Semantic Analysis of Massive Text under Multi-Model Strategy

Zekun Tao[1,a], Youwei Zhang[1, b,*], Feiyue Fang[1,c], Jing Li[2,d], Chuanwei Lu[2,e], Hongjian Wu[3,f]

[a]Zekun Tao: Zekun.tao@outlook.com

[b]Youwei Zhang*: wei_zhangyou@163.com

[c]Feiyue Fang: fangfei_yue@yeah.net

[d]Jing Li: 351800040@qq.com

[e]Chuangwei Lu: cw2033_lu@163.com

[f]Hongjian Wu: hongjianw@163.com

[1]Zhengzhou Xinda Institute of Advanced Technology, Zhengzhou, 450052, China;

[2]PLA Strategic Support Force Information Engineering University, Zhengzhou, 450052, China;

[3]Zheng Shu Network Technology Co., Ltd, Zhengzhou, 450052, China;

Abstract. The comment text generated by tourists' travel is one of the core contents of the research on semantic analysis of tourist destinations. Considering the phenomenon of fake reviews, simple copying, worthless information and irrelevant content, it prevents tourists from obtaining valuable information from online reviews. In this paper, the analysis of tourist reviews based on a multi-model fusion of natural language processing can solve the understanding problem with online reviews, and realize the analysis on characteristics of tourist destinations after machine processing. The method in this paper is experimentally verified on the data of question C in the ninth "Teddy Cup" Data Mining challenge, and the effective text is extracted for analysis of characteristics. It provides research ideas and methodological support for exploring the effectiveness and characteristic analysis of the text.

Keywords: DBSCN; TF-IDF; NLP; Word2Vec

1 Introduction

In 2021, China Internet Network Information Center (CNNIC) released the 47th *Statistical Reports on Internet Development in China* (hereinafter referred to as the report) in Beijing. The report shows that, as of December 2020, the number of Internet users in China had reached 989 million, an increase of 85.4 million compared with March 2020, and the Internet penetration rate reached 70.4% [1].

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322807

Although the virtual world is considered to replace the tourism industry in the future, tourism is still exceedingly popular, especially during public holidays. Once, hotels usually set up message books to record tourists' comments on their destinations. Now, people are getting used to arranging their holiday travel activities through travel apps because of the rapid development of mobile apps. Tourists also have a more convenient way to write comments – online comments. In this way, it is convenient for tourists to express their emotions, and beneficial for hotel managers to improve their services.

With the popularization of the travel apps, the rapid processing of a large number of complex tourist reviews has become an urgent issue to be solved. It is an irresistible trend to introduce Big Data Technology into the tourism industry, in order to optimize its efficiency of service, management and supervision. The analysis of tourist comments based on natural language processing in artificial intelligence technology will avoid the problem of "worthless text", and improve management efficiency. Therefore, the TOP20-hot words of the destination are extracted in this paper by mining the reviews of hotels, and formed an impression analysis of the hotels. Also, the validity of the online comment texts is analyzed, and a model is established to analyze the characteristics of the hotel.

2 Framework of research

The framework of overall research is shown in Fig.1, the detailed steps are as follows:

- Pre-processing of the raw data;
- Using the weight fusion of multiple models to remove "duplicated" texts (The duplicated texts here are point not only the same text, but also similar texts);
- Using TF-IDF, LDA model and other technologies to extract the features, and then visualize the features by knowledge graph.

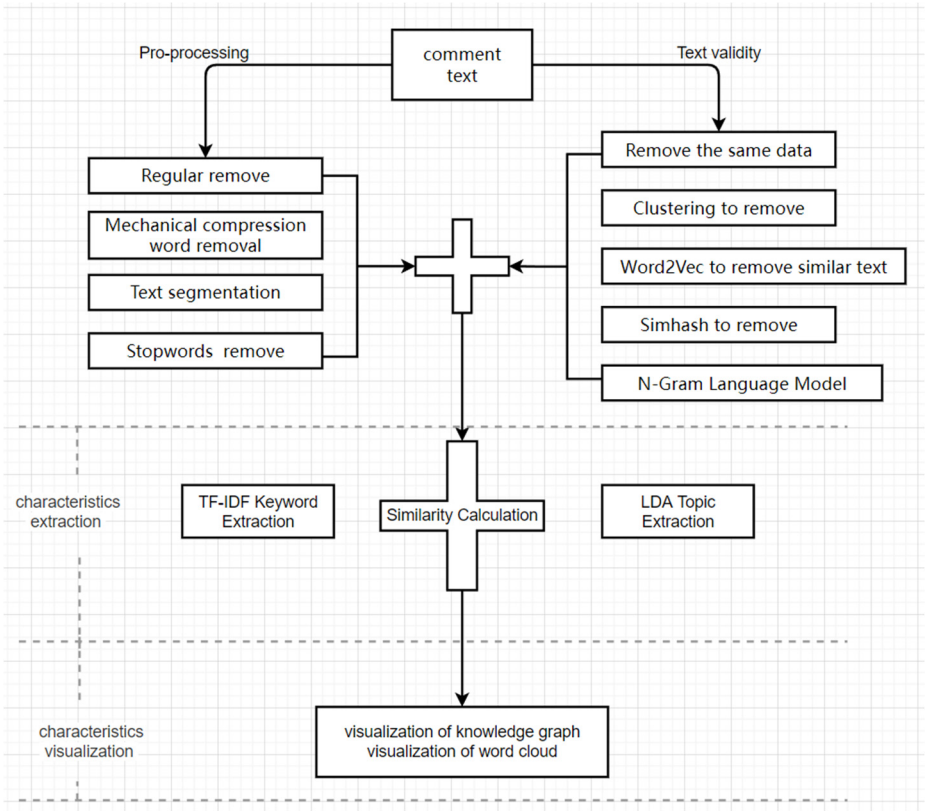

Fig. 1 Framework of overall research

2.1 DBSCAN

The Clustering method is an unsupervised learning technology, which does not rely on prior knowledge. It partitions the data according to their similarity. Cluster analysis can deal with various types of massive data to find hidden patterns, unknown relationships and other potentially useful information.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a typical density clustering algorithm [2]. Compared with K-Means and BIRCH, which are generally only applicable to convex sample sets, DBSCAN can be applied to both convex and non-convex sample sets.

Using DBSCAN density clustering, there is a data set $D = \{x_1, x_2, \ldots, x_m\}$, the related density concept description of DBSCAN and cluster shape are as follows Fig. 2:

1) $\varepsilon - \text{neighborhood}$: for any $x_j \in D$, its $\varepsilon - \text{neighborhood}$means a sub-sample set that includes points whose distance from x_j is not greater than ε in the sample set D. That is, $N_\varepsilon(x_j) = \{x_j \in D | distance(x_i, x_j) \leq \varepsilon\}$, the number of samples in this sub-sample set is recorded as $|N_\varepsilon(x_j)|$.

2) Core object: for any sample $x_j \in D$, if its $\left|N_\varepsilon(x_j)\right| of\ \varepsilon -$ neighborhood at least Min_$samples$, that is $\left|N_\varepsilon(x_j)\right| \geq$ Min_$samples$, is a core object.

3) Directly Density-Reachable: If x_i is in the $\varepsilon -$ neighborhood of x_j, and x_j is the core object, x_i is Directly Density-Reachable from x_j. Note that the opposite the causality is not necessarily true: x_j is Directly Density-Reachable from x_i, unless and x_i is also a core object.

The DBSCAN algorithm divides data points into three categories:

1) Core points: Points containing more than Min_$samples$ within radius ε.

2) Boundary points: The number of points within radius ε is less than Min_$samples$, but falls within the neighborhood of the core point.

3) Noise Points: Points that are neither core points nor boundary points.

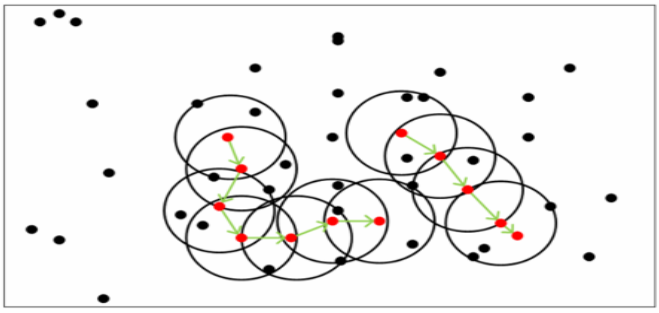

Fig. 2 Cluster shape

2.2 Word embedding based on Word2Vec model

Word2Vec is an efficient algorithm model that uses deep learning to represent words as valued vectors. It simplifies the text content processing into vector operations in K-dimensional vector space, and the semantic similarity of text is converted into the similarity of space vectors [3]. The core idea is to use training to map each word into a K-dimensional vector, and then judge the semantic similarity between them by the distance between the vectors. Word2Vec can be implemented by two methods: CBOW model and Skip-gram model, which are designed based on Hierarchical SoftMax and Negative Sampling respectively. Both models contain three layers: input layer, mapping layer and output layer.

This paper uses the Word2Vec model to vectorize reviews of hotels. The model can express contextual information of the word and the internal structure of the sentence, and the words are mapped into a low-dimensional, dense vector based on the neural network.

2.3 Similar words remove based on Simhash

The basic idea of Simhash is to transform the text into fingerprints, and then identify the text similarity by comparing the fingerprints. Most of the traditional methods of comparing text similarity are to convert text into a measure of feature vector distance, such as Euclidean

distance and Hamming distance. However, excessive resources are consumed in their calculations because of the weakness of algorithm in aspect of large dimension [4].

2.4 N-Gram Language Model

The N-gram model is a probability-based discriminant language model. Its input is a sentence (sequential sequence of words), and the output is the joint probability of a sentence [5]. PPL is an indicator used in the field of natural language processing (NLP) to measure the quality of language models. It mainly estimates the occurrence probability of a sentence based on each word, and uses the sentence length for normalization. The smaller the PPL means better language model. The formula (1) is as follows:

$$pp(s) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(\mathrm{w_i}|w_1 w_2 \dots w_{i-1})}} \tag{1}$$

2.5 Keywords extraction based on TF-IDF

TF (Term Frequency) represents the frequency of a word in a piece of text. The higher value of TF means a high frequency of word in a specific document; IDF (Inverse Document Frequency) is the total number of documents divided by the number of documents that contain the word, and then take the base 10 logarithm of the above result. The higher value of IDF means that fewer texts contain the word, and the word has a stronger representation. The formula for TF-IDF is as follows:

1) TF (Term Frequency)

TF represents the frequency of a word in the text, formula (2):

$$\mathrm{TF_{ij}} = \frac{\mathrm{n_{ij}}(\text{number of occurrences of a word in a text})}{\sum_k \mathrm{n_{k,j}}(\text{the number of total words in a text})} \tag{2}$$

2) IDF (Inverse Document Frequency)

The IDF value of a word is calculated by dividing the total number of documents by the number of documents containing the word, and then taking the logarithm of the obtained quotient. The fewer documents containing the term t and the larger IDF means that the term has a clear distinguishing ability between categories. ($\left|\{\mathrm{j}: \mathrm{t_i} \in \mathrm{d_j}\}\right| + 1$ is to prevent the word from not being in the text set so that the denominator is 0), formula (3)

$$\mathrm{IDF_i} = \log \frac{|\mathrm{D}|(\text{total number of texts})}{|\{\mathrm{j}:\mathrm{t_i} \in \mathrm{d_j}\}| + 1(\text{number of text containing the word})} \tag{3}$$

3) TF-IDF

A high-weight TF-IDF is generated by combining the high word frequency in the target text and the low text frequency of the word in the text set. Thus, TF-IDF tends to filter out common words and leave words with high importance, formula (4)

$$\mathrm{TF}-\mathrm{IDF}=\mathrm{TF}\times\mathrm{IDF} \tag{4}$$

2.6 LDA Topic Model

In 2003, Blei proposed the famous LDA (Latent Dirichlet Allocation) topic model, which is a probabilistic model for document topic generation. It mainly adds a Bayesian framework layer to the probabilistic Latent Semantic Analysis (pLSA) model. Its main goal is to use unsupervised learning methods to extract hidden topic information in a large number of documents, so that help readers quickly understand information of the document [6].

The essence of LDA topic model is to use the collinear features of text feature words to extract text topics. It has three clear layers, which are document layer, topic layer and feature word layer.

The main idea of LDA is as follows: the document set is based on the probability distribution of topics, and the topics are based on the probability distribution of feature words. Therefore, there is a probability formula (5), which expresses the probability of occurrence of word Wn in document Mm:

$$p(w_n|M_m)=\sum_{k\in K}p(w_n|K_k)p(K_k|M_m) \tag{5}$$

In the above formula, N is the total number of feature words, M is the number of documents, K is the total number of topics. Also, $n\in N$, $m\in M$, $k\in K$.

The LDA topic document generation process is shown in Fig. 3 below. For the "document-topic" multinomial distribution process, it is necessary for the distribution to obey the Dirichlet prior distribution with parameter α (the Dirichlet distribution is a multi-dimensional Beta distribution). The following formula (6) are the density functions of the Beta distribution.

$$f(p;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}P^{\alpha-1}(1-p)^{\beta-1} \tag{6}$$

The $B(\alpha,\beta)$ represents the Beta distribution with parameters (α, β), and p represents the probability of the event occurring. The K-dimensional Dirichlet distribution is shown in formula (7).

$$Dirichlet(\bar{p}|\bar{\alpha})=\frac{t(\sum_{k=1}^{K}\alpha_k)}{\prod_{K=1}^{K}t(\alpha_k)}\prod_{K=1}^{K}P_k^{\alpha_k-1} \tag{7}$$

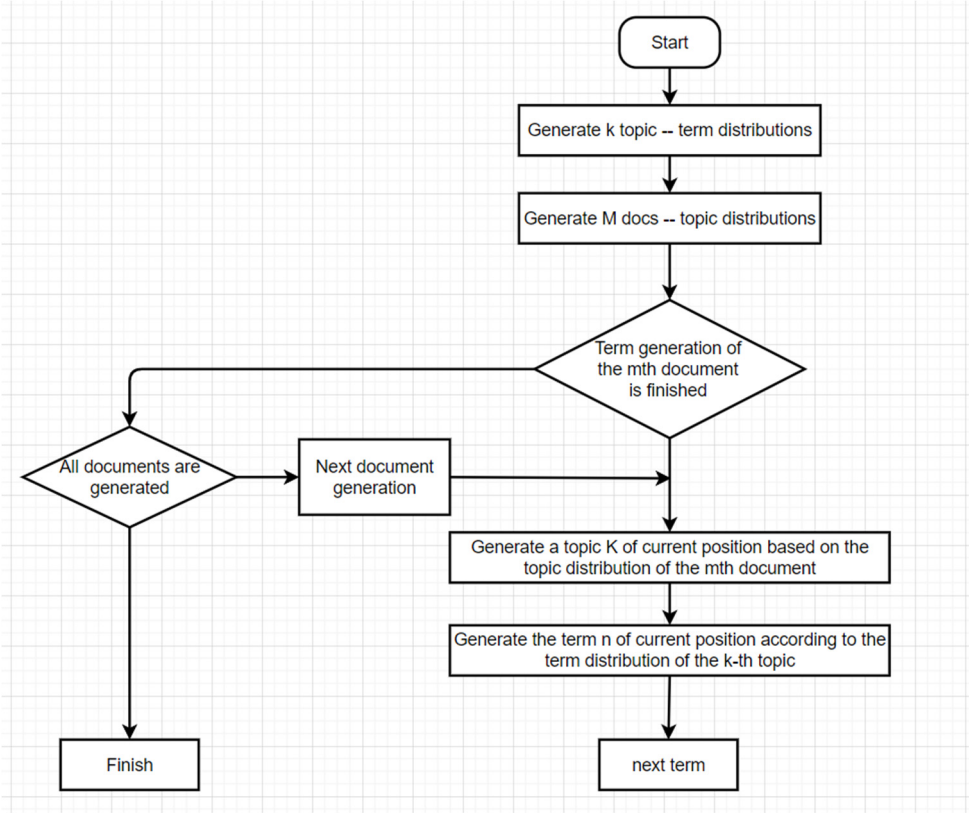

Fig. 3 LDA topic documents generation flow diagram

It can be seen that the special form of the Dirichlet distribution in the two-dimensional state is the Beta distribution, and the LDA model will determine the parameters (α, β). Each parameter is shown in Table1

Table1 Meaning of symbols in LDA model

Symbols	**Meanings**
∅	Terms distribution, ∅ $\sim Dirichlet\ (\alpha)$
θ	Topic distribution, θ $\sim Dirichlet\ (\beta)$
α	The prior distribution (Dirichlet distribution) parameters of the topic distribution θ
β	Prior distribution parameters for word distribution ∅.
N	The total number of feature words in the document, n∈N
M	Total number of documents, m∈M
K	Total number of topics, k∈K

The training process of the LDA topic model is mainly to train the parameters α and β, whose typical representatives are EM estimation and Gibbs sampling.

3 Experiments and Analysis

3.1 Experimental data sources and processing

The data in this paper comes from the hotel comments text data about tourist destination impressions, it is provided by the 9th National "Teddy Cup" Data Mining Challenge, C question. There are a total of 25,000 comments data, which involves 50 hotels. The data format is shown in Fig. 4 below:

	Name	Time	Content	Type
0	H01	2020-01-01	The hotel is great for families	standard room
1	H01	2020-01-01	Upgraded rooms and late check-out is great	standard room
2	H01	2020-01-01	In recent years, I come to Guangzhou every yea...	standard room
3	H01	2020-01-01	The hotel is very nice	standard room
4	H01	2020-01-01	Super five-star reviews	superior room

Fig. 4 Data Format

3.2 Characteristics analysis under the validity of tourism texts

After text pre-processing, it is necessary to deduplicate and cluster the text, and then calculate the frequency of each word in each comment in the corpus composed of all comments, so as to convert the unstructured text into a structured vector. Set $Min\_samples = 2,\ \epsilon = 0.9$ to cluster the data, and keep the longest one from the cluster, and delete the rest. The free data is also completely reserved. The first model data is obtained.

The inputs of second model data needs to be pre-processed again, and the vector after word segmentation is trained by the Word2Vec model for hotel reviews. According to the similarity calculation formula, judge the same sentence, deduplicate it in the original data set, and then get the second model data.

The Simhash algorithm uses the TF-IDF weight in the jieba library to process the result after the second step of word segmentation. Perform ordinary hash operations on the acquired features, and calculate the hash value, so that a binary length of n bits is obtained. It is a set as (hash: weight). On the basis of the obtained hash value, weight is performed according to the corresponding weight value (W=hash*weight). That is, if the hash is 1, it is multiplied by the positive weight. Otherwise, it is multiplied by the negative weight [7]. Then, the weighted results of each vector obtained above are summed to form only one sequence string. Judging each value of the accumulated result of the n-bit signature. If it is greater than 0, the value is set to 1, otherwise it is set to 0, so as to obtain the simhash value of the statement. In this way, the simhash value of a document is obtained. Finally, the similarity is judged according to the Hamming distance of the simhash values for different sentences. Perform similarity deduplication on the original dataset to obtain the third model data.

Finally, the N-Gram language model is used to calculate the sentence probability. Arrange its perplexity PPL of each sentence in order from small to large, and then take the data with the quartile below 75% to get the fourth model data.

The four models are synthesized, and the combination strategy in model fusion is adopted [8]: voting algorithm. Vote the datasets in models 1-4 to select relatively clean data. At this point, the data has reached a critical value and cannot be deduplicated, otherwise an error operation will occur.

After filtering out the efficient comment set, it is more effective to use topic keywords instead of original comments to analyze semantics and find hotel characteristics. Then, the similarity calculation idea is used to select the optimal words for the results of LDA topic model and TF-IDF weight extraction. The flowchart is shown on Fig.5.

The first step is the word cloud image of the weighted keywords of the TF-IDF algorithm. The hotel word cloud image of hotel A01 is shown in Fig. 6.

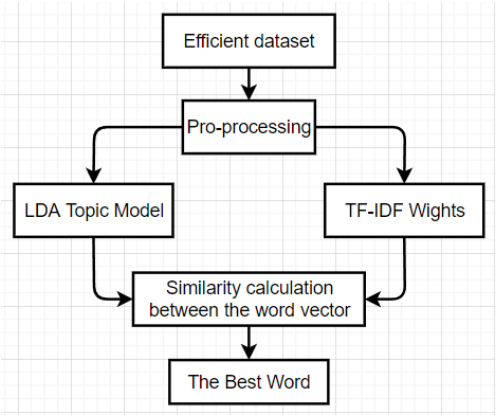

Fig. 5 Word vector similarity calculation

Fig. 6 Word Cloud of Hotel A01

The second step is LDA topic model extraction, using stop words and sentiment evaluation tables provided by BosonNLP data to construct a dedicated dictionary corpus for topic classification of the LDA model.

The words in the word matrix are scored according to the sentiment evaluation table. The "greater than 0 is positive, less than 0 is negative" rule differentiates the datasets. The corpora. Dictionary () function in the gensim library is used to process the desired dictionary, and the doc2bow () word bag technology converts the dictionary into the desired corpus. Finally, the subtopics are divided into three categories according to the corpus, and each category has positive and negative sides.

The third step is to vectorize the resulting keywords and subject words (arranged in order of weight) that from the first and second steps. Calculate the distance between the keyword vector and the topic word vector for each comment, and output the words with high results according to the distance. Here cosine similarity is used as the calculation method. A smaller angle means more similarity. The closer the cosine is to 1, the closer the angle is to 0, the more similar the two vectors are. The formula (8) is as follow:

$$cos\theta = \frac{\sum_{i=1}^{n} X_i \times Y_j}{\sqrt{\sum_{i=1}^{n} X_i^2}\sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{8}$$

The fourth step is to construct triples based on the combination of subject, keyword, hotel name, and hotel type. Converting the triple data of the neo4j database into SQL statements using the py2neo library, and uploading them to the neo4j data. Finally, a knowledge graph is generated, and as shown in Fig.7:

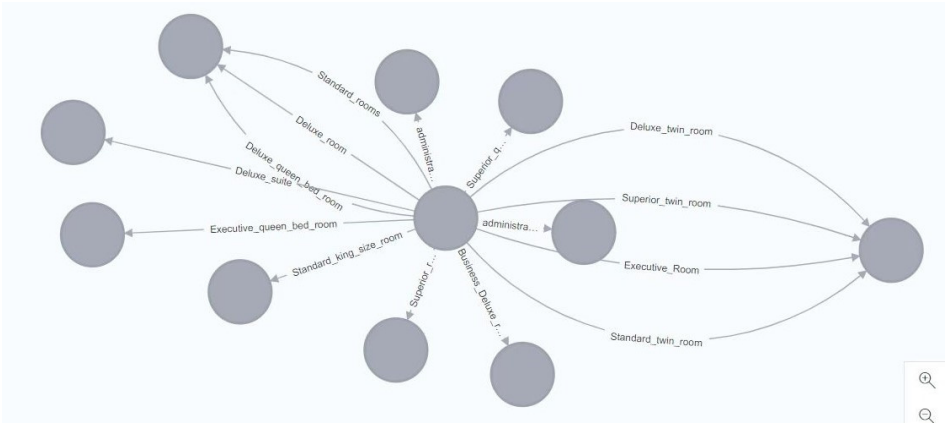

Fig. 7 Knowledge Graph

4 Conclusion

Due to the popularity of the Internet and the explosive growth of data, how to process data and obtain valuable information from data is of great significance. The core of this paper is to meet the mining and analysis requirements of a large number of tourist comment texts by constructing reasonable computational weights and rigorous models in the field of Internet cross-tourism.

Using the hotel review text set as the data source, natural language processing technology is used to provide assistance for the hotel's strategic management and development. The following conclusions are drawn:

Based on model fusion, the results of density clustering models, including Simhash algorithm model, Word2Vec model and N-Gram model, are voted to select a representative, authentic, and effective review text dataset, so that reduce the cost of information retrieval, improve the quality of comments, and enhance the vitality of the comments. Then the text is visually analyzed according to LDA topic model, knowledge graph and other technologies. Finally, each topic, weighted keywords and nodes of the knowledge graph are the characteristics of the text. However, the experiment in this paper has certain limitations because of using a single data source. In the future, the fusion of multimodal data will be considered for more comprehensive features.

References

[1] http://www.gov.cn/xinwen/2021-02/03/content_5584518.html

[2] Zhu Helong. Research on Performance Optimization and Parameter Selection of Density Clustering Algorithm [D]. Jiangxi University of Science and Technology,2020. DOI:10.27176/d.cnki.gnfyc.2020.000646.

[3] Adewumi Tosin, Liwicki Foteini, Liwicki Marcus. Word2Vec: Optimal hyperparameters and their impact on natural language processing downstream tasks[J]. Open Computer Science,2022,12(1).

[4] Jyoti Malhotra, Jagdish Bakal. Grey wolf optimization based clustering of hybrid fingerprint for efficient de-duplication[J]. Multiagent and Grid Systems, 2018, 14(2).

[5] Pooja Kherwa, Poonam Bansal. Semantic N-Gram Topic Modeling[J]. EAI Endorsed Transactions on Scalable Information Systems,2020,7(26).

[6] Cheng Haiqi. Short Text Topic Mining of Hotel Comments Based on Emotional Classification [D]. Zhejiang Gongshang University,2020. DOI:10.27462/d.cnki.ghzhc.2020.000516.

[7] Zhou Rujia. Research on Text Deconduplication Method Based on Semantic Fingerprinting and SimHash [D]. Jiangxi University of Finance and Economics, 2021.DOI:10.27175/d.cnki.gjxcu.2021.001236.

[8] Lin Zhixing, Wang Like, Cui Xiaoli, Gu Yongxiang. Fast Sentiment Analysis Algorithm Based on Double Model Fusion[J]. Computer Systems Science and Engineering,2021,36(1).