum_i^n d_i^2}{n(n^2-1)}$$

8.2 Correlation analysis

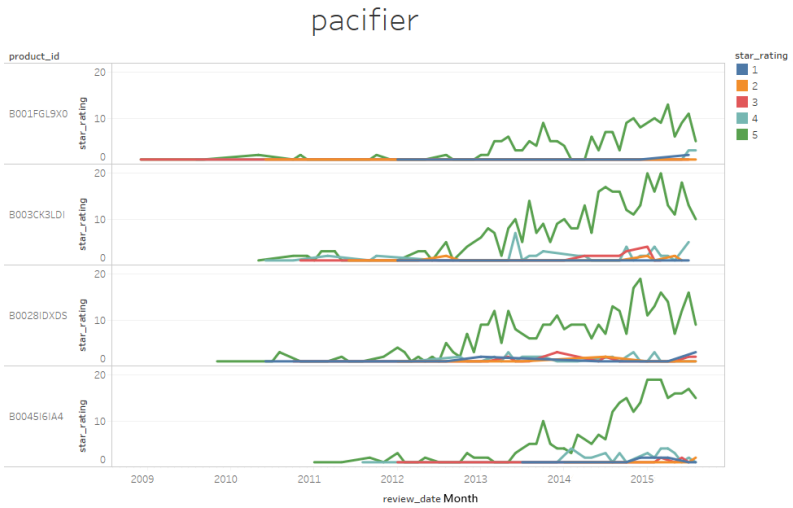
We have quantified reviews as a numeric indicator in the previous text mining section, so the correlation analysis will be performed between the review indicators and customer ratings.

Because the smaller the amount of review for a specific product, the less convincing it is, so for hair dryers, we filtered customer rating data with number of customer reviews > 300; for pacifiers, we filtered > 230 customer rating data; for microwave, we filtered user rating data> 77 customer rating data. In order to dig deep into the relationship between the rating levels before and after certain time node, we divide the rating $1, 2, 3, 4, 5$ stars into five obejive, using time (precision is month) as the horizontal axis and the number of different rating levels as the vertical axis. The data is visualized by drawing a line chart. Because the amount of data before 2012 is too small, the following discussion focuses on the changes after 2012.





Hair dryer and pacifier: After observing the images, we find that for different models of hair dryers and pacifiers, the 2-4 star rating curve fluctuated slightly, while the 1-star and 5-star rating curves fluctuated greatly, with multiple peaks appearing at different time periods. Therefore, we can speculate that users who buy hair dryers and pacifiers are more susceptible to 5-star reviews than other star reviews, and make similar rating decisions.



After observing the images, we find that for different models of microwave, the number of 1-5 star ratings all fluctuated. Therefore, we can guess that users who purchase microwave are not easily affected by the ratings of other customers, but are affected by the factors in the product itself.

Based on the conjectures made based on the images above, we also used Spearman correlation coefficients to analyze the correlation between the time and the number of five rating levels respectively using one specific product out of hair dryers, microwaves, and pacifiers.the data with the most customer reviews. We use SPSS to get the correlation coefficient Rs.

star-rating	Rs-h	Rs-p	Rs-m
1	0.847	0.678	0.382
2	-0.105	0.231	-0.485
3	0.328	0.305	0.256
4	0.543	0.582	0.344
5	0.981	0.943	0.534

Table 3: Time-related correlation coefficient

8.3 Conclusion

According to the table above, we can get generally consistent conclusions from previous image speculation. Extreme reviews (1 and 5 stars) of users who purchase pacifiers and hair dryers are easily affected by previous extreme reviews of users, while for microwave, reviews are less affected by previous users and are influenced by the pros and cons of the product itself.

9 Sensitivity Analysis

When constructing the Product Popularity Analysis Model, we used AHP to calculate the weights between various indicators with some subjective judgments. When changing weight vector w slightly, the RCI and PRI of the model change as follows: From the table above, we can find that the influence of weight vector is not big, which we can bear.

	CSI	PRI	rate of CSI change	rate of PRI change
microwave	0.460846	0.239457	0.21%	8.61%
hairdryer	0.394805	0.949224	0.19%	6.52%
pacifier	0.37505	0.859443	0.15%	6.83%

Figure 12: Sensitivity analysis results

10 Strengths and weaknesses

10.1 Strengths

- **Data processing**
  Our modeling confronts a large amount of data in the first place. Data processing, especially text-based review processing, is very significant for our research. Through this step, we have optimized our data to a great extent. This way, we can solve our problems more efficiently.

- **Accuracy and stability**
  We adopt the Arima model to predict the product reputation. This is a model that highlights time variables and can reflect the relationship with time accurately.

- **Full consideration**
  Combining the time series Arima model with winter linear and seasonal exponential smooth prediction models, fully analyzes and reasonably predicts the mean of different products.

- **Comprehensiveness**
  We make full use of various indicators, also the identified customer satisfaction model has high credibility.

- **Intuitive and intelligibility**
  We use a wealth of charts to present our research results.

10.2 Weaknesses

- **Subjectivity**
  In the customer satisfaction model, we are more subjective in determining the weight of index, the key component of customer credibility, which can lead to small deviations in our model results.

- **Limited use of data**
  In the time series model, we have used data from 2010 onwards, and the data we have has a small time span and limited ability to predict the future.

11 Conclusion

For **task 1**, we adopted TF-IDF to extract keywords and determine the similarity of the reviews. We clustered all reviews by Latent semantic analysis, and labeled them with 1 to 5 according to the clustering results. High values indicate good feedback, low values indicate poor feedback.

For **task 2**, we treat each customer as an object on the basis of quantifying text reviews. By analyzing the feasibility of the customer reviews and the timeliness of the reviews, we have established a more scientific customer satisfaction model. Data mining of text reviews yields "high-frequency words", which gives the product features that attract the most attention from customers. Specifically, the most popular product features of hair dryers are heating speed, price, and power. The product features of pacifiers are quality, appearance, and price. The most popular product features of microwave are quality, operation difficulty, Features.

For **task 3**, the models in the time series are combined to fully analyze and forecast the PRI of different products, and establish corresponding forecasting models. Finally, the mean forecast curves of these three products are obtained through December 2016. Overall, it is expected that the market competitiveness of hair dryers will be the largest in 2016, with pacifiers second and microwave third. However, the average of the three products in 2016 will have some fluctuations.

For **task 4**, we found that the reviews of hair dryers have a cyclical period of about three months; the reviews of pacifiers dropped to very low in March and May 2016, and other times were relatively stable. ; Microwave reviews will reach two peaks in January and August 2016, other times more volatile.

For **task 5**, we found that short-term seditious reviews (such as a series of one-star negative reviews and blindly positive reviews) will affect future customer reviews in the same direction, but in the long run, a particular product of customer reviews don't make a big difference because of specific emotions.

12 A letter to the marketing director of Sunshine Company

Dear sir or madam

In response to your questions regarding the online sales strategy and potentially important design features, we are writing to inform you of our work. We are confident to say that our team has proposed a product reputation index to quantify reputation of different brands in the three products in the market. A mathematical model is built to help predict the trend in the reputation over time. The optimal sales strategy is determined by sorting of products, which is according to PRI.

More specifically, the reputation index is constructed to represent the customer satisfaction index as well as review credibility index. We comprehensively adopted two evaluation methods (AHP and TOPSIS) to determine the weight of each index, then combined into the PRI. (Recommendation form at the end of the letter)

Next, in order to explore the relationship between product reputation and time, we adopted ARIMA models to fit the trend of PRI over time from 2010 to 2015 for each product, and successfully predicted the product reputation in the next period. Products with potential market value or lack of future market competitiveness were found separately.

One of the advantages of our research is the effective mining of customer-based text reviews. By extracting customer evaluation keywords, we adopt TF-IDF to calculate review similarity, and then using LSA for cluster analysis and objective rating. The combination of the two methods can scientifically establish indicators for evaluating the pros and cons of specific product reviews.

At the same time, we utilized the text data of the reviews obtained above to calculate the degree of attention of the three product customers to their specific features. We screened the top five

characteristics of the three categories of products. (Sorting list at the end of the letter)

In conclusion, our model successfully produced an online sales strategy of three categories of products. Since the model as well as the evaluation approach is fully data-motivated with no arbitrary criterion included, it is rather adaptable for determining sales strategies for the future. Based on research results, you can combine the profitability of the product and finally get the best solution for the company's online sales.

We have a strong belief that our model can effectively promote the Sunshine Companys sales strategy and help your team improve the product functions in accordance to the market needs.

	product_title
microwave	whirlpool wmc20005yw countertop microwave, 0.5 cu. ft., white
	whirlpool stainless look countertop microwave, 0.5 cu. feet, wmc20005yd
	pem31dmww%2d profile spacemaker il%2dcountertop microwave oven %2d white
	samsung mc11h6033ct countertop convection microwave with 1.1 cu. ft. capacity, slim fry technology, grilling element, ceramic enamel interior, drop down door, and eco mode in stainless steel
	jx7227sfss - deluxe built-in trim kit for 2.2 microwave ovens/ compatible with peb7226sf models/ stainless steel finish
pacifier	oh! baby® 2015, gift basket, special newborn set high quality plush bear and blanket, exclusive pacifier box, burp cloth, mittens, socks, body suit and legging. including gift card
	#1/16 serta sheep baby with pacifier
	mam air orthodontic pacifier, boy, 6+ months, 2-count
	pacifier - wubbanub infant plush pacifier - turtle
	mam love & affection orthodontic pacifier, i love mommy, boy, 0-6 months, 2-count
hair_dryer	turbo power turbo 1500 professional hair dryer
	remington d3190a damage control ceramic hair dryer, ionic dryer, hair dryer, purple
	conair 1875 watt turbo hair dryer and styler, 30.4 ounce
	remington compact ionic travel hair dryer, (colors vary) d5000
	panasonic nano-e nano care hair dryer eh-na95 | ac100v 50-60hz (japan model)

Ranking	Microwave	Hairdryer	Pacifier
1	durability	price	size
2	warranty	durability	appearance
3	price	weight	price
4	fuction	power	material
5	material	function	durability

Sincerely yours,

Team 2001159

References

[1] 2019 U.S. Top 500 Report.

https://www.digitalcommerce360.com/product/top-500/

[2] Duan W, Cao Q, Gan Q. Investigating Determinants of Voting for the "Helpfulness" of Online Consumer Reviews: A Text Mining Approach[J]. 2011, 50(2):511-521.

[3] Copeland J A. LSA Oscillator-Diode Theory[J]. 1967, 38:3096-3101.

[4] Jian G , Nailian H U , Xiang C , et al. ROCKBURST TENDENCY PREDICTION BASED ON AHP-TOPSIS EVALUATION MODEL[J]. Chinese Journal of Rock Mechanics and Engineering, 2014, 33(7):1442-1448.

[5] B K Nelson. Statistical methodology: V. Time series analysis using autoregressive integrated moving average (ARIMA) models[J]. Academic Emergency Medicine Official Journal of the Society for Academic Emergency Medicine, 1998, 5(7):739-744.

Appendices

Here are TF-IDF module written using $Python$ language.

Input Python source:

```
import nltk
jieba.load_userdict("newDict.txt")
from sklearn.feature_extraction.text import TfidfVectorizer
# print(dir(TfidfVectorizer))

def cut(txt_name1, txt_name2):
with open(txt_name1, 'r') as f1:
txt = f1.read()
txt_encode = txt.encode('utf-8')
txt_cut = nltk.cut(txt_encode)
result = ' '.join(txt_cut)
# print(result)
with open(txt_name2, 'w') as f2:
f2.write(result)
f1.close()
f2.close()

cut('nlp_test0.txt', 'nlp_test0_0.txt')
cut('nlp_test1.txt', 'nlp_test1_1.txt')

stopWords_dic = open('stop_words.txt', 'r')
stopWords_content = stopWords_dic.read()
stopWords_list = stopWords_content.splitlines()
stopWords_dic.close()

with open('nlp_test0_0.txt', 'r') as f3:
res3 = f3.read()
with open('nlp_test1_1.txt', 'r') as f4:
res4 = f4.read()

corpus = [res3, res4]
# print(corpus)
vector = TfidfVectorizer(stop_words=stopWords_list)
tf_idf = vector.fit_transform(corpus)
# print(tf_idf)

word_list = vector.get_feature_names()
weight_list = tf_idf.toarray()
# result1 = ''.join(word_list)
# result2 = ''.join(weight_list)
# print(result1, result2)
# with open('words_list.txt', 'w') as f3:
#    f3.write(result)
```

```
for i in range(len(weight_list)):
print("-------the tf-idf weight of ", i+1, "th word ------")
for j in range(len(word_list)):
print(word_list[j], weight_list[i][j])
```