Research on Star Rating Based on Poisson Regression

Qiang Lu[1], Longan Xiong[1], Yaoyao Liu[2]

[1]Jiluan Academy, Nanchang University, Nanchang, Jiangxi, 330031

[2]College of Information Engineering, Nanchang University, Nanchang, Jiangxi, 330031

Keywords: TF-IDF, LDA, Cosine Similarity Algorithm, python

Abstract: In this era when artificial intelligence is very popular, natural language processing has naturally attracted people's attention and has been widely used in many fields, including text analysis. In our article, we set up a model that can extract and quantify specific texts, an analysis model based on Rolling Time Windows, successfully extract and analyze various information contained in comments in online shopping, and prove that this analysis method is accurate enough and can be applied in reality by comparing with other models. As a creative way, we preprocessed the data mining, removed the incomplete parts from the given data, and then analyzed the user's comments using Term Frequency-Inverse Document Frequency (hereinafter called the “TF- IDF”) algorithm and Python-based Latent Dirichlet Allocation (hereinafter called the “LDA”) to classify and quantify them. In order to change the obtained data from discrete to continuous, we skillfully use Cosine Similarity Algorithm and linear weighting method to obtain very ideal numbers as comprehensive evaluation values. In order to further analyze the influence of user comments in it, we use the dynamic collaborative filtering recommendation model based on rolling time window, and use the model of the first point to process quantitative comments according to time series, and successfully analyze and obtain the curve of user satisfaction changing with time. From the above three graphs, we find that the comprehensive evaluation value of the hair dryer has been showing an upward trend, and the evaluation value is higher, while the comprehensive evaluation value of the microwave is lower, and shows a downward trend. So we think that the hair dryer is the potentially most successful product, and the microwave is the potentially most failed product.

1. Introduction

Born in this era of continuous technological advancement, you must have done online shopping, and most people will browse the Google"s reviews of this product before shopping and decide whether to continue buying according to the reviews [1]. Similarly, the manufacturers and the sellers of these goods also need to know the customer's satisfaction with the goods so that they can make a relatively correct decision for future development to obtain more benefits [2].

On Amazon, a nationwide well-known online shopping platform, the information contained in reviews is composed of these important parts: star_rating (1-5 star rating), helpful_votes (valid votes for reviews), vine (that is, whether the user is Amazon Member), verify_purchase (whether the user bought it at the original price), and the review text: review_body.[3] By analyzing these special combinations and data types, we can quantify the text evaluation, use weighted addition to obtain comprehensive evaluation value to represent customer satisfaction, and then obtain its time-based change rule so that the sunshine company selling goods can manufacture successful goods and obtain maximum benefits [4].

This article is based on some knowledge in the field of natural language processing in artificial intelligence, which is popular at present, to analyze and quantify the user's evaluation, and then to carry out weighted calculation on multiple indexes, in order to obtain a comprehensive evaluation value and analyze its relationship with time changes. In order to put forward reasonable development suggestions to sunshine company, we will use Python- based LDA model, TF-IDF algorithm, cosine similarity algorithm, linear weighting method, dynamic collaborative filtering recommendation model based on rolling time window, etc. to analyze user evaluation and obtain conclusions.

2. LDA Model Based on Python

The LDA model is a generative model. LDA is a matrix factorization technique. In vector space, any corpus (collection of documents) can be represented as a Document (Term-DT) matrix.

2.1 LDA document generation process

LDA assumes that documents are generated by a mixture of multiple topics, andn the process of generating each document is as follows:

(1) Generate the length of N document from a global Poisson distribution with parameter β;

(2) Generate a current document θ from the global Dirichlet distribution with parameter alpha;

(3) For each word of the current document length N: Generate a topic's subscript Zn from a polynomial distribution with θ as the parameter. Generate a word W_n from a polynomial distribution with θ and Z as parameters.

(4) These topics generate words based on their probability distribution. Given a document data set, LDA can learn which topics produced these documents.

For the document generation process, first, for each word in document n, first generate an index from θ_i in the document matrix M_1, telling us which row ϕ_m in the subject matrix M_2 is to generate the current word.

2.2 Training process (Gibbs sampling)

Gibbs Sampling first selects one dimension of the probability vector, and gives variable values of other dimensions to the value of the current dimension, and converges continuously to output the parameters to be estimated. The flow chart is as follows:

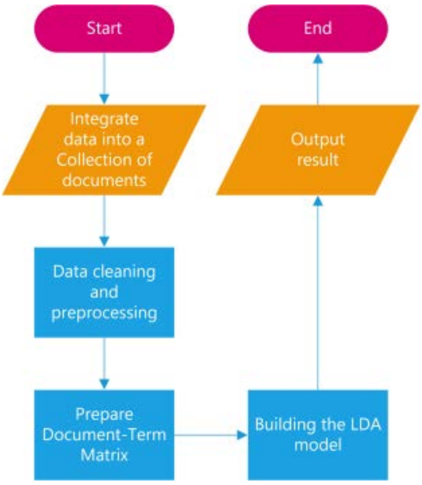
Start
End
Integrate data into a Collection of documents
Output result
Data cleaning and preprocessing
Prepare Document-Term Matrix
Building the LDA model

Fig. 1 Flow chart of LDA Running in the Python

LDA has a subject index for each word of each document. However, from the perspective of document clustering, LDA does not have a unified clustering label for documents, but each word has a clustering label, which is the theme. Each word of LDA may belong to a different category, and each document may belong to a different category. After a large number of iterations, the topic distribution and word distribution are relatively stable and better, and the LDA model converges.

3. TF-IDF in Python

3.1 TF-IDF concept

TF-IDF (Term Frequency-Invers Document Frequency) is a weighting technique commonly used in information processing and data mining. This technique uses a statistical method to calculate the importance of a word in the entire corpus based on the number of occurrences of the word in the text and the frequency of documents appearing in the entire corpus. Its advantage is that it can filter out some common but irrelevant words, while retaining important words that affect the entire text.

TF (Term Frequency) indicates how often a certain keyword appears in the entire article. IDF (Invers Document Frequency) indicates the frequency of calculating inverted text. Text frequency refers to the number of times a certain keyword appears in all articles in the entire corpus. Inverted document frequency is also called inverse document frequency, which is the reciprocal of document frequency and is mainly used to reduce the effect of some common words in all documents that have little effect on the document.

Calculation method:

(1) Calculate TF-IDF by multiplying the local component (word frequency) with the global component (inverse document frequency), and normalize the resulting document to a unit length.

(2) Word frequency (TF) = number of occurrences of a word in the article / total number of words in the article

(3) Inverse document frequency (IDF) = log (total number of documents in the corpus / number of documents containing the word + 1)

(4) TF-IDF = word frequency (TF) * Inverse document frequency (IDF)

3.2 Implementation of TF-IDF

(1) The CountVectorizer class converts the words in the text into a term frequency matrix.

(2) It uses the fit_transform function to count the number of occurrences of each word.

(3) Get_feature_names () can get the keywords of all text in the bag of words,

(4) You can see the result of the term frequency matrix by toarray ().

4. Cosine Similarity Algorithm

Similarity measure (Similarity), that is to calculate the degree of similarity between individuals, the smaller the value of the similarity measure, the smaller the similarity between individuals, the greater the value of similarity, and the greater the difference between individuals.

A good way to calculate the similarity between several different texts or short text dialogue messages is to map the words in these texts to vector space to form the mapping relationship between the words in the texts and vector data, and calculate the similarity of the texts by calculating the difference between several or more different vectors. The following is a detailed and mature vector space cosine similarity method to calculate similarity.

Cosine similarity measures the difference between two individuals by using the cosine of the angle between two vectors in vector space. The closer the cosine value is to 1, the closer the included angle is to 0 degrees, that is, the more similar the two vectors are, which is called "cosine similarity".

5. The relationship between time and evaluation

5.1 Polynomial Fitting Time and Comprehensive Evaluation Value

For time, we did the following:

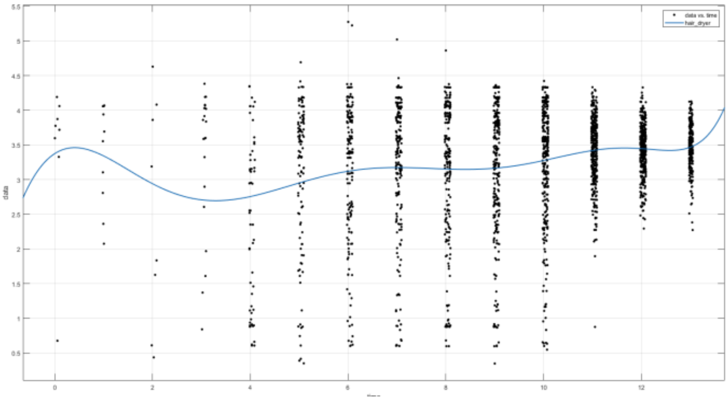

Fig. 2 comprehensive evaluation of hair_dryer

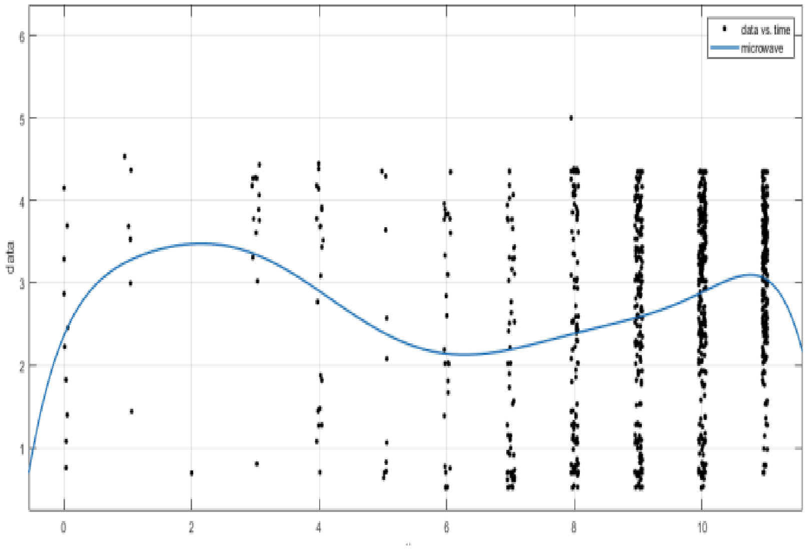

Fig. 3 comprehensive evaluation of microwave

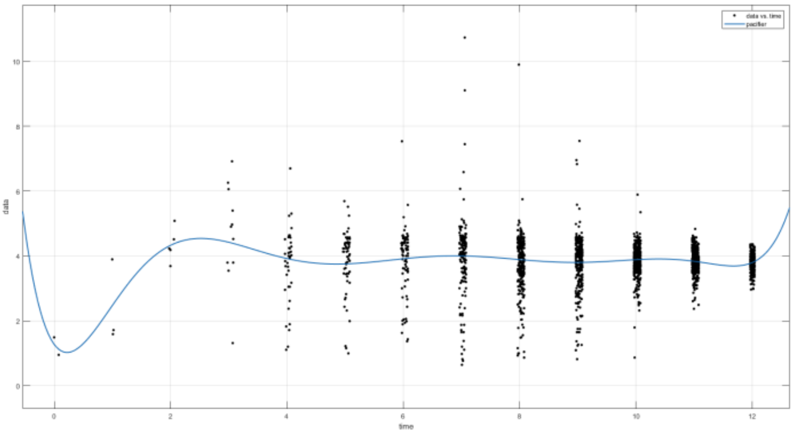

Fig. 4 comprehensive evaluation of pacifier

From the above three graphs, we find that the comprehensive evaluation value of the hair dryer has been showing an upward trend, and the evaluation value is higher, while the comprehensive evaluation value of the microwave is lower, and shows a downward trend. So we think that the hair dryer is the potentially most successful product, and the microwave is the potentially most failed product.

6. Dynamic Collaborative Filtering Recommendation Model and Algorithm Based on Rolling Time Window

The model algorithm processes interest scores at different times according to time series. The customer's emotional similarity is composed of components from different time periods, which improves the timeliness of the algorithm. It considers time series and can analyze the the sentiment contained in the time series and customer reviews, which facilitates us to analyze the association of comments before and after the time node.

uu closed plane is found in the dynamic model and mapped into *uu* items-time window two-dimensional interest degree matrix. Determine the similarity between the two users by calculating the distance between the *uu* two-dimensional matrices and then according to the similarity eventually, the nearest *KK* neighbors of each user can be found, that is, the *KK* users with the highest similarity with themselves. Finally the corresponding items are recommended to the target users according to the common interests of the *KK* users. When the time window is scrolled once, only the item-time window two-dimensional interest matrix needs to be shifted in columns and the latest column is updated, and then use the incremental algorithm to calculate the similarity after scrolling and obtain the recommended result.

Applying TensorFlow and a recurrent neural network (RNN) using LSTM units, we have obtaind some new quantized text data (partial):

Table. 1 some new keyword quantification data

Keywords	Quantified data	Keywords	Quantified data
great	0.602	works well	0.070
excellent	0.063	perfect	0.281
like	0.357	nice	0.223
good	0.411	happy	0.105
glad	0.013	best	0.074
useful	0.023	trash	0.005
don't buy	0.041	unsafe	0.012
bad	0.173	broke	0.181
can't	0.193	junk	0.082
useless	0.018		

We further calculated the similarity between the calculation results of the two methods, and the result was over 85%, so we have good reasons to believe that our method can be used in reality.

7. Conclusion

In our article, the construction of the IDA model based on Python and the implementation of the TF-IDF algorithm are emphasized. This is the key to convert text data into numerical values. Then, the cosine similarity algorithm is used to continue the numerical values. The weights of various indicators are calculated, and the comprehensive evaluation value is obtained after weighted summation.

Combining the comprehensive average with a reasonably processed time series, we analyzed and obtained that the specific stars will affect the subsequent reviews to different degrees, and the average correlation between reviews with a specific emotional color and rating level is 0.67. The comments after the node will be affected to varying degrees by the comments before the time node.

Like any model, the one present above has its strengths and weaknesses. Some of the major points are presented below.

References

[1] Zhuo Jinwu, Wang Hongjun. Methods and Practice of MATLAB Mathematical Modeling (Third Edition) [M]. Beijing: Beijing University of Aeronautics and Astronautics Press, 2018.

[2] Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python [M]. California: O’Rilly Media, Inc.2009.

[3] Getting started with the topic model LDA, https://blog.csdn.net/selinda001/article/details/80446766

[4] Wang Junkui. Research on the Usefulness of Online Reviews of E-commerce Websites [D]. Xidian University, 2014.