Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2002466

Summary

Sunshine Company is planning to introduce and sell three new products in the online marketplace, thus an efficient and reliable method to track and evaluate the products profile is crucial to their success. Our goal is to establish product profile evaluation models that are based on time and comments respectively.

To ensure the reliability of our model and simplify the modeling process, we firstly sorted and cleaned the redundant and invalid data, then we made some correction to the misplaced data.

After analyzing the given data set, we first **quantified** the relationship between star ratings, reviews, helpfulness voting and the overall reputation of the product. Considering the contribution of vine members, text length, and helpful votes to the reliability of the reviews, we gave them different weight, and integrated them into **the functional expression of the overall reputation of the product.**

Then we established a **product profile evaluation model** based on star ratings and reviews. We used **TextBlob** to evaluate the sentiment value of each review. Then we optimized the classical **TOPSIS Comprehensive Evaluation Model**, using **Entropy Weight Method (EWM)** to obtain an optimal weight of each indicator, and then calculated the relative score of each product.

Next, we established a **time-based evaluation model.** We used **Python** to analyze the word frequency of all the reviews and get corresponding key words of each product. According to these keywords, we conducted **Hierarchical Clustering Algorithm** to sort the products into different categories. Then we performed **Time Series Analysis** to see how the reputation of the products changes with time. When determining the most suitable time series model, we referred to the expert modeler in **SPSS**, and combined with the actual situation of the image. **Using this model, Sunshine Company can easily track the changes in the reputation of their products.**

Whats more, we further analyzed the fitting and prediction results given by the time series analysis model, and focused more on the **inflection points of the graph** to identify the best indicator of a potentially successful or failing product, and to see if specific star ratings will incite more reviews.

Based on our model and analysis, we figured out the **desirable design features** of each product, and also recommended a sales strategy to Sunshine Company, which is to **do more products promotion.**

Keywords : TOPSIS, Entropy Weight Method, Hierarchical Clustering Algorithm, NLP, Product Profile Evaluation Model

Contents

1 Introduction

1.1 Background

In the online marketplace it created, Amazon provides customers with an opportunity to rate and review purchases. There are two major ways to comment on the products they have purchases: First, "star ratings", which allow purchases to express their level of satisfaction with a product using a scale of 1 to 5, bigger number means you have higher satisfaction towards the product. Second, text-based messages, which called "reviews", that can allow customers to further express their opinions and information about the product. Besides, other customers can submit ratings on these reviews as being helpful or not in assisting their puechasing decision, which called "helpfulness rating". Companies can use these data to gain the insights into the market in which they particpate, the best timing of that participation, and the potential success of product design feature choices.

1.2 Restatement of Problem

Sunshine Company is planning to introduce and sell three new products in the online marketplace: microwave oven, baby pacifier, and hair dryer. We are required to figure out best online sales strategy and potentially important design features that would enhance their product desirability.

First, we should analyze the three product data sets provided to identify, describe, meaningful quantitative and/or qualitative patterns, relationships, measures, and parameters within and between star ratings, reviews, and helpfulness ratings. After the analysis, we should find out data measures based on ratings and reviews that are most informative for Sunshine Company to track once their three products are on sale.(refer to task 2a)

Then, we ought to figure out a time-based measure that can suggest whether a product's reputation is increasing or decreasing with time.(refer to task 2b)

Further, we need to find a combination of text-based measure and ratings-based measures that can best predict a product will succeed or fail.(refer to task 2c)

Lastly, we are also required to consider whether specific star ratings will incite more reviews, and analyze the relativity between star ratings and reviews.(refer to task 2d and 2e)

2 Assumptions and Symbol Table

2.1 Assumptions

To simplify the problem and make it convenient for us to establish a model, we make the following basic assumptions:

(1) The star ratings and reviews given by the customers are real.
(2) The data provided is enough to analyze the features of the products.
(3) We didn't consider the impact of the information beyond the given data.

2.2 Symbol Table

Table 1: Symble Table of Model I

Symbol	Definition
$x_{star,ij}$	original j^{th} star rating of i^{th} product
$x_{review,ij}$	original sentiment value of j^{th} review of i^{th} product
$y_{star,ij}$	final score of j^{th} star rating of i^{th} product
$y_{review,ij}$	final score of j^{th} review of i^{th} product
h_{ij}	real helpful votes of j^{th} review of i^{th} product
α_{ij}	helpfulness voting parameter of j^{th} review of i^{th} product
t_{ij}	text length coefficient of j^{th} review of i^{th} product
v_{ij}	vine member coefficient of j^{th} comment of i^{th} product
$y_{star,i}$	final score of star ratings of i^{th} product
$y_{review,i}$	final score of reviews of i^{th} product
A_i	final score of comments of i^{th} product
Z_j^+	the ideal solution of j^{th} indicator
Z_j^-	the negative ideal solution of j^{th} indicator
Z^+	the ideal solution
Z^-	the negative ideal solution
D_j^+	distance of i^{th} product from the ideal solution
D_j^-	distance of i^{th} product from the negative ideal solution
S_i	relative closeness of i^{th} product to the ideal solution
e_j	information entropy of j^{th} indicator
d_j	information utility value of j^{th} indicator
w_{star}	entropy weight of score of star rating
w_{review}	entropy weight of score of review

3 Data Preprocessing

3.1 Data Sorting and Cleaning

We classify the data given into four categories: redundant data, invalid data, normal data and quality data. The specific criteria for data categorization are as follows:

Firstly, for those Amazon Vine members, they have earned the trust in the Amazon Community for their accurate and insightful reviews, thus we can regard their star ratings as well as

Table 2: Symble Table of Model II

Symbol	Definition
C_k	the k^{th} class
μ_k	the location of the center of gravity
J	clustering coefficient
K	number of clusters

reviews as quality data, we will lay more emphasis on them in our later analysis.

Secondly, since a "Y" in the verified purchase column indicates Amazon verified that the person writing the review purchased the product at Amazon and didn't receive the product at a deep discount, we can view those that have "N" instead as invalid data. However, Amazon Vine members can get the products for free, thus we can only define those have "N" in both vine column and verified purchase column as invalid data, and we will delete them to make our analysis more rational.

What's more, we observe that there exist duplicate reviews of same product given by same customer, thus we regard these data as redundant data, and delete them to make it more concenient for us to search for our required data.

Lastly, for the remaining data, we see them as normal data, and sort them according to their product titles.

3.2 Data Correction

In data preprocessing, we observe that a small proportion of data are displaced, some information is put altogether in one column, which is supposed to belong to different columns. Therefore, we reallocate the data according to their data types.

4 Model I Product Profile Evaluation Model

4.1 Data Processing

4.1.1 Evaluation of Star Rating

Regarding the product desirability as reflected in the star rating, since star rating is already a quantitative indicator, we only need to do some simple processing which is as follows:

$$y_{star,ij} = \frac{x_{star,ij} - min(x_{star,ij})}{max(x_{star,ij}) - min(x_{star,ij})} \tag{1}$$

where $y_{star,ij}$ represents the final score of j^{th} star rating of i^{th} product, $x_{star,ij}$ represents the original j^{th} star rating of i^{th} product.

4.1.2 Evaluation of Review

Regarding the product desirability as reflected in the reviews, we try to convert it into a quantitative indicator. The specific procedures are as follows:

We use textblob to perform user sentiment analysis, and get the respective sentiment value of review title and body, then give different weight on the two to obtain the final sentiment value of the review. Note that the sentiment value can be either positive or negative. Positive value indicates that the review is positive, negative value indicates that the review is negative.

After the original sentiment value of each review is obtained, we further consider the reliability of the review in terms of helpfulness voting and the length of the review, and then perform weighting processing when finally calculating the sentiment value of the reviews of a certain product. The specific functional expression is shown below:

$$y_{review,ij} = x_{review,ij} \times (1 + \alpha_{ij}) \times t_{ij} \tag{2}$$

where

$$\alpha_{ij} = \frac{h_{ij}}{max(h_{ij}) - min(h_{ij})} \tag{3}$$

and

$$\begin{aligned} h_{ij} &= helpful\ votes_{ij} - unhelpful\ votes_{ij} \\ &= helpful\ votes_{ij} - (total\ votes_{ij} - helpful\ votes_{ij}) \\ &= 2 \times helpful\ votes_{ij} - total\ votes_{ij} \end{aligned} \tag{4}$$

$y_{review,ij}$ represents the final sentiment value of the j^{th} review of i^{th} product , $x_{review,ij}$ represents the original sentiment value of j^{th} review of i^{th} product, h_{ij} represents the real helpful votes of j^{th} review of i^{th} product, α_{ij} is the helpfulness voting parameter of j^{th} review of i^{th} product, t_{ij} is the text length coefficient of j^{th} review of i^{th} product.

4.1.3 Combination of Star Rating and Review : Solution to Task 1

In the previous steps, we have performed respective evaluation of star rating and review, next we are going to combine these two to obtain a final evaluation of the comment. The functional expression is as follows:

$$y_{star,i} = \sum_{j=1}^{m} \frac{v_{ij} \times y_{star,ij}}{m} \tag{5}$$

$$y_{review,i} = \sum_{j=1}^{m} \frac{v_{ij} \times y_{review,ij}}{m} \tag{6}$$

$$\begin{aligned} A_i &= f(x_{star,ij}, x_{review,ij}, h_{ij}, t_{ij}, v_{ij}) \\ &= w_{star} y_{star,i} + w_{review} y_{review,i} \end{aligned} \tag{7}$$

where $y_{star,i}$ represents the final score of star rating of i^{th} product, $y_{review,i}$ represents the final score of review of i^{th} product, A_i reprensents the final score of i^{th} product, v_{ij} represents the vine member coefficient of j^{th} comment of i^{th} product, if the comment is given by a vine member, $v_{ij} = 2$, otherwise, $v_{ij} = 1$.

4.2 TOPSIS Comprehensive Evaluation Method

4.2.1 Introduction

TOPSIS(Technique for Order Performance by Similarity to Ideal Solution), which was first proposed by C.L.Hwang and K.Yoon, is a useful technique in dealing with multiattribute or multi-criteria decision making(MADM/MCDM) problems in the real world. It is a commonly used comprehensive evaluation method, which can make full use of the original data, the results of which can accurately reflect the gap between different schemes. This method has no strict restrictions on data distribution and sample content, and data calculation is also simple and practicable. A complete and efficient procedure for establishing the model will be provided later.

4.2.2 Model Establishment

Step One : Normalization

Since the different indicators differ greatly in dimension, we should first normalize the value of these two factors to eliminate the impact of different index dimensions. The normalized value X_{ij} of the decision matrix X can be any linear-scale transformation to keep $0 \leq X_{ij} \leq 1$.

We consider the normalized value of X_{ij} is the value of the corresponding element X_{ij} divided by the operation of its column elements,i.e.,vector normalization, then(take factor as example).

$$Z_{m1} = \frac{A_n}{\sqrt{A_1^2 + A_1^2 + \cdots + A_n^2}}$$

Step Two : Construct Decision Matrix X

The structure of the matrix can be expressed as follows:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ z_{i1} & z_{i2} & \cdots & z_{ij} & \cdots & z_{in} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mj} & \cdots & z_{mn} \end{bmatrix}_{m \times n}$$

where each row represents the alternative X_i, i=1,2,...m, each column represents the indicator j, j=1,2,...,n. In our model, each row represents each product, and each column represents each indicator in the evaluation.

Step Three : Determine the Ideal and Negative Ideal Solutions

$$\begin{aligned} Z^+ &= \{Z_1^+, \ldots, Z_n^+\} \\ &= \{(max\{z_{11}, z_{21}, \ldots, z_{n1}\}, max\{z_{12}, z_{22}, \ldots, z_{n2}\}, \ldots, max\{z_{1m}, z_{2m}, \ldots, z_{nm}\})\} \\ Z^- &= \{Z_1^-, \ldots, Z_n^-\} \\ &= \{(min\{z_{11}, z_{21}, \ldots, z_{n1}\}, min\{z_{12}, z_{22}, \ldots, z_{n2}\}, \ldots, min\{z_{1m}, z_{2m}, \ldots, z_{nm}\})\} \end{aligned}$$

Step Four : Calculate Seperate Distance from the Ideal and Negative Ideal Solutions

Calculate the seperate distance from the ideal and the negative ideal solutions, D_i^+ and D_i^-, respectively, for the group.

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(Z_j^+ - z_{ij})^2},\ for\ alternative\ i,\ i = 1, \ldots, m$$

$$D_i^- = \sqrt{\sum_{j=1}^{n}(Z_j^- - z_{ij})^2},\ for\ alternative\ i,\ i = 1, \ldots, m$$

Step Five : Calculate the relative closeness S_i to the Ideal Solution

Calculate the relative closeness of i^{th} alternative to the ideal solution and rank the alternatives in descending order. The relative closeness of ith alternative A_i with respect to positive ideal solution can be expressed as

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-},\ i = 1, \ldots,\ m$$

where $0 \leq S_i \leq 1$, the larger the index value, the better the performance of the alternative.

4.3 The Optimization of TOPSIS Based on Entropy Weight Method

4.3.1 Introduction of Entropy Weight Method

In TOPSIS Model, we have acquiesced that the weight of each indicator is the same. However, in the case analysis of the real world, it's obvious that different indicators will have different impact on the results, thus should be given different weight.

In Analytic Hierarchy Process(AHP), there is one way to determine the weight of different indicators, that is to ask a so-called "expert". However, we obviously don't have an expert to turn to right now and this kind of method is also rather subjective. Therefore, we figured out a more objective method to help us determine the weight of different indicators, which is the Entropy Weight Method(EWM).

4.3.2 Model Establishment

Step One : Restandarization

Judge whether there is a negative number in the matrix X, and if there is one, another standardization method should be used for the positive matrix X in TOPSIS to standardize it into matrix $\tilde{Z}$, its standardized formula is:

$$\tilde{z}_{ij} = \frac{x_{ij} - min\{x_{1j}, x_{2j}, \ldots, x_{nj}\}}{max\{x_{1j}, x_{2j}, \ldots, x_{nj}\} - min\{x_{1j}, x_{2j}, \ldots, x_{nj}\}}$$

Step Two : Calculate the Probability of Each Element

Calculate the proportion of the i_{th} sample under the j_{th} index, and regard it as the probability used in the calculation of relative entropy Suppose that there are n objects to be evaluated, m evaluation indexes, and the nonnegative matrix obtained through the previous processing is:

$$\tilde{Z} = \begin{bmatrix} \tilde{z_{11}} & \tilde{z_{12}} & \ldots & \tilde{z_{1j}} & \ldots & \tilde{z_{1n}} \\ \tilde{z_{21}} & \tilde{z_{22}} & \ldots & \tilde{z_{2j}} & \ldots & \tilde{z_{2n}} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ \tilde{z_{i1}} & \tilde{z_{i2}} & \ldots & \tilde{z_{ij}} & \ldots & \tilde{z_{in}} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots \\ \tilde{z_{m1}} & \tilde{z_{m2}} & \ldots & \tilde{z_{mj}} & \ldots & \tilde{z_{mn}} \end{bmatrix}_{m \times n}$$

We then calculate the probability matrix P, where p_{ij} of each element in P is calculated as follows:

$$p_{ij} = \frac{\tilde{z}_{ij}}{\sum_{i=1}^{m} \tilde{z}_{ij}}$$

It's easy to verify : $\sum_{i=1}^{m} p_{ij} = 1$, the sum of probability corresponding to each index is 1.

Step Three : Calculate the Information Entropy and Get the Entropy Weight

We first calculated the information entropy of each index, then calculated the information utility value, and normalized it to get the entropy weight of each index.

For the j_{th} index, the calculation formula of information entropy is:

$$e_j = -\frac{1}{lnn} \sum_{i=1}^{m} p_{ij} ln(p_{ij}) \ (j = 1, 2, \ldots, m)$$

Definition of information utility value: $d_j = 1 - e_j$, the larger the information utility value is, the more information it can provide. By normalizing the information utility value, we can get the entropy weight of each index:

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \ (j = 1, 2, \ldots, n)$$

4.4 Solutions to Task 2a

Since we have hair dryer, pacifier and microwave to consider, and each of them has hundreds of different products, thus here we only take hair dryer as an example. The raw score of each hair dryer is shown in Table 3.

Table 3: Scores of Different Hair Dryer

product parent	product title	star rating	review
466064538	conair 1875 watt turbo hair dryer and styler, 30.4 ounce	4.93103	2.88248
453645026	turbo power 1500 hair dryer	4.71429	2.74276
407404113	panasonic nano-e nano care hair dryer eh-na95 — ac100v 50-60hz (japan model)	4.84783	2.69185
977457747	babyliss pro babfrv2 volare ferrari designed professional luxury mid sized hair dryer, red, 2000 watts	4.77273	2.52336
945323010	babyliss pro hair dryer ceramix xtreme pink edition - babbpk2000	4.71429	2.50608
⋮	⋮	⋮	⋮
932955324	hot tools professional 1061 hard hat 1200 watt salon hair dryer	3.26087	1.3213
862140913	blo and go by laurie coleman - portable hair dryer holder	3.73913	1.20012
496940864	ovente seductive ceramic ionic tourmaline lightweight professional hair dryer, matte black	3.75	1.13611
981727854	andis 80345 styler hair dryer	3.27586	1.04681
955015830	revlon 1875w rectractable cord, fold go hair dryer	3.23333	0.856151

The matrix X in the TOPSIS model and the normalized matrix Z are shown below:

$$X = \begin{bmatrix} 4.93103 & 2.88248 \\ 4.71429 & 2.74276 \\ 4.84783 & 2.69185 \\ 4.77273 & 2.52336 \\ 4.71429 & 2.50608 \\ \vdots & \vdots \\ 3.26087 & 1.3213 \\ 3.73913 & 1.20012 \\ 3.75 & 1.13611 \\ 3.27586 & 1.04681 \\ 3.23333 & 0.856151 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 1 \\ 0.872333156623668 & 0.931047722260304 \\ 0.950992519290805 & 0.905923470472959 \\ 0.906756199564116 & 0.822773103479247 \\ 0.872333156623668 & 0.814245366867868 \\ \vdots & \vdots \\ 0.0162219473405196 & 0.229552555384639 \\ 0.297932496907581 & 0.169749828384236 \\ 0.30433527713966 & 0.138160683679699 \\ 0.0250515403192556 & 0.0940908411220488 \\ 0 & 0 \end{bmatrix}$$

Based on the Entropy Weight Method, we then calculated the propability matrix P, which is as follows:

$$P = \begin{bmatrix} 0.00944978 & 0.012505415 \\ 0.009290336 & 0.011678382 \\ 0.009034421 & 0.011899251 \\ 0.009146415 & 0.010947401 \\ 0.009034421 & 0.010872433 \\ \vdots & \vdots \\ 0.007186465 & 0.004928925 \\ 0.006446048 & 0.006183163 \\ 0.006249101 & 0.005732357 \\ 0.006277827 & 0.004541504 \\ 0.006196323 & 0.003714345 \end{bmatrix}_{m \times 2}$$

Using the probability matrix, we then calculated the information entropy of each indicator, and obtained the entropy weight. The information entropy of star rating $e_{star} = 0.999319665$, the information entropy review $e_{review} = 0.996331899$, thus, the entropy weight of star rating $w_{star} = 0.156455179$, the entropy weight of review $w_{review} = 0.843544821$.

After calculating the the entropy weight of each indicator, now we can apply formula 7 to calculate the final score of each product, part of the result is shown in Table 4:

Table 4: Final Score of Different Hair Dryers

product parent	product title	final score
466064538	conair 1875 watt turbo hair dryer and styler, 30.4 ounce	0.0161
453645026	turbo power 1500 hair dryer	0.0148
407404113	panasonic nano-e nano care hair dryer eh-na95 — ac100v 50-60hz (japan model)	0.0147
977457747	babyliss pro babfrv2 volare ferrari designed professional luxury mid sized hair dryer, red, 2000 watts	0.0134
945323010	babyliss pro hair dryer ceramix xtreme pink edition - babbpk2000	0.0132
⋮	⋮	⋮
932955324	hot tools professional 1061 hard hat 1200 watt salon hair dryer	0.0034
862140913	blo and go by laurie coleman - portable hair dryer holder	0.0031
496940864	ovente seductive ceramic ionic tourmaline lightweight professional hair dryer, matte black	0.0027
981727854	andis 80345 styler hair dryer	0.0014
955015830	revlon 1875w rectractable cord, fold go hair dryer	0

5 Model II Time-Based Evaluation Model

5.1 Model Introduction

We first use Python to analyze the word frequency of all the reviews and get the corresponding keywords of each product, and for all the keywords extracted from each product, we have used Python draw a word cloud which is shown in Figure 1. Each keyword is taken as an index. If the product has this keyword, it will be recorded as 1. If it does not have this keyword, it will be recorded as 0. Then we can use these data to perform clustering process to sort these products into different categories. Later, time series analysis is carried out to see the reputation of which kinds of products is increasing with time, so that we can know which characteristics are increasingly popular with customers.

Notice : Since we have thousands of different products to consider, here we only conduct hierarchical clustering on those top-selling products to simplify our computing process. Also, top-selling products have a large number reviews, thus the analysis of them will be reliable.

5.2 Hierarchical Clustering Algorithm

The merging algorithm of system clustering combines the closest two data points by calculating the distance between each two data points, and iterates the process repeatedly until all the

Figure 1: Word Cloud of Hair Dryer, Pacifier, and Microwave

data points are combined into one group, and in the meantime generates a clustering pedigree, a sketch map is shown in Figure 2:

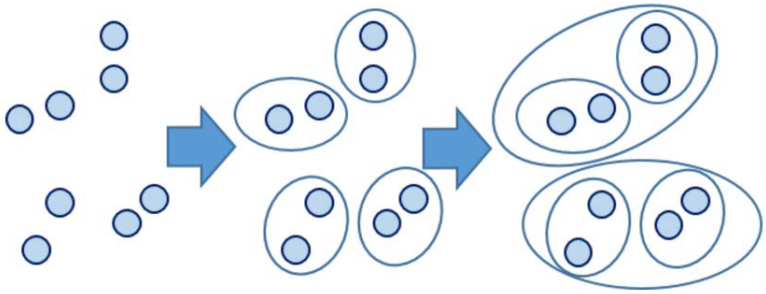

Figure 2: A Sketch Map of Hierarchical Clustering (by qingfeng)

Procedures of hierarchical clustering (as shown in Figure 3):
(1) Each object is regarded as a class, and the minimum distance between two objects is calculated;
(2) The two classes with the smallest distance are combined into a new class;
(3) Recalculate the distance between the new class and all classes;
(4) Repeat two or three steps until all classes are finally merged into one class;

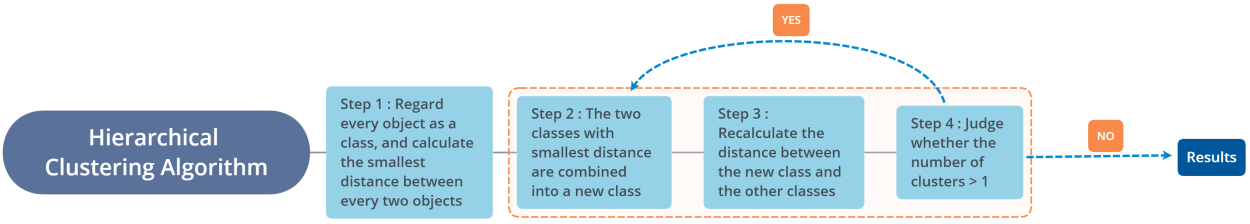

Figure 3: Procedure of Hierarchical clustering

Elbow method: roughly estimate the optimal number of clusters by line chart of clustering coefficient.

The sum of the distortion degree of each class: the distortion degree of each class is equal to the sum of the square of the distance between the center of gravity of this class and its internal members. Suppose that we divided a total of n objects into K classes ($K \leq n-1$, i.e. there are at least two objects in one class). The k^{th} class (k = 1, 2,..., K) is represented by C_k, and the location of the center of gravity of this class is recorded as μ_k, then the distortion degree of the k_{th} class is: $\sum_{i \in C_k} |x_i - \mu_k|^2$. Define the total degree of distortion for all classes:$J = \sum_{k=1}^{K} \sum_{i \in C_k} |x_i - \mu_k|^2$, J is also called the clustering coefficient. Then we draw the line graph of clustering coefficient: the abscissa is the number of clustering categories K, and the ordinate is the clustering coefficient J.

The line charts of clustering coefficient of hair dryer, pacifier and microwave are shown in Figure 4. By observing the line chart of clustering coefficient of hair dryer, we can notice that when K = 4, the downward trend slows down, and when K changes from 1 to 4, the change of distortion degree is maximized, thus we can set the number of clusters of hair dryer $K_{hd} = 4$. Then according to the clustering pedigree graph (see Figure 5), we can obtain the clustering results. Similarly, we can set the number of clusters of pacifier $K_p = 6$, the number of clusters of microwave $K_m = 4$.

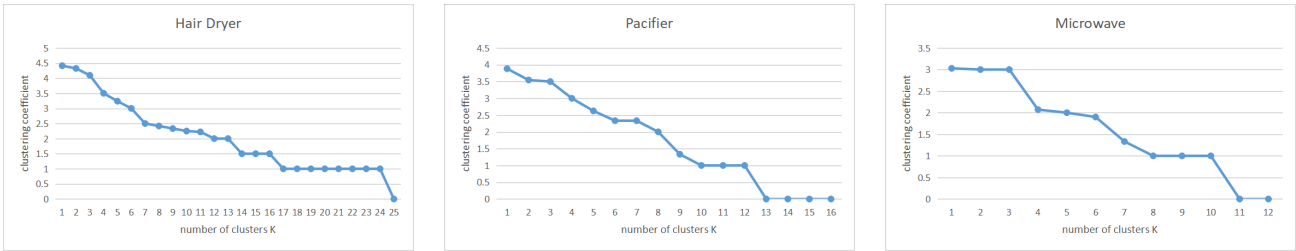

Figure 4: Line Chart of Clustering Coefficient

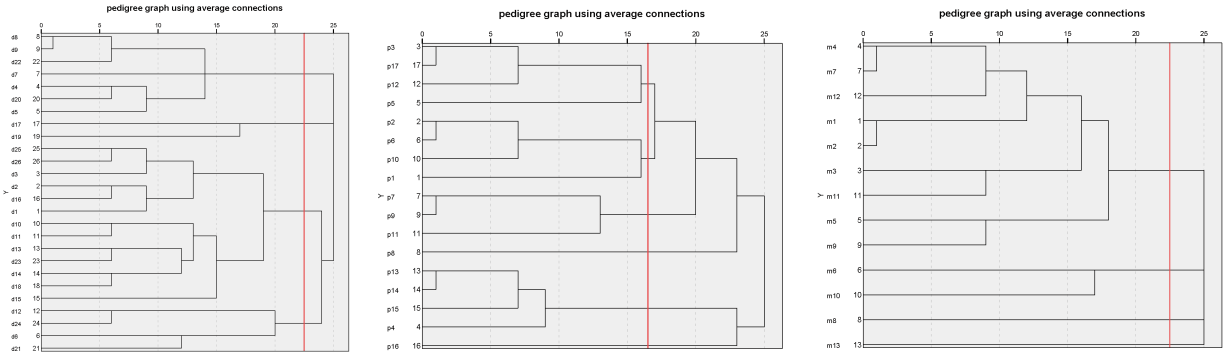

Figure 5: Clustering Pedigree Graphs of Hair Dryer, Pacifier, and Microwave

5.3 Time Series Analysis Model : Simple, Holt, ARIMA, and Simple Seasonal Model

After conducting the hierarchical clustering, we have obtained several classes of products which corresponds to different keywords, then we perform time series analysis on a quarterly

basis on every class of products to see how their reputation, i.e. the scores, changes with time.

When determining the appropriate time series model, we refered to the recommended model given by the expert modeler in SPSS, and selected the most suitable time series model for each image based on the actual situation of the image, such as simple exponential smooth model,Holt exponential smooth model, simple seasonal exponential smooth model, ARIMA model. The obtained fitting and prediction results are shown in Figure 6, 7 and 8.

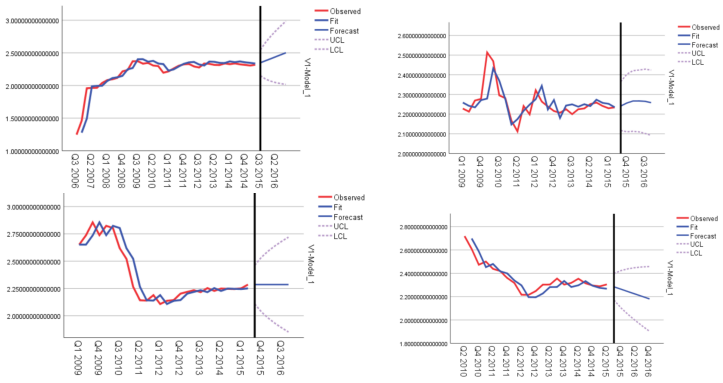

Figure 6: Reputation of Hair Dryers Relative to Time

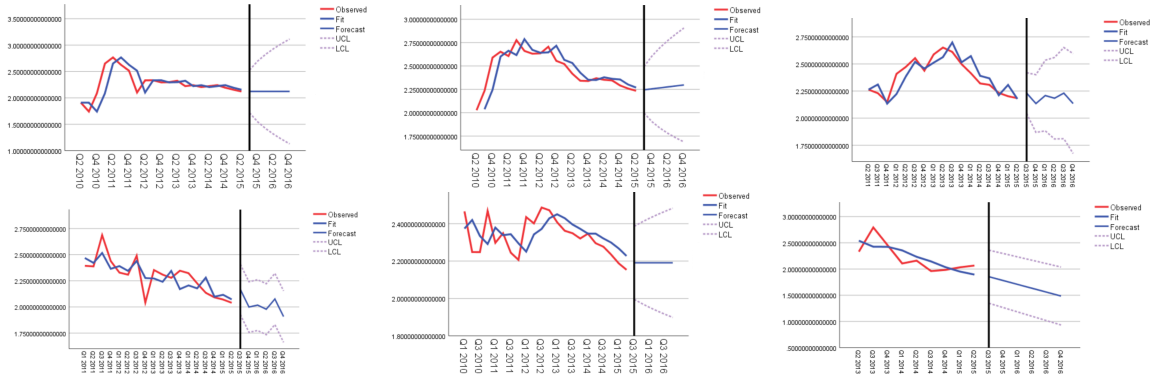

Figure 7: Reputation of Pacifiers Relative to Time

5.4 Results Analysis : Solutions to Task 2b

In the fitting and prediction results of above time series analysis model, we can directly observe from the graphs that the reputation of some products are increasing with time, while some are decreasing with time, and we can also see the prediction of reputaion in later days. Thus this is a quick, efficient and reliable method for Sunshine Company to adopt to see if their products' reputation is increasing or decreasing in the online marketplace, once they've collected the data set.

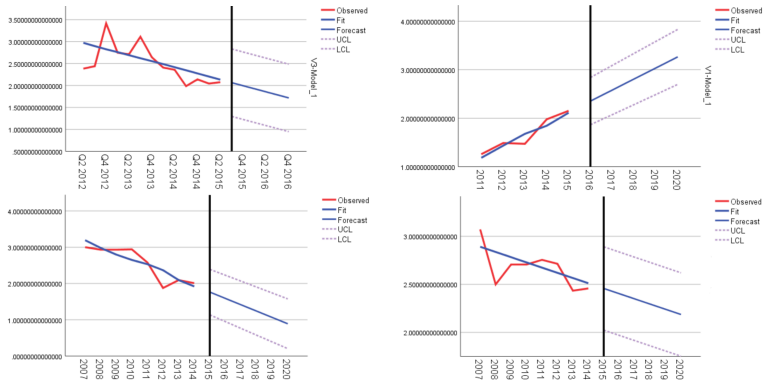

Figure 8: Reputation of Microwave Relative to Time

5.5 Analysis of the Inflection Points : Solutions to Task 2c and 2d

After careful analysis of some significant inflection points of the line charts, we obtained some useful information of important indicators of a potentially successful or failing product:

For potentially failing product:

1. Customers reflect on the same problems of the product during a short period of time.
2. Customers show their love for some features of the product, but still gave poor comments because of other intolerable defects.
3. After one customer gave a poor comment on the product, other customers followed to comment on some other shortcomings.
4. The picture of the product shown online does not match the real object.

For potentially successful product:

1. The reputation of the product continues to increase with time.
2. At the beginning, the reputation of the product is relatively low, but with the promotion online or in the hotel, the reputation of the product increased dramatically, which indicates that the products with more promotion are more likely to succeed.

Next, we analyzed the reviews after a series of low star ratings, to see if specific star ratings will incite more reviews.

We classify the products of hair dryer, pacifier and microwave oven according to the product parent, and find out the three products with the largest number of 1 and 2 star ratings respectively. After we collected 9 groups of commodity data, we then tried to find out the continuous low star ratings appeared in the data (here we only think that more than 2 successive low star ratings can be called continuous low star ratings). During this process, we observe that it is rare to have low star ratings continuously, and most of the low star ratings are mixed with the high star ratings. However, in the analysis of 10 groups of consecutive low star ratings, we find that after a series of low star ratings, there will be more continuous positive reviews about the product.

6 Further Discussion

6.1 Relativity between Star Rating and Review: Solutions to Task 2e

When we analyze the customer's evaluation, we find that specific quality descriptors, such as "enthusiastic", "disappointed", will appear in some reviews. However, statistics show that there are four reviews that use the word "enthusiastic". Reading the reviews directly, we can find that two of them are strongly recommending the products, while the other two are poor reviews. These four comments correspond to three five stars and one two stars, especially the two very similar comments correspond to five stars and two stars respectively. Then we selected 50 comments that includes the word "disappointed", and their corresponding star ratings are shown in Figure 9.

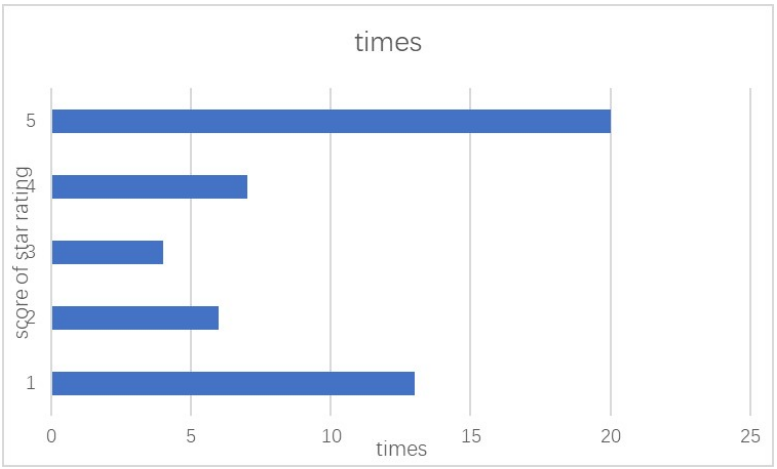

Figure 9: Corresponding Star Ratings of the Reviews that Include "disappointed"

From above analysis, we can roughly guess that there is no strong relativity between star rating and review. In order to get a more accurate and reliable conclusion, we use the score of star rating and score of review calculated previously, and use MATLAB to get the fitting curve, which is shown in Figure 10, and get the following results:

The fitting coefficient R of each graph is 0.3429, 0.2623, 0.2769 in turn. From this, we can draw the following conclusion: there is a certain relation between star rating and comment rating, but the absolute value of fitting coefficient is small, thus the correlation between them is not strong.

The possible reasons for the phenomenon are:

1. They have received malicious comments from the trade due to competition.

2. People tend to give a higher score in the star rating, but the comments can better reflect the real feelings of customers. Also, star rating is simply given from 1-5, while text review is more comprehensive in product evaluation, with less limitation and greater differentiation.

6.2 Desirable Design Features and Recommended Sales Strategy

Desirable Design Features:

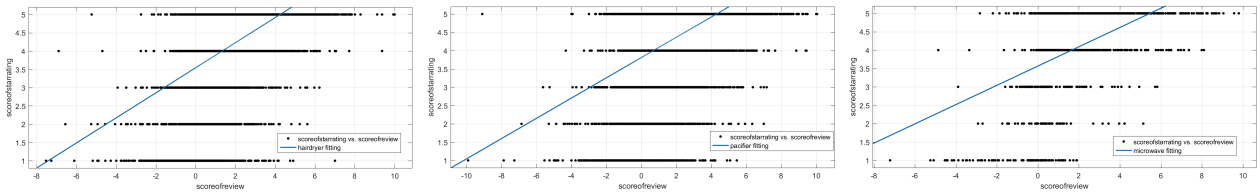

Figure 10: Relativity of Star Rating and Review

In the fitting and prediction results of above time series analysis model, we can readily observe that the reputation of some products are increasing quickly with time. We picked out the key words of those products whose reputation are increasing, and after careful analysis, we have obtained following results:

For hair dryer, customers favor features like **portable, durable, cheap and high power**.

For pacifier, customers favor features like **light-weight, visible in the dark, attractive appearance and soft texture**.

For microwave, customers favor features like **durable, large inner space, easy to clean and excellent after-sale service**.

Recommended Sales Strategy:

Apart from above desirable features, we have also figured out a useful sales strategy for Sunshine Company to adopt, that is to **do more promotion**. When we are analysing a kind of hair dryer whose reputation continues to increase with time, we find out that actually at the beginning, the reputation of this product is relatively low. But with its promotion in the five star hotel, both the reputation and the sales volumn of the product increased dramatically. Some customers said in the review that, after they encountered the hair dryer in the hotel, they found it very nice, so bought one online after they went home. It's clear to see that promotion will contribute a lot to the success of a product. Other than promotion in the hotel, Sunshine Company can also ask some bloggers to promote their products, which proves to very useful in the real world.

7 Sensitivity Analysis

1. Sensitivity analysis on vine member coefficient

The difference shown below means that when the vine member coefficient increased by 1, the average score of star rating and review will increase by no more than 0.0143. So the variation of this index has minor influence on the results.

2. Sensitivity analysis on text lenghth coefficient The difference shown below means that when the range of text lenghth changes, the average score of star rating and review will increase by no more than 0.01. So the variation of this index has minor influence on the results.

Table 5: Sensitivity Analysis on Vine Member Coefficient

	vine member coefficient t_{ij}	1	2	3
hair dryer	average score of star rating	4.1886	4.1931	4.1974
	difference	-0.0045	0	0.0043
	average score of review	1.8894	1.8882	1.8871
	difference	0.0012	0	-0.0011
pacifier	average score of star rating	4.3397	4.3396	4.3395
	difference	0.0001	0	-0.0001
	average score of review	1.9057	1.9054	1.9052
	difference	0.0003	0	-0.0002
microwave	average score of star rating	4.1226	4.1269	4.1310
	difference	-0.0143	0	0.0041
	average score of review	1.8610	1.8564	1.8520
	difference	0.0046	0	-0.0044

Table 6: Sensitivity Analysis on Text Length Coefficient

	text length coefficient t_{ij}	0-150-400	0-200-400
hair dryer	average score of review	1.9873	1.8894
	difference	0.0979	0
pacifier	average score of review	2.0061	1.9054
	difference	0.1497	0
microwave	average score of review	1.9155	1.8610
	difference	0.0545	0

8 Strengths and Weaknesses

8.1 Strengths

1. Our model is innovative, we have optimized classical TOPSIS model, using Entropy Weight Method to determine the optimal weight on each indicator, which makes our model more objective and scientific.

2. We made full use of the data given to us. In our product profile evaluation model, we have considered different influence factors, such as we give different weight on the comments given by Amazon vine members, we also analyzed the contribution of the text length and helpful votes to the reliability of the review, and integrate them into the final functional expression.

3.In addition to solving the problems given, we also conduct further analysis of the data and model, and then figured out customers' favored features of each product. Furthermore, we also recommend a sales strategy to the Sunshine Company, that is to do more promotion of the products.

4. Classical mathematical models and algorithms are used in our model. We draw on the wisdom of our predecessors fully in the process of modeling and problem solving and we combine the classic mathematical models and algorithms with our model to make our model more scientific and accurate.

8.2 Weaknesses

1. In the time series analysis, we only considered those top-selling products in order to simplify our computational process, which will have an influence on comprehensiveness of the modeling results.

2. We assume that the star ratings and reviews given by the customers are real, however, there may exist some malicious comments on the products given by the trade.

9 Conclusion

In this paper, by analyzing the data set, we figured out a evaluation method on the product profiles which combines text-based measure and ratings-based measure. Besides, we identify a time-based measure to track the change of the reputaion of products. Here we conclude our findings.

Firstly, as to quantify the relationship between star ratings, reviews, helpful votes and the overall reputation of the product, we have set up a functional expression of them which is shown in formula 7.

Secondly, as to identify a data measure based on star ratings and reviews, we have combined TOPSIS and Entropy Weight Method to establish a product profile evaluation model to obtain the relative score of each product. And Sunshine Company can also adopt this measure to see the reputation of their products relative to other similar products.

Then, as to find a time-based measure to track the reputation of the products, we established a time-based evaluation model using time series analysis. And after analysis of the graphs we have drawn, we figured out the indicator to predict a successful or failing product.

At last, we conduct further discussion about the relativity between star ratings and reviews, and find out that there exists certain but not strong correlation between them. Also, we figured out customers's favored features of each product, and also recommended a useful sales strategy to Sunshine Company.

Letter

March 10, 2020

Dear Sunshine Company:

Thanks very much for hiring our team as consultants to help you identify online sales strategy and potential important design. By analyzing the data, we have constructed a comprehensive model which can give you meaningful advice to craft excellent products.

From the model, first we will calculate the score of each product and find the most welcomed product according to the star-rating and reviews. By building this sorting system, its very easy for your company to simply view the customer feedback and judge if your new design is desirable to the public. We use computers to access the word frequency, establish the relationship to specific products, then we can know some universal advantages of some kind of products, which may tell you how to improve production and show market tendency. Also, an organization of product score is to build according to the time, providing your company a more precise method to track your data and meet the future requirements.

Our model can offer you valuable information from many different aspects. For example, your new product microwave could be traced not only from the sorting and review, but also from its own character and property. The following introduces of our models in more details.

Natural Language Processing (NLP) is basis our module. Text-based comments are more valuable reference than star ranking when developing business strategies. But, its hard to measure text-based comments using mathematics. NLP just gives us a way to quantify abstract information from text written by customers. It turns long texts into word frequency, numerically expressed emotions and so on, which can be used to build mathematics models to solve many problems.

To obtain the properties of products that customers pay attention to, we calculate the word frequency of text-based comments. The property of owning the most frequency is necessary for successful products. And we give every product a binary string created by referring to those important properties. With the binary strings, we can cluster products.

Based on the previous analysis, we use some practical assumptions to decide the coefficients of our classification system to divide products into different types. By analyzing the inter-data difference, we use some math methods to calculate the weight between star-rating and reviews. Since then, we work out a very scientific indicator to reflect the overall quality of this product.

Through our work. we also use a practical model to indicate the connection between products and the time. Its very important to keep abreast of customer needs since public opinion could change over time. Our model could help you get a deep understanding of the sorting system and let your company stay ahead of the competitors.

Whats more, we can provide you some business strategies. A reasonable price is in requirement of all the customers of the three products. For the hair dryers, we found that most customers prefer those with light weight and retractable cords since convenience is very important. Power and hair protecting are also essential. As for the microwaves, two things that customers complain most are the design of doors and service. But most of the microwaves can satisfy them in the aspect of space. The third product, pacifiers, belong to babies. So, the customers care about whether the products are suitable and for babies, and if they are easy to clean.

In conclusion, our model gives a method measure if a product is successful combing both star rating and text comments. At the same time, we show how attitudes of customers toward products owning different features change with time. Hopefully these can help your company produce better products for customers and be successful.

Sincerely,
Team #2002466

References

[1] Belenson S M, Kapur K C. An algorithm for solving multicriterion linear programming problems with examples[J]. Journal of the Operational Research Society, 1973, 24(1): 65-77.

[2] Ren Y, Wang R, Ji D. A topic-enhanced word embedding for Twitter sentiment classification[J]. Information Sciences, 2016, 369: 188-198.

[3] Sahni T, Chandak C, Chedeti N R, et al. Efficient Twitter sentiment classification using subjective distant supervision[C]//2017 9th International Conference on Communication Systems and Networks (COMSNETS). IEEE, 2017: 548-553.

[4] Rodak J, Xiao M, Longoria S. Predicting helpfulness ratings of amazon product[J].

[5] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision[J]. CS224N project report, Stanford, 2009, 1(12): 2009.

Appendix

```
#include <iostream>
#include <fstream>
#include <string>
#include  <cmath>
using namespace std;

struct sample{
    string name;
    string name_;
    string a1;
    string a2;

};


int main() {
    sample *arr1 = new sample[100000];
    ifstream inFile;
    inFile.open("try.txt");
    if (!inFile.is_open()){
        cout << "Could not open the file " << endl;
        cout << "Program terminating.\n";
        exit(EXIT_FAILURE);
    }
    else{
        cout<<"right.";
    }
    int i = 0;
    for(;; i++) {
        if (!inFile.eof()) {
            getline(inFile, arr1[i].name,'\t');
            getline(inFile,arr1[i].name_,'\t');
            getline(inFile, arr1[i].a1,'\t');
            getline(inFile,arr1[i].a2);
        } else { break; }
    }
    long long name;
    double total = 0;
    int count = 0;
    for (int j = 0;j < 20000; ++j) {
            if(name == stoll(arr1[j].name)){
                total += stod(arr1[j].a1);
                count++;
            }
```

```
45.                 else if(name != stoll(arr1[j].name) && j != 0){
46.                     if(count > 5){
47.                         cout<<total/count<<endl;
48.                     }
49.                     name = stoll(arr1[j].name);
50.                     total = stod(arr1[j].a1);
51.                     count = 1;
52.                 }
53.                 else if(j == 0){
54.                 total += stod(arr1[j].a1);
55.                 count++;
56.                 name = stoll(arr1[0].name);
57.                 }
58.         }
59.     }
```

```
01. #calculate the quantity of sale
02. import xlrd
03. import xlwt
04. 
05. workbook = xlrd.open_workbook(r'E:\MCMICM\Csource\mic_pare_com.xlsx')
06. sheet_name = workbook.sheet_names()[0]
07. sheet = workbook.sheet_by_index(0)
08. f = xlwt.Workbook()
09. sheet1= f.add_sheet('refer',cell_overwrite_ok = True)
10. reference = sheet.cell(1,0).value
11. count = 0
12. count_row = 1
13. for rown in range(sheet.nrows):
14.     if rown == 0:
15.         sheet1.write(rown,0,'product_parent')
16.         sheet1.write(rown,1,'counts')
17.     elif sheet.cell(rown,0).value == reference:
18.         count = count + 1
19.     else:
20.         sheet1.write(count_row,0,reference)
21.         sheet1.write(count_row,1,count)
22.         reference = sheet.cell(rown,0).value
23.         count = 1
24.         count_row = count_row + 1
25. f.save('mic_count.xlsx')
```

```
26.
27.  #calculate the length of reviews
28.  import xlwt
29.
30.  f = open("E:\MCMICM\Csource\paci_combine.txt","r",encoding = 'utf-8')
31.  mf = xlwt.Workbook()
32.  sheet1= mf.add_sheet('refer',cell_overwrite_ok = True)
33.  sheet1.write(0,0,'lenth')
34.  i = 1
35.  f.readline
36.  for line in f:
37.      if len(line) <= 150:
38.          sheet1.write(i,0,1)
39.      elif len(line) <= 400:
40.          sheet1.write(i,0,1.5)
41.      else:
42.          sheet1.write(i,0,2)
43.      i = i + 1
44.  mf.save('paci_combine_150.xlsx')
45.
46.  #generate word cloud
47.  import xlwt
48.
49.  f = open("E:\MCMICM\Csource\paci_combine.txt","r",encoding = 'utf-8')
50.  mf = xlwt.Workbook()
51.  sheet1= mf.add_sheet('refer',cell_overwrite_ok = True)
52.  sheet1.write(0,0,'lenth')
53.  i = 1
54.  f.readline
55.  for line in f:
56.      if len(line) <= 150:
57.          sheet1.write(i,0,1)
58.      elif len(line) <= 400:
59.          sheet1.write(i,0,1.5)
60.      else:
61.          sheet1.write(i,0,2)
62.      i = i + 1
63.  mf.save('paci_combine_150.xlsx')
```

```
64.
65.  #turn text into numerical value
66.  from textblob import TextBlob
67.  import xlrd
68.
69.  workbook = xlrd.open_workbook(r'E:\MCMICM\Csource\dryer_body.xlsx')
70.  sheet_name = workbook.sheet_names()[0]
71.  sheet = workbook.sheet_by_index(0)
72.  print(sheet.name,sheet.nrows,sheet.ncols)
73.  rows = sheet.row_values(3)
74.  print(rows)
75.  for rown in range(sheet.nrows):
76.      rows = sheet.row_values(rown)
77.      st = "".join(rows)
78.      tag = TextBlob(st)
79.      print(tag.sentiment.polarity)
```