Problem Chosen	2022 MCM/ICM Summary Sheet	Team Control Number
C		2019057552

Dig Reviews and Discover Wealth

Summary

With the technoledge and economy developing, Electronic Commerce gets more and more popular. There is no doubt that E-commerce companies led by Amazon have brought great changes to people's life and shopping styles. In order to help the sunshine company better prepare for online sales, we analyzed the existing review data of three types of Amazon products and obtained a series of conclusions, which can help the sunshine company make scientific decisions.

For problem 1, Firstly, we analyze the data, including preprocessing, visualization and descriptive statistics. Then by establishing the **LDA** theme model, we extract the theme of each comment, summarize the main content of the comment, grasp the customer comment tendency and make scientific decisions. In general, there are more five-star comments on the three products, it is necessary to pay attention to the problems reflected by some customers.

For problem 2(a), utilizing the theme of effective comments, we build five classic examples through a self-built corpus, the corresponding scores are 0.1, 0.3, 0.5, 0.7 and 0.9. we also use **Word2Vec** to analyze the similarity between comments and classic examples. The satisfaction rate S of each comment is the product of the score of classic examples with the largest similarity and the largest similarity. Synthesizing satisfaction rate, star rating and helpful_ ratio, we use the K-means algorithm to divide the customer's comments on products into four categories: "terrible","bad", "good" and "great". The advantages of products can be found through the "great" comments, while the "terrible" comments are often the most valuable, which is the key to improving products and influencing decision-making. See the text for detailed results.

For problem 2(b), in order to reflect the increase or decrease of product reputation, we define reputation rate R, complete the construction of R evaluation model through the **EWM-TOPSIS** model, and take the final score as R. According to the daily date data, and the importance to time measurement, the product review data can be regarded as a time series. A product reputation prediction model is constructed through the **ARIMA** model to quantitatively analyze whether the reputation increases or decreases.

For problem 2(c), since there is no specific measure of product success or failure, we use One Class **SVM** model to distinguish normal products from abnormal products (potentially successful or failed products). It is found that the reputation rate of successful products is higher than that of normal products, while the reputation rate of failed products is lower than that of normal products.

In response to problem 2(d), we need to explore the correlation between comment stars and comments. Using the Pearson correlation coefficient, we found that the higher the star of the product, the more likely customers are to give high praise, and the more high praise. On the contrary, the lower the star of the product, the more likely customers are to give bad reviews, and the more bad reviews will be.

Aiming at problem 2(e), we utilize the **Apriori** algorithm to mine frequent itemsets, and it is found that there is a strong correlation between stars and specific comments. For example, pacifier, high ratings are strongly associated with "perfect" and "nice", while low ratings are often accompanied by "break" and "bad".

Based on the analysis of the whole paper, we wrote a letter to the marketing director of the sunshine company, summarized our analysis and results, and suggested that sunshine company should pay attention to customer evaluation while improving product quality, and adjust the sales strategy in time through customer evaluation in order to obtain more profits.

Keywords: LDA, Word2vec, K-Means, EWM-TOPSIS, ARIMA, SVM, Apriori

Contents

1 Introduction

1.1 Problem Background

While online marketplace is becoming more and more popular, the vast majority of people like shopping online. As the same time, everyone can give some text-messages(review) and star rating(1∼5) for products they buy, these can provide some useful informations for potential customers. If this review provide unuseful information for potential customers, he/she can star rating for this review(helpfulness rating), these information is the most important part for online marketplace.

Sunshine Company plan to introduce three new products in the online marketplace, and he wants to know about the above three indicators of microwave oven, a baby pacfier, and a hair dryer. They plan to hire a team to analyze the data given, and give some message about these:

- inform their online sales strategy.
- identify potentially important design features that would enhance product desirability.

1.2 Restatements of the Problem

Considering the background information and restricted conditions identified in the problems statement, we need to solve the following problems.

1. According to the three product data set provieded, give the
2. According the analysis result of item 1, you should solve these problems:
   (a) Identify data measures based on ratings and reviews that are most informative for Sunshine Company to track.
   (b) Identify and discuss time-based measures and patterns within each data set that might suggest that a product's reputation is increasing or decreasing in the online marketplace.
   (c) Determine combinations of text-based measure(s) and ratings-based measures that best indicate a potentially successful or failing product.
   (d) Do specific star ratings incite more reviews? For example, are customers more likely to write some type of review after seeing a series of low star ratings?
   (e) Are specific quality descriptors of text-based reviews such as 'enthusiastic', 'disappointed', and others, strongly associated with rating levels?

1.3 Our Work

Our workflow flow chart is shown in Figure 1, which systematically shows the method to solve the problem and the relationship between each step.

In Question 1, given three data sets are given, data analysis is carried out first. Comments with VINE N and verified_purchase N are removed, review stars are visualized, and helpful_ratio is constructed by sigmoid function. Description statistics for review stars and Helpful_ratio. The LDA topic model is constructed to extract the topic of each comment, and the result is displayed as a word cloud.

In Question 2(a), the theme of effective comments is firstly used to construct 5 classic example sentences through self-built corpus, and the corresponding scores are set as 0.1, 0.3, 0.5, 0.7 and 0.9. Word2vec is used to analyze the similarity between the comments and the classic example sentences. The satisfaction rate of each comment is defined as the product of the score of the classic example with the greatest similarity and the maximum similarity. Then combining satisfaction rate, star rating

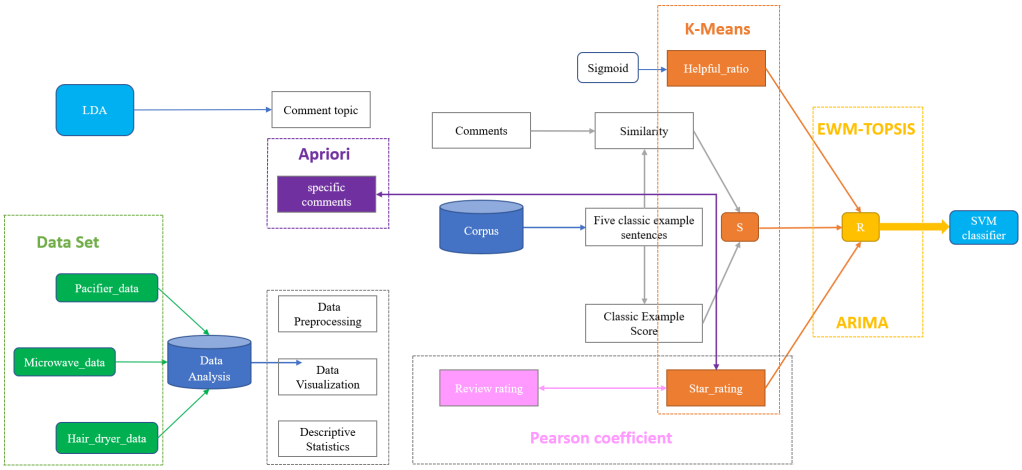


Figure 1: Workflow

and helpful_ratio, k-means algorithm was used to classify customers' comments on products.

For question 2(b), the reputation rate R is defined, and the evaluation model of R is constructed through EWM-Topsis model, and the final score is R. Product review data has date attributes and is regarded as a time series. A prediction model of product reputation rate is constructed through ARIMA model for quantitative analysis.

For question 2(c), the One Class SVM model was first used to distinguish normal products from abnormal products (potentially successful or failed products), and then the SVM model was trained to give the judgment of successful or failed products.

For question 2(d), it is necessary to explore the correlation between review stars and reviews. Pearson correlation coefficient is used to explore the correlation between stars and the number of reviews, as well as between stars and review levels.

For question 2(e), the Apriori algorithm is used to mine frequent item sets and work out strong association rules between stars and specific comment words.

Finally, we wrote a letter to the Marketing director of Sunshine Company, summarizing our analysis and results, and giving our own reasonable suggestions.

2 Reasonable Assumptions

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

- **Assumption 1**: Data preprocessing is effective.

  Justification: Data preprocessing has eliminated outliers as much as possible, and ensuring the effectiveness of preprocessing is to ensure the accuracy of the model and the reliability of the results.

- **Assumption 2**: Customers read reviews of this product carefully before purchasing.

  Justification: Make sure that product reviews have an impact on whether or not customers buy the product, as well as giving review stars and specific reviews.

- **Assumption 3**: Data files are trusted.

Justification: Malicious comments from customers or other spam will affect the model and results, but we have no way of knowing the malicious comments in the data file.

3 Notations

Symblo	Description
helpful_ratio	Likelihood of this review being a helpful review
β_i	Topic i
$W_{j,n}$	The nth word in document j
$Z_{j,n}$	The topic to which the nth word in document j belongs
η	Hyperparameters for topic distributions
α	Hyperparameters for document topic distribution
S	Satisfaction rate
R	Reputation rate

Note: The above table does not list all symbols, the meaning of the symbol is subject to the use of the symbol.

4 Model 1: LDA Topic Extraction and Data Analysis

4.1 Data Preprocessing

Since there are no vacant values in the data file, they can be ignored.Other data processing, such as data standardization and feature coding, will be mentioned in the specific model. Below, only abnormal data will be judged and processed.

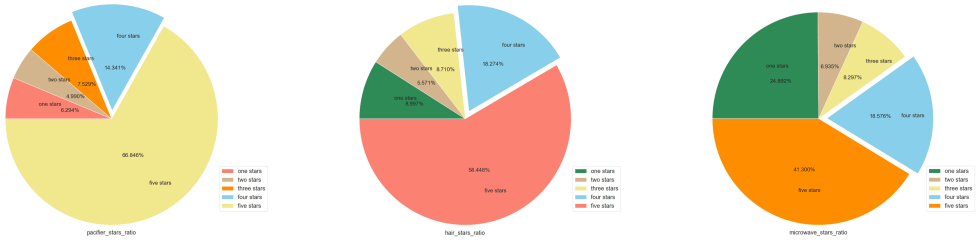
For each sample comment, marketplace, product_id, and other irrelevant attributes can be deleted. Verified_purchase for Y indicates that the customer has purchased the product for close to the original price, and Vine for Y indicates that the customer has gained trust in the Amazon community due to the accuracy and insight of the review.Vine members are trustworthy, but we are suspicious of non-Vine members.Comments with Vine N and verified_purchase N lack credibility if the reviewer is not a Vine member and did not purchase the product at close to its original price or at all (since reviews may lack authenticity if they are heavily discounted).Table 1 gives an example of an untrustworthy review that exists in all three data sets.

Table 1: Examples of untrustworthy comments

marketplace	customer_id	...	vine	verified_purchase	...	review_date
US	39431051	...	N	N	...	8/31/2015

4.2 Data Visualization

Since the actual situation of the possible data of the three products was different, we visualized the three products and analyzed the results respectively.First, visualize the review stars, as shown in Figure 2.



(a) pacifier_star_ratio (b) hair_dryer_star_ratio (c) microwave_star_ratio

Figure 2: The proportion of five stars in the star_rating of each product

It can be seen from Figure 2 that five-star reviews are always the most among the three products. The star distribution of pacifier and Hair_dryer is roughly the same, while the two-star reviews are the least, indicating that the quality of the two products is generally satisfactory, but it still needs to be strengthened.However, one-star reviews in microwave are close to 24.892%, second only to five-star reviews, indicating that there may be quality problems and other aspects, and the specific content of one-star reviews should be paid attention to.

To take advantage of comments' useful voting, we define a variable that measures how helpful a comment is: $helpful\_Ratio$, which represents the likelihood that the comment will be helpful.Intuitively, the expression of help rate is $helpful\_votes/total\_votes$, but considering that $total_votes$ cannot be calculated when it is 0, and comments cannot simply be considered to be unhelpful, **sigmoid** function $y = \frac{1}{1+e^x}$ is used to solve this problem.The sigmoid function image is shown in Figure 3.

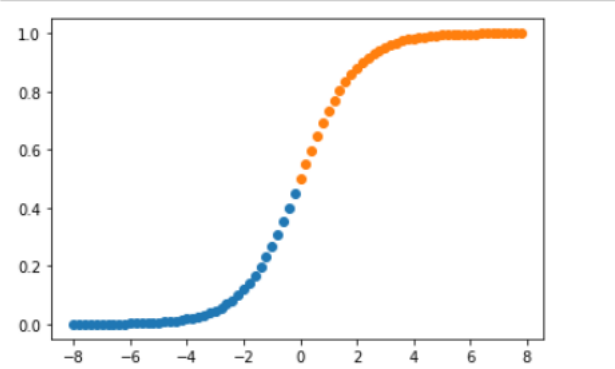


Figure 3: Image of sigmoid function

Make helpful_ratio=y,,x=helpful_votes-(total_votes-helpful_votes), and you get .After calculation, the helpful_ratio is greater than 0.5 as helpful comments, less than 0.5 as unhelpful comments, equal to 0.5 not sure whether helpful.Figure 4 shows the large distribution of helpful comments.

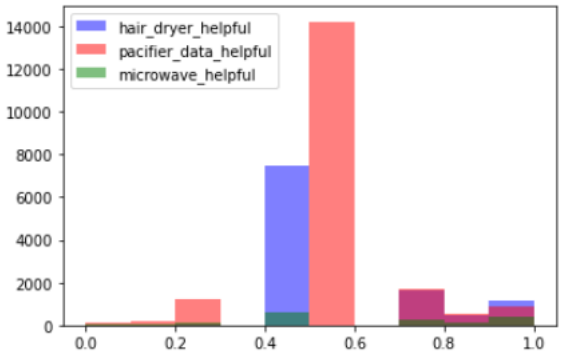


Figure 4: Distribution of helpful comments

It can be found from Figure 4 that most of the help rates of pacifier and hair_dryer are concentrated around 0.5, and the help rate is basically not too low. As the microwave data volume only has 1432 comments after data cleaning, basically at 0.5 and 0.8-1.0, indicating that there are many helpful comments.

4.3 Descriptive Statistics

The following describes the statistics of stars and helpful_ratio. Microwave is taken as an example. It can be seen from Table 2 that the star rating is between 1 and 5, divided into 5 levels, with

Table 2: Star description statistics

Statistics	Numerical value
mean	3.45
std	1.65
min	1
25th percentile	1
50th percentile	4
75th percentile	5
max	5

Table 3: Helpful_ratio Description statistics

Statistics	Numerical value
mean	0.72
std	0.21
min	0.5
25th percentile	0.5
50th percentile	0.73
75th percentile	0.95
max	1

an average value of 3.45, indicating that the average microwave star rating is between 3 and 4, and microwave evaluation can be considered as general only by the star rating. Table 3 shows that the mean helpful_ratio of Microwave is 0.72, and the score is 0.5-1, indicating that the data cleaning effect is very good, and there are almost no comments with a help rate lower than 0.5.

4.4 Topic Extraction Model Based on LDA

Model Construction

Latent Dirichlet Allocation(LDA)[1] topic model proposed by Blei in 2003, is a three-layer Bayesian probability model extended on the probabilistic implicit semantic index (pLSI), is a document generation probability model.The LDA model consists of three layers: term, topic and document. The basic idea is to treat a document as a mixture of its implied topics.Documents to topics follow a polynomial distribution, and topics to words follow a polynomial distribution.The purpose is to identify topics, that is, the document vocabulary matrix is divided into document topic matrix and topic vocabulary matrix.LDA[2] has had a huge impact in the field of natural language processing

and statistical machine learning, and has quickly become one of the most popular probabilistic text modeling techniques in machine learning.

Suppose in the document set, the number of documents is A, the number of words in the document is B, and there are C topics in the document set. β_i represents the i-th topic, ζ_i represents the topic distribution of document J, $W_{j,z}$ and $Z_{i,j}$ represent the n-th word in document J and its topic respectively, η and α represent the hyperparameter of topic distribution and the hyperparameter of document topic distribution respectively.Figure 5 shows the graphical structure of the LDA model.

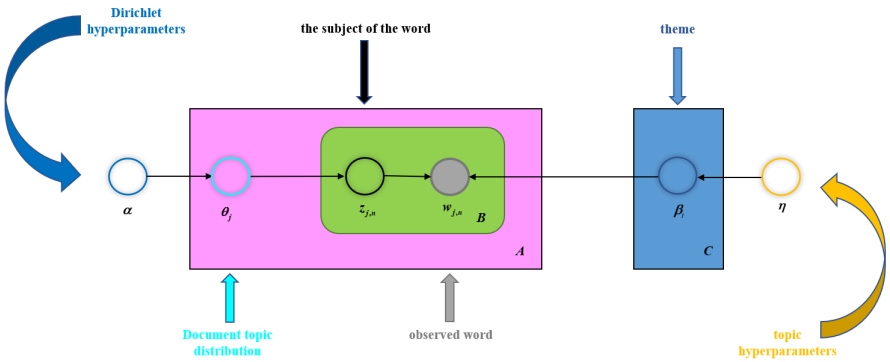


Figure 5: Graphical structure of LDA model

The joint probability density of implied variables and observed variables in the LDA model is:

$$\begin{aligned} p(\theta,\beta,w,z \mid \eta,a) &= p(\theta \mid \alpha)p(\beta \mid \eta)p(w \mid \beta,z)p(z \mid \theta) \\ &= \prod_{-1}^{D} p\left(\theta_j \mid \alpha\right) \prod_{\zeta=1}^{c} p\left(\beta_1 \mid \eta\right) \prod_{n=1} p\left(w_{1,n} \mid z_{j,n},\beta_z\right) p\left(z_{j,n} \mid \theta_j \right. \end{aligned} \tag{1}$$

Among them, subject to dirichlet distribution has $P(\Theta_i|\alpha)$ and $P(\beta_i|\eta)$, but $P(w_{j,n}|z(j,n),\beta_{i,n})$ and $P(z_{j,n}|\Theta_j)$ obey the multinomial distribution.

Variational reasoning and Gibbs sampling are usually used for parameter estimation of LDA model.Both are approximate estimation methods, each has its own advantages and disadvantages, depending on the situation to choose appropriate methods.In general, Gibbs sampling is easy to realize but inefficient, while variational reasoning is complicated but efficient.

Results and Analysis

Based on the LDA topic analysis algorithm, we extracted the subject words of the comments, used the wordcloud library of python to display the wordcloud, and gave an explanation.

As can be seen from the Figure 6, most of the customer comments on the three products are positive, with Fivestars accounting for a large proportion. From the word cloud, it can be seen that the favorable comments mainly contain words such as love, great, cute and nice, while the negative comments are related to words such as problem, broke and problem.

(a) Wordcloud of pacifier (b) Wordcloud of hair_dryer (c) Wordcloud of microwave

Figure 6: Subject word cloud extracted by LDA

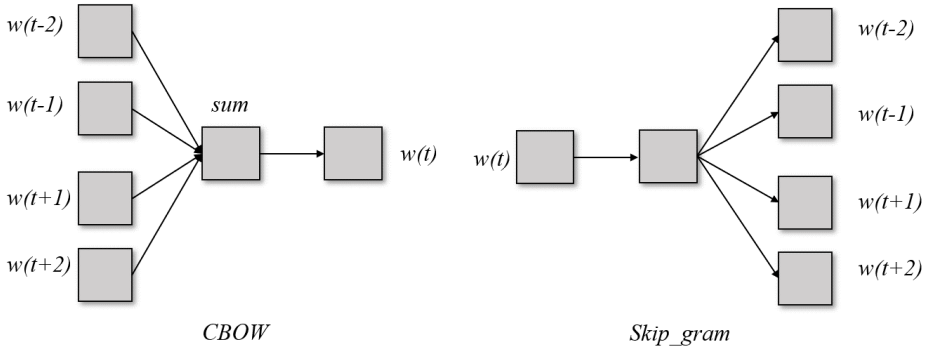
5 Model 2: Screening Valuable Reviews Based on K-Means And Word2vec

5.1 Model Construction

First, we have used LDA to get the topic of each valid comment. Next, we use the self-built corpus of comments to construct 5 classic example sentences, and the corresponding scores are 0.1, 0.3, 0.5, 0.7 and 0.9. Similarity analysis is conducted between all valid comments and these 5 classic example sentences.The score of each comment is the product of the score of the classic example with the greatest similarity and the maximum similarity. We call the score satisfaction rate and mark it as S.

Word2vec Algorithm

Word2vec[3] is an implementation of the model proposed by Mikolov et al. It can train word vector quickly and effectively.It contains two training models, CBOW and Skip_gram respectively, and their schematic diagrams are shown in FIG 7(a) and 7(a). It can be seen from Figure 7 that both



(a) Wordcloud of pacifier (b) Wordcloud of hair_dryer

Figure 7: CBOW and Skip_gram structure framework

models have three layers, namely, input, hidden and output layers.If a word is used as input to predict the context, then the model is called Skip-Gram model; if the context of a word is used as input to predict the word itself, then it is the CBOW model.

Table 4: word2vec word vector training framework

model	CBOW	Skip_gram
Hierachy Softmax	CBOW+HS	Skip_gram+HS
Negative Sampling	CBOW+NS	Skip_gram+NS

Table 4 shows two optimization methods of Word2VEc to improve training efficiency and four word vector training frameworks combined with two training models.

K-Means Clustering Algorithm

K-means[4] algorithm is an unsupervised learning and clustering algorithm based on partition. Generally, Euclidean distance is used as an indicator to measure the similarity between data objects. The similarity is inversely proportional to the distance between data objects, and the larger the similarity, the smaller the distance.Figure 8 is the k-means algorithm flow chart. We perform K-means clustering according to satisfaction, Helpful_ratio and star rating.

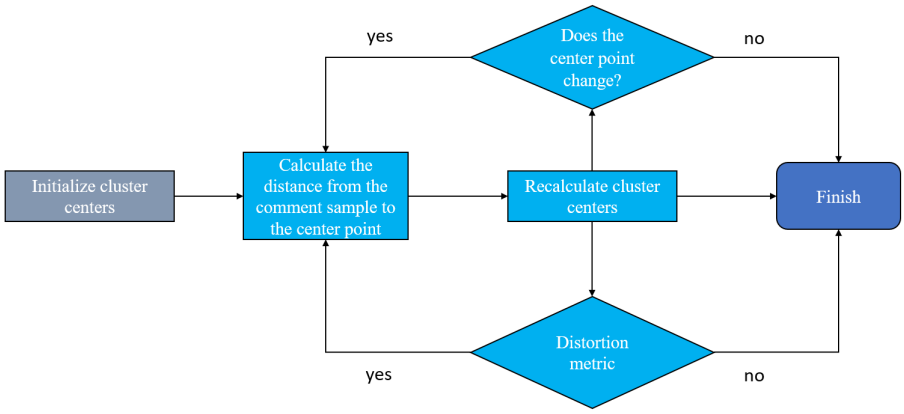


Figure 8: K-means flow chart

5.2 Results and Analysis

The following analysis only takes Hair_Dryer as an example, and the analysis of other two products is similar.Table 5 shows satisfaction rates for some of the comments.

Table 5: Satisfaction rate for some reviews of hair_dryer

marketplace	customer_id	review_id		S
US	51995766	R230LCPQDOFJJZ	...	0.87
US	39431051	R21NN9ONVZITI0	...	0.79
...	...	...	...	...
US	9924936	R3N0F2FKJOMGKK	...	0.95

Then, satisfaction, Helpful_ratio and stars are used for K-means clustering. The number of clusters is set to 4, and the given numbers 1,2,3 and 4 respectively represent very good, good, poor and very poor.Sort out the clustering results, add category numbers to the data shown in Table 5, and show the clustering effect of satisfaction, Helpful_ratio and star rating together, as shown in Table 6.

Table 6: some reviews of hair_dryer

review_id	...	star_rating	helpful_ratio	S	category
R230LCPQDOFJJZ	...	5	0.5	0.87	2
R21NN9ONVZITI0	...	1	0.5	0.79	4
...	...	...	...	...	
R3N0F2FKJOMGKK	...	5	0.95	0.95	1

It is not difficult to find from Table 6 that reviews with low star rating and low helpful_ratio are usually in category 3 and 4, that is, "poor" and "very poor" reviews, indicating that customers are dissatisfied with the product and give comments on business trips; while reviews with high star rating and high helpful_ratio are basically in category 1 and 2, indicating that customers are satisfied with the product and give favorable comments.Similarly, reviews in category 2 tend to be favorable, which we define as "good", and reviews in category 3 tend to be negative, which we define as "poor".This is consistent with the actual situation, which shows that the calculation of model and satisfaction rate is reasonable.

6 Reputation Prediction Based on EWM-TOPSIS and ARIMA

6.1 Model Construction

EWM-TOPSIS Algorithm

According to the requirements of the title, we need to find time-based measures and patterns to reflect the increase or decrease of product reputation, so we need to define reputation rate R, which is constructed by EWM-Topsis algorithm, that is, R is the comprehensive score, and then we need to use time series model ARIMA to predict reputation.

TOPSIS[5] is a comprehensive evaluation method to determine the advantages and disadvantages of evaluation objects. It makes full use of known initial data information and the results can well reflect the differences between evaluation schemes, which is widely applied in many fields.Considering that the weight coefficients of each index may not be equal, entropy weight method (EWM) is introduced to determine the weight, which is more objective than AHP.

The steps of ewM-TopSIS model construction in this paper are as follows:

Step 1: Construct decision matrix A. Since there are three indicators including satisfaction, help rate and star rating, m=3 and n depends on specific products.

$$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{2}$$

Step 2: The decision matrix index is forward.The extremely small, intermediate and interval indicators are transformed into extremely large indicators, which do not need to be transformed because they are all extremely large indicators.

Step 3: standardize the decision matrix, set the standardized matrix as B, and each element in B:

$$b_{1j} = \frac{1}{a_{ij}} \sqrt{\sum_{i=1}^{n} a_{ij}} \tag{3}$$

If there is A negative number in matrix A, the Min-Max normalization method is used:

$$b_{ij} = \frac{a_{ij} - \min\left\{x_1, \ldots, x_{\pi j}\right\}}{\max\left\{x_1, \ldots, x_{\pi j}\right\} - \min\left\{x_1, \ldots, x_{\pi y}\right\}} \tag{4}$$

Step 4: The index weight is determined based on EWM method, and the non-negative matrix is obtained by **Step 3**, and the probability matrix P is calculated, where the value of each element is:

$$p_{nj} = \frac{b_{ij}}{\sum_{i=1}^{n} b_{ij}} \tag{5}$$

and $\sum_{i=1}^{n} b_{i,j} = 1$. Then calculate the information entropy of the j-th indicator E_j:

$$E_j = -\frac{1}{\mathrm{In}(n)} \sum_{-1}^{n} p_{1j} \mathrm{In}\left(p_{ij}\right) \tag{6}$$

then calculate the information utility V_j:

$$V_j = 1 - E_j \tag{7}$$

Finally, the weight of each index is obtained by normalization W_j:

$$W = \frac{V_j}{\sum_{j=1}^{m} V_j} (\mathrm{j} = 1, 2, \ldots, \mathrm{m}) \tag{8}$$

Step 5: Calculate the score and normalize the treatment.Define maximum value A+ and minimum value B−:

$$\begin{aligned} B^- &= \left(B_1^-, B_2^+, \ldots, B_m^+\right) = \left(\max\left\{b_{11}, b_{21}, \ldots, b_{n1}\right\}, \ldots, \max\left\{b_{1m}, b_{2\pi}, \ldots, b_{nm}\right\}\right) \\ B^- &= \left(B_1^-, B_2^-, \ldots, B_m^-\right) = \left(\min\left\{b_{11}, b_{21}, \ldots, b_{n1}\right\}, \ldots, \min\left\{b_{1m}, b_{2m}, \ldots, b_{nm}\right\}\right) \end{aligned} \tag{9}$$

at the same time define the distance between the ith comment object and the maximum value $\mathrm{D_i}+$ and the minimum value $\mathrm{D_i}-$ as follows:

$$D_i^* = \sqrt{\sum_{-1}^{m} W_j \left(b_{ij} - B_j^+\right)^2}, \quad D_i^- = \sqrt{\sum_{-1}^{m} W_j \left(b_{ij} - B_j^-\right)^2} \tag{10}$$

Then the unnormalized score of the ith comment object S_i calculated as:

$$S_i = \frac{D_1^-}{D_i^+ + D_i^-} \tag{11}$$

Finally, the normalization can obtain $S_1 = S_1 / \sum_{i=1}^{n} S_1$ represents the final score of the i-th comment.

ARIMA Time Series Analysis

ARIMA model, short for autoregrestive moving average model, is a famous time series prediction method proposed by Box and Jenkins in the early 1970s[7]. ARIMA(p,d,q), p represents the order of autoregression, d represents the order of difference, and q represents the order of moving average.The model can be expressed as:

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t \tag{12}$$

L is a lag operator and d is a positive integer.

The difference expression is as follows:

$$\begin{cases} \phi(B)\nabla^d X_t = \theta(B)\varepsilon_t, \\ E(\varepsilon_t) = 0, \mathrm{Var}(\varepsilon_t) = \sigma^2, E(\varepsilon_t\varepsilon_s) = 0, S \neq 1, \\ E(X_s X_t) = 0, \nabla S < t. \end{cases} \tag{13}$$

Time series model can be transformed into stationary time series model by difference operation.

$$\nabla^d X_1 = \sum_{i=1}^{d}(-1)^d C_d^i X_{t-1}, \text{among them, } C_d^j = \frac{d!}{i!(d-i)!} \tag{14}$$

For d in the model, when $d=0$, ARIMA(p,d,q) model is actually ARIMA(p,q) model; When $p=0$, ARIMA(p,d,q) model is denoted as IMA(p,d); When $q=0$, ARIMA(p,d,q) model is simply denoted as ARI(p,d).When d=1,p=q=0, ARIMA(p,d,q) model is called random walk model, namely:

6.2 Results and Analysis

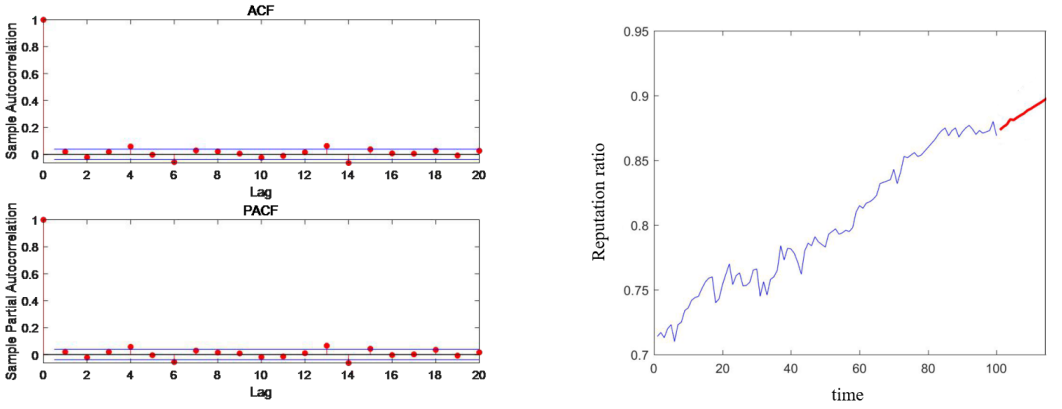
After the program is run, the reputation rates of pacifier, hair_dryer and microwave are obtained respectively. We only take pacifier as an example, and the results obtained are shown in Table5.

Table 7: Reputation rate for pacifier data

marketplace	customer_id	review_id		R
US	40626522	R1A3ZUBR8TSAKY	...	0.00006046
US	15312194	RG9XY3EKPUCL1	...	0.00003773
...	...	...	...	...
US	20849759	rzhlaecz5ztnc	...	0

It can be seen from the three tables that R is small because of the large amount of data, which will result in a small R after normalization.Even after data pretreatment, there will still be comments with a reputation rate of 0, which indicates that it is not reliable to only look at the star rating, but also refer to the satisfaction rate and help rate.

Next, SPSS's expert modeler was used for time series analysis, as shown in Figure 9. In order to show the increasing and decreasing trend of comment reputation rate, we selected the comment with product_ID B013RF851A from pacifier dataset as an example.It can be seen from FIG 9(b) that the reputation rate of this review keeps increasing trend, which requires differential processing. The first-order difference is determined by testing and significance test, that is, d =1. From the ACF and PACF images in FIG 9(a), it can be found that both are trailing.So the model is ARIMA(3,1,3).FIG 9(b) also shows that the reputation rate of the sample review has an increasing trend.



(a) ACF and PACF (b) Sequence diagram and prediction results

Figure 9: Time series analysis of a single comment sample

7 Product Classification Model Based on SVM

7.1 Model Construction

Since the division of failed products and successful products is not clear, we need to use existing data and reputation rate to establish a SVM model. In order to facilitate the following description, we name the products as normal products and abnormal products, and use SVM to identify potential successful or failed products.

SVM[6] in the SLT (StatisticalLearningTheory) based on a machine learning method, has been widely used in pattern classification, function estimation, regression analysis and other fields.It contains three ideas: optimal hyperplane technique, soft interval and inner product kernel function.

The SVM Classifier

For the dichotomous problem, SVM separates two different classes by training a hyperplane, namely successful products and failed products in this problem.Assuming that the hyperplane can accurately classify training samples, the hyperplane can be described by the following equation:

$$W^T X + b = 0 \tag{15}$$

Where $W = (W_1, W_2, \ldots, W_d)$ represents the normal vector, b represents the displacement term, and the partition of the hyperplane is determined by W and b. Give the training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m), y$ $\{+1, -1\}\}$, For $(x_i, y_i) \in D$, there are:

$$\begin{cases} w^T x + b \geq +1, y_1 = +1 \\ w^\top x + b \leq -1, y_1 = -1 \end{cases} \tag{16}$$

We define a product as a successful product $y_i = +1$ and a failed product $y_i = -1$.We hope to find some of the largest interval partitioning hyperplane, which under the constraints problem:

$$\begin{aligned} &\min_{w;b} \frac{1}{2}\|w\|^2 \\ &st.(x_i + b) \geq 1, i = 1, 2, \ldots, m. \end{aligned} \tag{17}$$

Usually uses dual problem solving and SMO methods to deal with dual problems. If some sample division errors can be allowed, "soft interval" is used to introduce relaxation variable solution.

One Class SVM

Since the conditions for judging successful or failed products are not given in this paper, OneClassSVM needs to be introduced to distinguish successful products from failed products. The goal of OneClassSVM is to separate data points from the origin as much as possible in the feature graph space. Firstly, the data points are projected onto the feature space using gaussian kernel, and then the quadratic programming problem is solved to separate the projected data points from the origin.

7.2 Results and Analysis

OneClassSVM was first used to determine outliers of reputation rates that represent the presence of potentially successful or failed products.LSVM is then trained using OneClassSVM as a potential success or failure threshold for the product.Taking the Pacifier data set as an example, we found the corresponding successful products and failed products.Partial results are shown in Table 8:

Table 8: Potentially successful or failing products

product_id	b00f8nkkzo	B003CK3LDI	B00B7U61RI	b00hnjo5uw
Classification result	success	fail	success	fail
product_id	B00LZKBP2Q	B00PWKC32G	b004fq086g	B00793CZAE
Classification result	success	fail	fail	success

After judging the potential successful or failed products, corresponding to the calculated reputation rate, it is found that the reputation rate of successful products is higher than that of normal products, while the reputation rate of failed products is lower than that of normal products.

8 Correlation Analysis Model Based on Pearson's Coefficient

8.1 Model Construction

Pearson's correlation coefficient, also known as Pearson product moment correlation coefficient and simple correlation coefficient[8], is mainly used to calculate the correlation between two variables, and the value range is [-1,1]. Suppose the existing variables A and B, then the specific calculation formula is as follows:

$$\rho_{AB} = \frac{\mathrm{Cov}(A,B)}{\sigma_A \sigma_B} = \frac{\sum_{-1}^{n}(A-E(A))\,(B_1-E(B))}{\sqrt{\sum_{-1}^{n}(A-E(A))^2}\sqrt{\sum_{-1}^{n}(B_1-E(B))^2}} \tag{18}$$

When r is located at [0.8,1.0], it indicates that variable A is strongly correlated with variable B; when r is located at [0.6,0.8], it is strongly correlated; when r is located at [0.4-0.6], it is moderately correlated; when r is located at [0.2-0.4], it is weakly correlated; when r is located at [0,0.2], it is extremely weakly correlated or no correlated. Pearson correlation coefficient variables need to satisfy the linear relationship and pass the significance test, because this is a linear correlation test method, in addition to the test of normality.

In this article, we will consider the correlation between stars and the number of reviews, as well as the correlation between stars and review categories.

8.2 Results and Analysis

Since the principles are the same, we use hair_dryer as an example, and the statistics of star_rating and the number of comments are shown in Table 9.

Table 9: star_rating and comments

star_rating	number of comments
1	908
2	586
3	929
4	1971
5	6400

The correlation coefficients are shown in Table 10. Due to a large amount of data, we select some comment data to display the star_rating and comment category of some comments.

Table 10: Data display of some comments

review_id	...	S	star_rating	category
B00VRN7SB8	...	0.68	1	4
B00092M2XW	...	0.97	5	1
...	...	...	...	...
B003FBG88E	...	0.84	2	2

Through calculation, the correlation between star_rating and the number of comments is 0.8056, which is in a strong correlation, indicating that the higher the star_rating is, the more the number of comments will generally be. It can be found that 1-star and 2-star reviews are not a linear increasing trend. The correlation coefficient between star_rating and review category is -0.9453, because review rating and star rating are approximately opposite and strongly correlated, indicating that our review rating index is very successful.

To sum up, the higher the star_rating, the more comments will generally be, and the more favorable comments will be. The lower the star_rating, the less the number of comments will be, and the more negative comments will be.

9 Association Rules Mining Model Based on Apriori Algorithm

9.1 Model Construction

Apriori algorithm was proposed by Agrawal et al in 1993 for the shopping basket problem[8]. Its basic idea is to generate candidate item sets first and then select frequent item sets from them. According to ”if an item set is not a frequent item set, then all item sets containing this item set are not frequent item sets”, the frequent item set is first found.Then continue digging for frequent 2 items, and so on until you find all frequent items.Its construction process mainly includes two steps: generation and filtering, as shown in Figure 10.

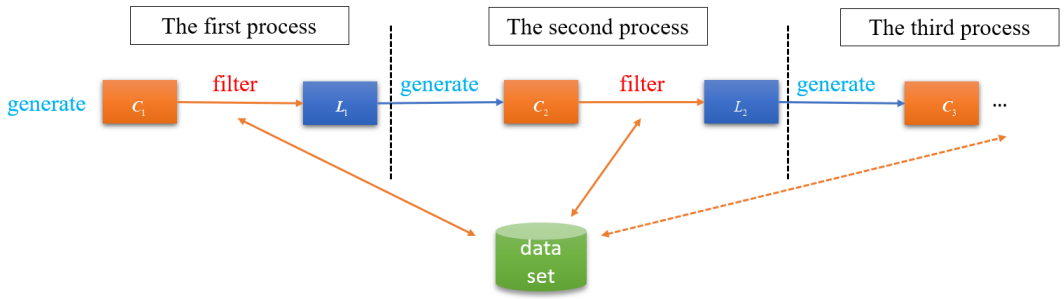


Figure 10: Apriori algorithm schematic diagram

Before using the Apriori algorithm, you need to introduce concepts such as association rules, support, and confidence. Let $X = x_1, x_2, \ldots, x_n$, let A and B be itemsets, and the association rule $A \subset B$ is satisfied, A and B are disjoint. The definition of support is as follows:

$$\text{support}(AB) = P(A \cup B) = \frac{\frac{D_{A \to B}}{D} \times 100\%}{D} \tag{19}$$

D represents the total amount of data, while $N = A \cup B$ is the amount of data containing both A and B, meaning the proportion of data covered by the association rule, that is, the probability of A and B appearing in the total amount of data at the same time.

The definition of confidence is as follows:

$$\text{confidence}\ (AB) = P(B \mid A) = \frac{D_A \to B}{D_A} \times 100\% \tag{20}$$

D_A is the amount of data in A, the probability that the amount of data that contains item set A also contains item set B.

A strong rule is defined to be greater than both the support threshold (min_suppt) and confidence threshold (min_conf). Item sets are called item sets, k-item sets are items containing K items, and frequent K-item sets are denoted as L_k.

In this case, perform the following steps:

Step 1: One-hot encoding of data.

Step 2: Define association rules, support, and confidence.

Step 3: Mining association rules using Apriori.

Step 4: Calculate the specific review words that are most relevant to each star class.

9.2 Results and Analysis

In the case of hair_dryer, high ratings are strongly associated with ”love”, ”great”, etc. while low ratings are often accompanied by ”cost”, ”heavy”. From the strong association rules of the three products, it is found that the higher the star rating, the specific review words tend to be common praise words or praise words of the product. The lower the star rating is, the specific comments are basically bad comments. Some comments directly point out the problems of the product. This kind of bad comments is an important way to improve sales strategy and understand the shortcomings of the product.

10 Model Analysis

10.1 Strengths and Weaknesses

Strengths

1. LDA topic model can analyze text and extract topics well.
2. Combined EWM and TOPSIS, comprehensively evaluated the reputation rate of the product by taking full consideration of various indicators, which effectively solved the one-sidedness of the single index of star rating, help rate and satisfaction rate.
3. Creatively propose normal and abnormal products, use OneClassSVM to get abnormal products, and then distinguish potential successful or failed products.
4. Apriori algorithm can effectively help us obtain the association between specific review words and stars.

Weakness

1. Topics extracted by LDA are not necessarily valid.
2. The results of Apriori algorithm need to be selected manually, and the required results cannot be given according to the needs.

10.2 Sensitivity Analysis

There are two hyperparameters in the LDA model, η and α the hyperparameter representing the topic distribution and the document topic distribution respectively. And based on the following joint probability density:

$$p(\theta,\beta,w,z \mid \eta,a) = \prod_{-1}^{D} p(\theta,\mid \alpha) \prod_{\mid=1}^{C} p\left(\beta_1 \mid \eta\right) \prod_{-1}^{N_H} p\left(w_{j,n} \mid z_{2,n},\beta_2\right) p\left(z_{j,n} \mid \theta_j\right) \tag{21}$$

When η and α is changed, that is, when conditions in conditional probability are changed, the joint probability will change and the solution result will also change. We adjust the original η=0.6 and α=0.3 to η=0.4 and α=0.7 to observe whether the topic extracted from comments has changed significantly, and take satisfaction rate of comments as the judgment standard. Take the comments in the hair_dryer section as an example, the results are shown in the following Table 11:

Table 11: Comment subject sensitivity analysis

review_id	...	star_rating	helpful_ratio	S	s_new	rate of change
R230LCPQDOFJJZ.		5	0.5	0.87	0.84	D 3.45%
R21NN9ONVZITI0.		1	0.5	0.79	0.75	D 2.53%
...	...	...	...	...	...	...
R3N0F2FKJOMGKK		5	0.95	0.95	0.96	I 1.05%

Note: We use I for increase and D for decrease.

It is not difficult to find that the satisfaction rate of comments does not increase or decrease significantly and the change rate is small, indicating that the model is insensitive to the change of hyperparameters and the model has good robustness.

11 Conclusion

E-commerce platforms led by Amazon have been integrated into our lives. In order to help Sunshine Company prepare for online sales, we combined data analysis with LDA model to analyze the characteristics of data. Secondly, we built a self-built corpus to construct classic example sentences, analyzed them with Word2vec, and then clustered the comments into four categories: Good, good, poor and very poor, corresponding to grades 1, 2, 3 and 4, respectively. Then, EWM-Topsis is used to obtain reputation R, ARIMA is used to predict the increase and decrease of R, and the trend can be obtained by observing the sequence diagram. SVM was used to distinguish between successful and unsuccessful products using reputation as a criterion. In addition, we also use the Apriori algorithm, mining strong association rules, to get three products of high and low star corresponding specific comment words, such as hair_dryer, high star with "love", "great", low star with "cost", "heavy" and so on. Finally, write to the director of Sunshine company with advice on "focusing on bad reviews and improving quality".

12 Letter

Dear Sunshine Company Marketing Director:

We are honored to present to you our team's data analysis and results, and to give you sound recommendations for your three products.

The first step is data preprocessing. The comments with VINE N and verified_purchase N are rejected. Step 2: Data visualization, visualize the review stars and construct helpful_ratio using the sigmoid function. Step 3: Descriptive statistics, descriptive statistics of review stars and Helpful_ratio. Then the LDA topic model is constructed to extract the topic of each comment. It was found that the effect was very good after data cleaning, and there were almost no comments with a help rate lower than 0.5. Five-star reviews were always the most among the three products, and the distribution of five-star reviews was roughly the same between pacifier and Hair_dryer, while the two-star reviews were the least, indicating that the quality of the two products was generally satisfactory, but still needed to be strengthened. However, the number of one-star reviews in microwave is second only to that of five-star reviews, indicating that there may be quality problems and other aspects, so we need to pay attention to the specific content of one-star reviews.

Then, by using effective review subject, considering the need structure similarity, need self-built corpus, this can comment on the best use of existing information, on this basis to build five classic example, set up corresponding score of 0.1, 0.3, 0.5, 0.7, 0.9, use Word2vec to comment and classic example similarity analysis, The satisfaction rate of each comment is defined as the product of the score of the classic example with the greatest similarity and the maximum similarity. Then combining satisfaction rate, star rating and helpful_ratio, k-means algorithm is used to classify customers' comments on products, so as to obtain valuable comments. Customers' comments on products are divided into four categories: Reviews in the "very good" category reveal the advantages of a product, while reviews in the "very bad" category are often the most valuable, and are key to improving a product and influencing decision making.

In addition, in order to reflect the increase or decrease of product reputation, we defined the reputation rate R, completed the construction of the evaluation model of R through EWM-Topsis model, and took the final score as R. Since the data has date data every day, and Sunshine company also attaches great importance to time measurement, product review data can be regarded as a time series, and a prediction model of product reputation rate is constructed through ARIMA model to quantitatively analyze whether reputation is increasing or decreasing. Moreover, in order to define the evaluation criteria of potential success or failure, we first used One Class SVM model to distinguish normal products from abnormal products (potential success or failure products), and then trained SVM model to give judgment of successful or failed products, and achieved good results.

Finally, we discuss the correlation between review stars and reviews and find strong association rules between review stars and specific review words. We first use Pearson correlation coefficient to explore the correlation between stars and the number of reviews, as well as between stars and review levels. It is found that the higher the star_rating is, the more reviews there are, and the more favorable reviews there are; the lower the star_rating is, the fewer reviews there are, and the more negative reviews there are. In order to find strong association rules, we used Apriori algorithm and found that for hair dryers, high stars often correspond to ”hot” and ”Quick”, while low stars often correspond to ”waste money”. Therefore, hair dryers should be hot enough to dry hair quickly, but not too expensive. For pacifiers, high stars tend to correspond to ”clean” and ”soft”, while low stars tend to correspond to ”hard” and ”small”. Therefore, baby pacifiers should be soft, the right size and fit the baby. For a microwave oven, a high star usually corresponds to easy or fit, while a low star usually corresponds to Repair. Therefore, microwave ovens should be easy to use and provide excellent after-sales service.

The above is our team’s analysis results and suggestions for you. We sincerely hope that our research results and suggestions can help your products succeed. Looking forward to your reply, thank you!

Sincerely

Team 2019057552

References

[1] Zou Xiaohui, Sun Jing. LDA Topic Model[J]. Intelligent Computer and Application, 2014(5). DOI:10.3969/j.issn.2095-2163.2014.05.031.

[2] Tong Z, Zhang H. A text mining research based on LDA topic modelling[C]//International Conference on Computer Science, Engineering and Information Technology. 2016: 201-210.

[3] Zhou Lian. The working principle and application of Word2vec[J]. Science and Technology Information Development and Economy, 2015(2):145-148. DOI:10.3969/j.issn.1005-6033.2015.02.061.

[4] Yang Junchuang, Zhao Chao. Review of K-Means Clustering Algorithm Research [J]. Computer Engineering and Applications, 2019,55(23):7-14,63. DOI:10.3778/j.issn.1002-8331.1908-0347 .

[5] Wei Jie, Li Quanming, Chu Yanyu, et al. Optimization of room-and-pillar stope layout scheme based on EWM-TOPSIS model [J]. Journal of Hefei University of Technology (Natural Science Edition), 2021,44(5):691-695. DOI :10.3969/j.issn.1003-5060.2021.05.020.

[6] Ji Hua. Support Vector Machine (SVM) Learning Method Based on Statistical Learning Theory[J]. Science Times, 2006(11):33-37.

[7] Peng Yue. Introduction of ARIMA Model[J]. Electronic World, 2014(10):259-259. DOI:10.3969/j.issn.1003-0522.2014.10.252.

[8] Liu Dongyang, Liu En. Improvement of Apriori Algorithm[J]. Science Technology and Engineering, 2010,10(16):4028-4031. DOI:10.3969/j.issn.1671-1815.2010.16.054.

Appendices

Here are simulation programmes we used in our model as follow. Just show the main code, other code in the file shown below.

1. Data Visualization Mercer-Data-visualization.py
2. K-Means Clustering Mercer-Kmeans-clustering.py
3. Data Processing Mercer-Data-processing.py
4. LDA Mercer-LDA.py

Appendix A K-Means Clustering Algorithm

```
from numpy import *
import matplotlib.pyplot as plt

def loadDataSet(fileName):
    dataMat = []
    fr = open(fileName)
    for line in fr.readlines():
        curLine = line.strip().split('\t')
        fltLine = map(float, curLine)
        dataMat.append(fltLine)
    return dataMat

def distEclud(vecA, vecB):
    return sqrt(sum(power(vecA - vecB, 2)))

def randCent(dataSet, k):
    n = shape(dataSet)[1]
    centroids = mat(zeros((k,n)))
    for j in range(n):
        minJ = min(dataSet[:,j])
        maxJ = max(dataSet[:,j])
        rangeJ = float(maxJ - minJ)
        centroids[:,j] = minJ + rangeJ * random.rand(k, 1)
    return centroids

def kMeans(dataSet, k, distMeans =distEclud, createCent = randCent):
    m = shape(dataSet)[0]
    clusterAssment = mat(zeros((m,2)))
    centroids = createCent(dataSet, k)
    clusterChanged = True
    while clusterChanged:
        clusterChanged = False;
        for i in range(m):
            minDist = inf; minIndex = -1;
```

```
            for j in range(k):
                distJI = distMeans(centroids[j,:], dataSet[i,:])
                if distJI < minDist:
                    minDist = distJI; minIndex = j
            if clusterAssment[i,0] != minIndex:
                clusterChanged = True
            clusterAssment[i,:] = minIndex,minDist**2
        for cent in range(k):
            ptsInClust = dataSet[nonzero(clusterAssment[:,0].A == cent)[0]]
            centroids[cent,:] = mean(ptsInClust, axis = 0)
return centroids, clusterAssment
```

Appendix B Data Preprocessing

```
import seaborn as sns
sns.axes_style("darkgrid")
explode =[0,0,0,0.3,0]
labels =['one stars','two stars','three stars','four stars','five stars']
colors = ['salmon','tan','darkorange','skyblue','khaki']
patches,l_text,p_text=plt.pie(x=pacifier_stars_ratio,labels=labels,
        explode=explode,colors=colors, autopct='%.3f%%',pctdistance=0.4,
         ↪ labeldistance=0.7,startangle=180,center=(4,4),radius=3.8,counterclock=
         ↪ False,

for t in p_text:
    t.set_size(17)

for t in l_text:
    t.set_size(17)
plt.xticks(())

plt.yticks(())
plt.title('pacifier_stars_ratio',y=-1.18,fontsize=18)
plt.legend(patches,
           labels,
           fontsize=18,
           loc="center left",
           bbox_to_anchor=(2, 0, 1, -1))

plt.show()


from wordcloud import WordCloud
import matplotlib.pyplot as plt
from PIL import Image
wordcloud =
 ↪ WordCloud(mask=mask,background_color='#FFFFFF',scale=1,).generate(a)
```

```
image_produce = wordcloud.to_image()
wordcloud.to_file("new_wordcloud.jpg")
image_produce.show()
```

Appendix C LDA

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

hair_dryer_data = pd.read_excel('hair_dryer.xlsx')
microwave_data = pd.read_excel('microwave.xlsx')
pacifier_data = pd.read_excel('pacifier.xlsx')

import re


#
def clean_text(text):
    text = text.replace("<br />", " ")
    text = text.replace("<br", " ")
    text = re.sub(r'[^\x00-\x7F]+', ' ', text)
    text = re.sub(r"([.,!:?()])", r" \1 ", text)
    text = re.sub(r"\s{2,}", " ", text)
    text = text.replace("-", " ")
    return text


hair_dryer_data['review_body'].apply(clean_text)

import nltk
from sklearn.feature_extraction.text import TfidfVectorizer,
 ↪  CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation

n_features = 1000
tf_vectorizer = CountVectorizer(strip_accents='unicode',
                                max_features=n_features,
                                stop_words='english',
                                max_df=0.5,
                                min_df=10)
tf = tf_vectorizer.fit_transform(microwave_data['review_headline'])

lda = LatentDirichletAllocation(n_components=15,
                                max_iter=150,
                                learning_method='online',
                                learning_offset=50, random_state=0)
lda.fit(tf)
```