Contents

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Symblo	Description

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

Table 1: Examples of untrustworthy comments

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

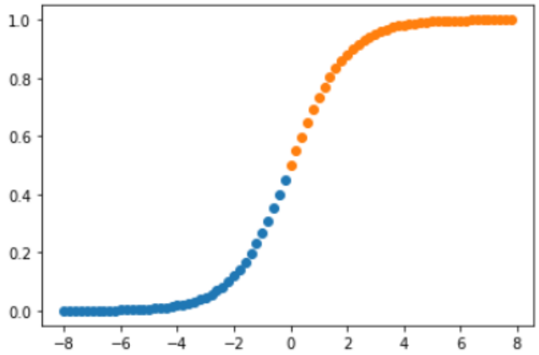
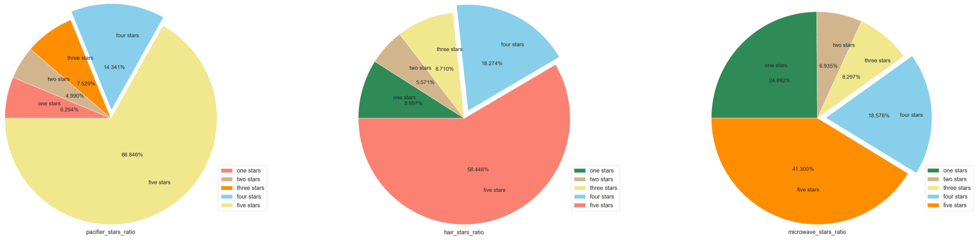
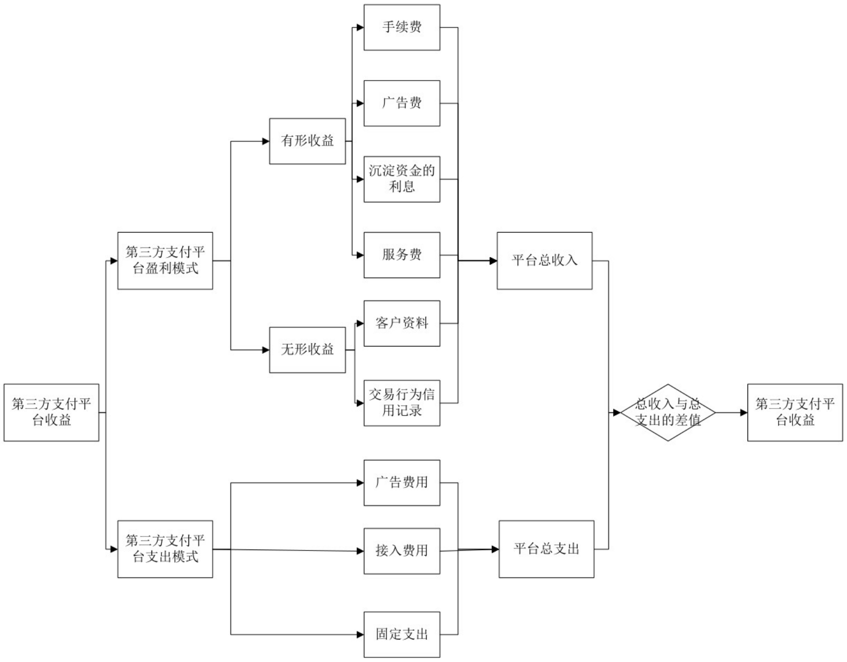


Figure 2: Image of sigmoid function



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

Figure 1: The proportion of five stars in the star_rating of each product



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 3, 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

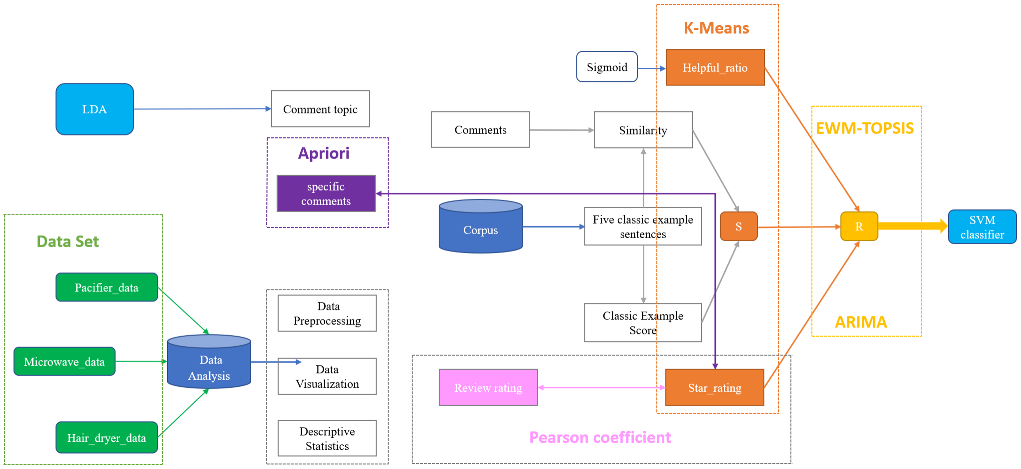


Figure 3: Workflow

the greatest similarity and the maximum similarity. Then combining satisfaction rate, star rating 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.

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.

```
import csv
import pymysql
# with open("players_stats3.csv", "r") as file:
#     reader = csv.reader(file)
#     next(reader)
#     for i in reader:
#         print(i)
db = pymysql.connect(host="localhost",
                     user="root",
                     password="271xufei.",
                     db="test",)
cursor = db.cursor()
cursor.execute("show tables");
print(cursor.fetchall())
cursor.execute("create table if not exists hh("
               "id int,"
               "name varchar(4))")
cursor.execute(f"insert table hh values({'1'}, {'chae'})")
#
# with open("players_stats3.csv", "r") as file:
#     teams = {}
#     reader = csv.DictReader(file)
#     for item in reader:
#         team = item['Team']
#         # if team in teams:
#         #     teams[team] += 1
#         # else:
#         #     teams[team] = 1
#         if team not in teams:
#             teams[team] = 0
#         teams[team] += 1
#     for team in sorted(teams, key=lambda x: teams[x], reverse=True):
#         print(team, teams[team])
#
# print("b" in {"a": "b"})
# print("a" in {"a": "b"})
```

Algorithm 1: IntervalRestriction

```
    Data: $G = (X,U)$ such that $G^{tc}$ is an order.
    Result: $G' = (X,V)$ with $V \subseteq U$ such that $G'^{tc}$ is an interval order.
    begin
        $V \longleftarrow U$
        $S \longleftarrow \emptyset$
        for $x \in X$ do
            $NbSuccInS(x) \longleftarrow 0$
            $NbPredInMin(x) \longleftarrow 0$
            $NbPredNotInMin(x) \longleftarrow |ImPred(x)|$
        end
        for $x \in X$ do
            if $NbPredInMin(x) = 0$ and $NbPredNotInMin(x) = 0$ then
                $AppendToMin(x)$
            end
        end
1       while $S \neq \emptyset$ do
REM         remove $x$ from the list of $T$ of maximal index
2           while $|S \cap ImSucc(x)| \neq |S|$ do
                for $y \in S - ImSucc(x)$ do
                    { remove from $V$ all the arcs $zy$ : }
                    for $z \in ImPred(y) \cap Min$ do
                        remove the arc $zy$ from $V$
                        $NbSuccInS(z) \longleftarrow NbSuccInS(z) - 1$
                        move $z$ in $T$ to the list preceding its present list
                        {i.e. If $z \in T[k]$, move $z$ from $T[k]$ to $T[k-1]$}
                    end
                    $NbPredInMin(y) \longleftarrow 0$
                    $NbPredNotInMin(y) \longleftarrow 0$
                    $S \longleftarrow S - \{y\}$
                    $AppendToMin(y)$
                end
            end
            $RemoveFromMin(x)$
        end
    end
```