| 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