Publication Date: 15 May 2024
Archs Sci. (2024) Volume 74, Issue 2 Pages 150-158, Paper ID 2024221.
https://doi.org/10.62227/as/74221

Research on the Value and Language Features of Chinese Language and Literature Texts Based on Text Mining Technology

Yiyi Ru[1,*]
[1]Zhengzhou College of Finance and Economics, Zhengzhou, Henan, 450000, China.
Corresponding authors: (e-mail: 13939086290@163.com).

Abstract The linguistic characteristics of a literary work are the way of thinking embodied in the author's use of language. From the textual value of Chinese language literature, this paper analyzes the spiritual connotation of Chinese language literature from two dimensions: reading and education. Based on the web crawler technology, we obtain the text data of Chinese language literature from three writers, Bajin, Yu Zheng and Qiong Yao, preprocess the data through data cleaning, Chinese word segmentation, de-duplication, etc., and extract the feature values of the text by using the TF-IDF algorithm. Then the text documents are mapped onto vectors using the VSM model, and the parameters of the LDA topic model are estimated by the Gibbs sampling algorithm in order to better obtain the topic changes of the Chinese language literature texts. This paper carries out linguistic feature verification from the lexical and similarity features of Chinese language literary texts. It is found that the difference in lexical density between Ba Jin's Cold Night and Resting Garden is only 2.1 percentage points, and the frequency of the verb "to say" is 1,213 times and 735 times respectively. The average sentence lengths of Yu Zheng and Qiong Yao fluctuate within the range of [18.49,34.27], and Qiong Yao's works have a higher thematic concentration than Zheng Zheng's works. Analyzing the linguistic features of Chinese language literary texts based on text mining techniques helps to understand the authors' language usage methods and helps to promote innovative expression paths in literary texts.

Index Terms tf-idf algorithm, web crawling techniques, Gibbs algorithm, LDA topic model, VSM model, literary text

I. Introduction

Text mining refers to the process of extracting the required information or knowledge from a large amount of unstructured text data and transforming it into structured data [1], [2].

Chinese Language and Literature is a discipline set up by colleges and universities to help students learn Chinese language and Chinese literature [3], through the study of Chinese Language and Literature, students can master the basic knowledge of the specialty as well as news, history, art, philosophy and other aspects [4], in addition to having the ability to read canonical books, scientific research, practical work and so on [5], and understand the most cutting-edge achievements of colleges and universities as well as the development prospects. Chinese literature is a part of traditional culture [6], the design and development of its course content needs to use traditional culture as the basic theory and ideological support, and in this process, through the teaching of the course and indirectly play a role in publicizing traditional culture, invariably become a propaganda carrier [7].

On the other hand, China's relevant laws clearly pointed out that the content of Chinese language and literature courses in colleges and universities must be integrated into the content of excellent traditional culture, and produce positive education for students, and through the process of education gradually call on more students to actively participate in the cause of promoting traditional culture, and further promote the development of Chinese language and literature majors [8], [9].

There are two major characteristics of the discipline construction of Chinese language and literature majors in undergraduate universities: first, the discipline has a trapezoidal structure, that is, the discipline has a deep basic platform but does not deliberately pursue high-end scientific research projects and scientific research achievements, and second, it does not take the narrow concept of "literature" in research universities, but adopts the broad concept of "literature". Due to the dual characteristics of "symmetry" and "asymmetry" between the major and the profession of Chinese language and literature in applied universities, the positioning of the cultivation of this professional talent is neither "subject expertise" nor "professional craftsmen" of higher vocational colleges, but "vocational professionals", and the cultivation of "harmonious

development of people", not "useful tools" [10].

Literature [11] proposes a framework that integrates citation analysis and text mining using academic papers and patents as data resources to monitor the evolutionary path of nanogenerator technologies and predict their trends. Literature [12] provides a comprehensive review of telecollaboration practices over the past two decades and focuses on five key themes that have emerged in the telecollaboration literature, namely patterns, tasks, challenges, technologies, and emerging trends. Literature [13] utilized a web-based survey to conduct a comprehensive evaluation of the new curriculum reform of Chinese language and literature to conclude that there are social and political standards in the curriculum goal setting of Chinese language and literature, and that the cultivation of creative ability is neglected. Literature [14] puts forward the improvement measures in the teaching process of Chinese language and literature, starting from enriching the teaching content, clarifying the teaching plan, innovating the teaching method, improving the assessment method and so on, to fully stimulate the enthusiasm of students to learn Chinese. Literature [15] analyzes the optimization of teaching Chinese language and literature. Multimedia-assisted teaching of Chinese language and literature can promote the teaching process and tap students' inner knowledge blind spots. Literature [16] discusses some of the challenges facing humanities and social sciences-CELJ publishing, and highlights some of the contradictions of internationalization in the Chinese context. Literature [17] proposes a quantitative method for communication methods and communication effects, and further analyzes the role of national culture in cultural communication.

In this paper, we firstly sorted out the text value of Chinese language literature, mainly from the reading value and educational value of literary texts, and explored the specific role of the text value of Chinese language literature on the growth of readers' ideology. Secondly, we utilize web crawler technology to obtain the literary text data of Ba Jin's Cold Night and Rest Garden, and also crawl five classic literary works of Yu Zheng and Qiong Yao, and then combine the TF-IDF algorithm with the data pre-processing to extract the eigenvalues to establish a Chinese language literary text corpus. Then the vector mapping of the literary text data is carried out by the vector space representation model, the themes of the literary texts are analyzed by combining with the LDA theme model, and the parameters of the LDA theme model are estimated by using the Gibbs sampling algorithm. Finally, for the linguistic features of Chinese language literary texts, the exploration was carried out in four dimensions, namely, lexical density, word frequency features, sentence similarity and text topic concentration, in order to deeply analyze the changes of linguistic features of different Chinese language literary texts.

II. Exploring the Value of Chinese Language and Literature Texts

The development of Chinese language literature text can not be separated from reading education, the text value contained in Chinese literature can better cultivate readers' social values, literary aesthetic ability, and build a bridge of spiritual communication between readers and writers. It promotes the readers to deeply understand the spiritual core of Chinese language literature texts, enhances the readers' literary literacy ability, and provides support for the dissemination of Chinese language literature texts.

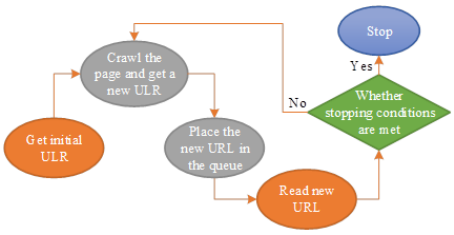

Figure 1: Reading value of literary text

A. Reading Value of Literary Texts

For the value of Chinese language literature text, which is essentially a value activity about education, only from the value level clarification, adjustment, so as to determine the reasonable value presuppositions based on sound value rationality, the reading of Chinese language literature text can be released to the maximum extent. For the reading value of Chinese language literature text, its main performance is shown in Figure 1, including the original value of the text, the value of teaching materials and teaching value of three dimensions. The original value of the literary text is the information value that exists as a separate social reading individual, the value of the teaching material is the value of creatively retaining and adding value to part of the original value of the Chinese language literary text, and the teaching value is the highest value that is based on the original value and the value of the teaching material.

B. Educational Value of Literary Texts

The educational value of Chinese language literature texts is based on the education of Chinese national culture and tradition and socialist education, helping students establish positive ideological values and cultivate patriotism. In the ever-changing development of the times, it is necessary to increase the exploration of the educational value of Chinese language literary texts from the purpose of reading for all. As far as Chinese language literature text is concerned, its educational value is to lead readers to read good books and draw nutrition from literature. The rich social cognitive value, literary aesthetic value and special political and ideological educational value of Chinese language literary texts have an irreplaceable educational role in the growth of readers' ideological concepts.

(1) Social cognitive value of Chinese language literature texts

The social value of Chinese language literature refers to the value significance of literature as a social phe-

nomenon and ideology to the whole society. The social value of literature is a value system running through literary symbols. Literature is an important content of human spiritual activities, but due to the participation of its complex individual psychology and socio-historical factors, it will be presented with different value meanings.

(2) Literary Aesthetic Value of Chinese Language Literature Texts
Chinese language literature is a typical nationalized and popularized cultural product, and it is a rare artistic model with both quality and quality, whose grand artistic expression and lofty ideal spirit provide readers with artistic aesthetic evaluation standards and bring them a rich aesthetic experience. The process of expression and ideology are immersed in each other and penetrate into each other, with ideology infiltrating aesthetics and ideology being conveyed through aesthetics.

(3) Ideological Education Value of Chinese Language Literature Texts
The ideological education of Chinese language literature is the most important value of Chinese language literature, and the rich ideological education value contained in the text, such as the spirit of nationalism, revolutionary heroism, idealism, and the spirit of enterprising and striving, etc., is crucial to the ideological and moral construction of the readers.

III. Chinese Language Literature Text Corpus and Text Mining

The text mining of Chinese language literature helps to understand the development of literature, explore the language characteristics between different Chinese language literary texts can understand the writing habits of literary writers, and provide a new direction for the creation of innovative literary texts. With the development of modern technology, relying on text mining technology to realize the depth of the Chinese language literature text becomes possible, so that the value of the literary text can be more intuitively displayed in front of the readers, thus helping the readers to more effectively grasp the connotation of the literary text, and enhance the spiritual communication between the readers and the writers.

A. Creation of a Corpus of Literary Texts

1) Web Crawling Method to Obtain Data

In the era of big data, with the rapid increase of network information resources, the traditional search engine has been unable to meet the needs of some people for information acquisition, web page as a carrier of massive information, has become the main source of people to obtain information, however, through the manual way to obtain a large amount of information in the web page, it is usually both time-consuming and laborious, therefore, the web crawler technology is particularly important. Web crawler, also known as web

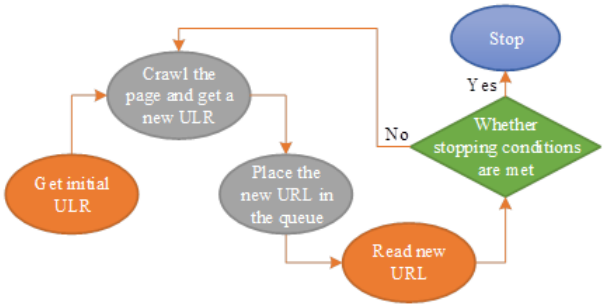

Figure 2: Flow chart of web crawler technology

information collector, is a program or script that automatically crawls web page information according to certain rules, and is an important part of search engine. Web crawler technology process shown in Figure 2, usually used generally for the whole network crawler, which can crawl the entire web page data, and save it to the appropriate database. The working principle is mainly from the pre-set URL, first crawl the URL list of the initial web page, the HTML markup on its web page to get the required data, so as to download the information into the storage. At the same time, a new list of URLs is extracted and added to the queue to be analyzed, and the above steps are repeated until the system stopping conditions are met.

In this paper, we will use general-purpose web crawlers to collect data on two literary works of Ba Jin's Cold Night and Rest Garden, and the literary works of two authors, Yu Zheng and Qiong Yao, in order to carry out linguistic characterization of different literary works. The specific steps of using web crawler technology to carry out text data acquisition of Chinese language literature are as follows:

(1) Determine the content of the Chinese language literary text to be crawled and the time node.
(2) Use the advanced search function of the browser to input keywords and time nodes, start parsing the initial URL, and download the current webpage information into the storage.
(3) Set up page-flip collection, i.e., after the information collection of the current page is completed, page-flip to get the new URL and parse and download the information of the current page.
(4) Repeat the above steps to complete the page-flip information acquisition and store all the data.

2) Pre-Processing of Literary Text Data

(1) Data Cleaning

When using crawler software for Chinese language and literature text data acquisition, due to the large amount of collection will inevitably produce redundant data, in the final crawling results will be duplicated, invalid and other data. Therefore, in order to further convert the unstructured data into structured data for the next sentiment analysis, the crawled data are firstly subjected to data cleaning, including checking whether the missing values need to be supplemented, removing duplicated data, deleting invalid data and so on.

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

(2) Text Segmentation

Segmentation is a key step in Chinese text processing, for Chinese text, a single word is the unit of Chinese text, and a sentence is composed of multiple words, sentences and sentences between the form symbols as a division, and words and words are coherent, only words and words together to reflect the semantics of a sentence. Therefore, before classifying a Chinese text, it is necessary to divide a complete sentence into individual words based on a certain word separation algorithm, and then group the individual words into words according to certain rules.

(3) De-duplication

Deactivated words are words that have no effect on text categorization, and instead of improving the accuracy of text categorization, such words may even affect the accuracy of the categorization results. In different task scenarios, the types of deactivated words are likely to be different, and it is possible that a word defined as a deactivated word in one scenario may not be a deactivated word in another scenario. Therefore, in order to improve the computational efficiency and classification accuracy of the classification model at the cost of minimizing the memory footprint during text classification, it is necessary to actively remove these words or phrases before performing text classification. In this paper, regular expressions are added to the deactivated words to optimize the removal of deactivated words, which speeds up the computational efficiency of the algorithm and improves the accuracy of the classifier.

3) Literary Text Eigenvalue Extraction

Word Frequency-Inverse Document Frequency Keyword Extraction Algorithm (TF-IDF) is widely used in the field of text keyword extraction. In a document, the importance of words is called word weight, and the larger the weight the more it can represent the topic of the document. In TF-IDF algorithm, TF represents the frequency of keywords appearing in a document, and IDF mainly reflects the distribution of documents containing keywords in the total document.TF-IDF algorithm is based on the statistical way of calculating word weights, mainly using two statistical features of words to calculate the weight of words, which are the word frequency of the word and the inverse document frequency of the word.

In TF-IDF algorithm, TF is the word frequency of the feature word appearing in the document and IDF is the inverse document frequency, then it can be expressed as:

$$TFIDF = TF * IDF = tf * \log\left(\frac{N}{n_t} + 0.01\right), \quad (1)$$

where, tf is the word frequency representing the keyword, N is the total number of documents, and n_t is the number of documents related to the keyword. The traditional TF-IDF algorithm does not consider the distribution information between feature words as well as ignores the incomplete categorization of words. Aiming at the defects of traditional TF-IDF algorithm, this paper proposes an improved TF-IDF-AG algorithm, which calculates TF-IDF as follows:

$$TF - IDF - AG = tf * \log n_t * \log\left(\frac{N}{n_t} + 0.01\right) * \left[1 - \frac{\sqrt{\frac{\sum_{j=1}^{n}(tf-\bar{t})^2}{k-1}}}{\widehat{tf}}\right], \quad (2)$$

where tf denotes the word frequency, $\widehat{tf}$ denotes the average number of occurrences of the feature word in each document, N is the total number of documents, n_t is the number of documents associated with the keyword, and k is the total number of documents in the other categories.

Based on the above steps, this paper establishes a corpus of Ba Jin's Cold Night and Resting Garden literary texts, as well as a corpus of the works of two literary writers, Yu Zheng and Qiong Yao, with the aim of providing support for analyzing the expression of linguistic features in Chinese language literary texts.

B. Literary Text Theme Mining Techniques

1) Text Vector Representation of Literature

Vector space model (VSM) is to map a text document onto vectors, i.e., a paragraph of text is regarded as a vector, and the feature words contained in the text are regarded as the dimensions of the vector. The whole text is then transformed into a vector consisting of many one-dimensional vectors, and the problem of calculating text similarity can be transformed into the problem of calculating the cosine distance between two text vectors.

Given a n-dimensional text vector d, $\overrightarrow{d}_n = \{v_1, v_2, \ldots, v_n\}$, each v in the vector represents the weight score of each feature word item in the lexicon computed by the TF-IDF-AG algorithm. Using the cosine distance formula, the similarity between document d_1 and document d_2 can be expressed as:

$$s(d1, d2) = \frac{\vec{d_1} \cdot \vec{d_2}}{\left|\vec{d_1}\right| \left|\vec{d_2}\right|}, \quad (3)$$

if the document has more feature words, it will lead to the dimensional disaster of the vector, which not only has high computational complexity but also the results are not representative. Therefore, when extracting features from documents, feature merging can be performed, i.e., combining feature words with similar meanings to reduce the dimensionality and improve the accuracy of the model.

2) LDA Topic Text Features

Implicit Dirichlet Distribution (LDA) model is a document generation model, its model structure is shown in Figure 3.The basic principle of the LDA model is to convert the text data into the form of probability distributions, so that the probability distributions can be used to describe the topic structure of the text data.The LDA model assumes that each document consists of a number of topics, and each topic consists of a number of words. From the perspective of generative

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

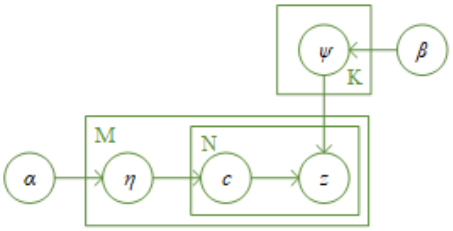

Figure 3: Structure of LDA model

modeling, each word in a document needs to be obtained through a process of "selecting a certain topic with a certain probability, and selecting a certain word from this topic with a certain probability".

where the two prior parameters are denoted by α and β, the distribution of topics extracted from parameter α is denoted by η, the topics extracted from η are denoted by c, the distribution of words corresponding to topic c extracted from parameter β is denoted by ψ, and the final generated word is denoted by z. In this model, the word z is sampled by taking the topic c and the parameter β as the basis, and the topic c is extracted from the parameter α, and their joint probability distribution is:

$$P(z, c|\alpha, \beta) = P(z|c, \beta)P(c|\alpha), \tag{4}$$

where prior parameter β follows an independent multinomial distribution with respect to parameter ψ, is updated as follows using parameter ψ:

$$P(z|c, \psi) = \prod_{i=1}^{|z|} P(z_i|c_i) = \prod_{i=1}^{|z|} \psi_{s_i,z_i}. \tag{5}$$

Expand the above equation as follows:

$$P(z|c, \psi) = \prod_{k=1}^{k} \prod_{j=1}^{v} P(z_i|c_i)\, \psi_{z,j}^{n_k^z}, \tag{6}$$

where n_k^z denotes the number of occurrences of word z in theme k.

3) Gibbs Sampling Parameter Estimation

In LDA subject modeling, there are usually two approaches for solving parameter estimation, namely the variational maximum expectation (VEM) algorithm and the Gibbs sampling algorithm. The variational maximum expectation algorithm is often used to compute the maximum likelihood estimation of the parameters, but due to the complexity of the iterative process of this method and its dependence on unobservable hidden variables in order to complete the inference, Gibbs sampling algorithm is used in this paper to estimate the parameters in the LDA thematic model.

Gibbs sampling is a practical method for parameter estimation in thematic models, which evolved from and is a special case of the Markov chain Monte Carlo method (MCMC), and is usually used to obtain a series of observation samples that are approximately equal to a specified multidimensional probability distribution, e.g., it can be used to obtain the joint probability distribution of two or more random variables.The specific procedure of Gibbs sampling is as follows:

The formula and process of sampling is represented below:

$$\begin{aligned} P\left((z_d + z_l) = k|\vec{z}_{-d,l}, \vec{w}\right) &= \frac{P(\vec{w},\vec{z})}{P(w,\vec{z}_{-d,l})} \\ &= \frac{P(\vec{w}|\vec{z})}{P(\vec{w}_{-d,l}|\vec{z}_{-d,l})P(w_{d,l})} \cdot \frac{P(\vec{z})}{P(\vec{z}_{-d,l})} \\ &\propto \frac{\left(n_{k,-d,l}^j + \beta_j\right)}{\sum_{j=1}^{W}\left(n_{k,-d,l}^j + \beta_j\right)} \cdot \frac{\left(n_{m,-d,l}^k + \alpha_k\right)}{\sum_{k=1}^{K}\left(n_{m,-d,l}^k + \alpha_k\right)}, \end{aligned} \tag{7}$$

where n_k^j denotes the number of words j belonging to the knd topic in the corpus, n_m^k denotes the number of topics k in the mth document, and $\neg d, l$ denotes the removal of words from the topics and tagged topics that are currently being sampled. The right hand side of this formula is $P\left(topic|doc\right) \bullet P\left(word|topic\right)$, which is the probability of a path of $doc \Rightarrow topic \Rightarrow word$. Since the number of topics and labeled topics is k in total, the physical meaning of the Gibbs Sampling formula is to sample these k paths.

After sampling convergence, the set of document-topic probabilities and topic-vocabulary probability parameters Θ and Φ can be obtained based on the state of the Mahalanobis chain, and after the topics of the entire set of comment documents have stabilized, the document-topic distribution $\theta_{m,k}$ and topic-vocabulary distribution $\varphi_{k,j}$ can be computed as respectively:

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k}, \tag{8}$$

$$\varphi_{k,j} = \frac{n_k^j + \beta_j}{\sum_{j=1}^{W} n_k^j + \beta_j}. \tag{9}$$

After obtaining the probability distributions of themes and vocabularies, the vocabularies with the highest probability values under each theme are selected to form the theme pool according to the word co-occurrence analysis method. Judge whether the vocabulary comes from the labeled topics, and if it comes from the labeled topics, add N_{avg} to the number of its occurrences, so as to propose the improved topic and vocabulary probability distributions as follows:

$$\theta_{m,k}^{avg} = \frac{n_m^k + \alpha_k + N_{avg}}{\sum_{k=1}^{K} n_m^k + \alpha_k + N_{avg}}, \tag{10}$$

$$\varphi_{k,j}^{avg} = \frac{n_k^j + \beta_j + N_{avg}}{\sum_{j=1}^{W} n_k^j + \beta_j + N_{avg}}. \tag{11}$$

In summary, the specific process of Gibbs sampling algorithm is as follows:

For all vocabularies in the document set, randomly assign a topic number, sample the topic information according to the Gibbs sampling formula above, while constantly updating the parameters, so that the cycle iteratively repeat the above sampling process for all vocabularies in the document set. When the vocabulary corresponding to the topic information approximate convergence, according to the number of topics generated by the document and the number of topics corresponding to the number of vocabulary, the document-topic

(including labeled topics) probability distribution and topic-vocabulary probability distribution can be estimated. The results are averaged to obtain the final parameter estimates.

IV. Linguistic Characteristics of Chinese Language Literature Texts

The linguistic features of a text are the embodiment of the author's speech characteristics in the process of writing, and these features, like an individual's "linguistic fingerprint", are an unconscious and profound reflection of the author's linguistic style characteristics. These features can be described to some extent by the statistical analysis of quantitative features, which in turn can more intuitively observe and rationally analyze the author's personal language style and speech expression habits. The text mining technology combined with literary text corpus to analyze the linguistic features of Chinese language literary texts, which helps the study of linguistic features to be more scientific and objective, and helps readers to know and understand Chinese language literary texts more intuitively.

A. General Characterization of the Textual Vocabulary

This section is based on the LDA topic text feature model combined with the corpus Wordsmith software to conduct a general analysis of the lexical features of Bajin's two literary works, Cold Night and Resting Garden, at the lexical level. Firstly, the use of real words in the novels is analyzed by using the lexical density index, and then the linguistic style of the literary texts is examined by combining with the high-frequency word disk analysis to explore the correlation between thematic text features and the theme of the texts.

1) Lexical Density of Literary Texts

Lexical density refers to the proportion of real words in the corpus, which is calculated as real words ÷ total words × 100%. Real words are the main bearers of information in language, the more real words there are in a text, the more information it carries, and at the same time, the reading difficulty of the readers will be elevated accordingly, and vocabulary density is an important quantitative index reflecting the style of using words in a text.

This paper takes the division method in Modern Chinese as the standard, and divides nouns, verbs, adjectives, distinguishing words, numerals, quantifiers, adverbs, pronouns, exclamations and onomatopoeia into real words according to the grammatical functions of the words. After searching and counting, the results of lexical density of the novels Cold Night and Resting Garden are obtained as shown in Table 1.

From the distribution of vocabulary density of two literary works, Cold Night and Rest Garden, the total vocabulary density of these two novels is very close to each other. The total vocabulary density of Cold Night is 86.98%, and the total vocabulary density of Rest Garden is 84.88%, and the vocabulary density of the text of Cold Night is slightly higher than that of the text of Rest Garden by 2.1 percentage points. This indicates that the text of Cold Night carries slightly more information than that of Rest Garden, but the difference is small. From this, it can be inferred that the authors produced works of fiction of comparable difficulty during this period, and that the readers' experience of the difficulty of reading these two novels would not be significantly different. In terms of words with different lexical properties, Cold Night and Resting Garden have the same ordering of lexical densities on various lexical properties. According to the order of vocabulary density from high to bottom, they are verbs, nouns, pronouns, adverbs, adjectives, number words, quantifiers, exclamations, and onomatopoeia, in which the vocabulary density of verbs is the highest in both literary works, and it reaches 27.31% and 28.25% respectively. This shows that there is a strong consistency in the writer's choice of various types of vocabulary in the creation of these two novels. The above analysis shows that real words, as the main bearer of information in language, its lexical density can indeed reflect the complexity of the content of the novels to a certain extent.

2) Word Frequency Characteristics of Literary Texts

Word frequency refers to the number of words that each word form appears in the text, and word frequency statistics is a basic statistical tool in corpus analysis. Word frequency statistics can reflect the distribution of words that frequently appear in the author's creative process. In particular, the use of high-frequency words can reflect the author's personal style in the use of words and the writing style of a certain work. Using the Wordlist function of Wordsmith software to count the high-frequency words of the two novels, we get the distribution of the top twenty high-frequency words of the two literary works, Cold Night and Resting Garden, as shown in Table 2.

Based on the word frequency distribution of high-frequency words, the following conclusions were drawn:

(1) There are certain commonalities in the use of the top 20 high-frequency words in the two novels, "Cold Night" and "Rest Garden", which reflect some common characteristics of the two novels, both novels use a large number of dialogue forms and focus on the language description of the characters. Verbs and nouns such as "say", "dao", "speak", "answer", "ask" and "speak" account for a large proportion of the high-frequency words in the two novels. Both novels focus on the physical appearance of the characters. The two negative words "no" and "no" occupy the top five positions in the high-frequency words of the two novels, and the word frequency of the two negative words reaches 859 and 527 times respectively in "Cold Night", and 563 and 375 times respectively in "Rest Garden".

(2) The highest-frequency verb in both novels is "say", with word frequencies of 1,213 and 735 respectively, indicating that character dialogues account for a larger proportion in both novels. The sensory verb "see" ranks second only to "say" in the two novels. It can be inferred that the author's perception of the characters in the two novels mainly focuses on visual depiction. In addition, in terms of the use of psychological

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

	Cold Night		Recreation Park	
	Quantity	Proportion	Quantity	Proportion
Nouns	16312	20.79%	11489	21.49%
Verbs	21428	27.31%	15104	28.25%
Adjective	4969	6.33%	3326	6.21%
Numerals	2067	2.63%	1345	2.52%
Quantifier	463	0.59%	288	0.54%
Adverb	7652	9.75%	4472	8.36%
Pronoun	15004	19.12%	9253	17.31%
Exclamation	263	0.34%	88	0.16%
Onomatopoetical	91	0.12%	21	0.04%
Total	68249	86.98%	45386	84.88%
Total word	78463		53471	
Total vocabulary density	86.98%		84.88%	

Table 1: Text vocabulary density distribution

verbs, there are significantly more high-frequency words in psychological verbs in Cold Night than in Rest Garden. In Cold Night, "think" appears 381 times, "know" appears 174 times, and "feel" appears 132 times, while there is only one mental verb in the top 20 high-frequency words in Rest Garden. "think", which occurs 63 times. From this, it can be inferred that a large number of inner descriptions of characters are used in Cold Night to enrich the characterization and infer the plot development of the novel.

B. Text Similarity Characterization

1) Sentence-Level Similarity

The linguistic features of literary works are not only reflected at the character and vocabulary levels studied earlier, but also at the sentence level. Distinguishing the linguistic styles of the two authors can be examined through four stylistic features: text reading difficulty, narrative rhythm of the work, emotional value and linguistic richness. The reading difficulty of the text is directly related to the average sentence length and average paragraph length, the narrative rhythm of the work is embedded in the proportion of sentences, and the emotional value and linguistic richness can be reflected by the proportion of word choices and the distribution of special sentence patterns. In this part, the average sentence length linguistic features will be statistically analyzed, aiming to analyze the similarity between Yu Zheng and Qiong Yao's literary texts.

Based on the literary works of Yu Zheng and Qiong Yao collected in the corpus, the average sentence length and sentence length dispersion of their literary works were counted to determine the stylistic characteristics of Yu Zheng and Qiong Yao in terms of the use of long and short sentences and the rhythm of narration. The statistical results of average sentence length and sentence length dispersion are shown in Table 3.

On the whole, the average sentence lengths of Yu Zheng's five works are significantly larger than those of Qiong Yao's five works, and the average sentence lengths of these ten works fluctuate between 18.49 and 34.27. Among them, the average sentence length of The Last Gege is the longest at 34.27, and the average sentence length of Returning the Pearl is the shortest at 18.49, indicating that the former text is more difficult to understand and less readable, while the latter text is less difficult to understand. In terms of the choice of long and short sentences, Yu Zheng tends to use long sentences, and the sentence expression is no different from that of modern texts. Qiong Yao's works are more dominated by short sentences, and the language achieves a poetic effect, probably because the authors are deeply influenced by the culture of classical poems and words, and their language is concise, with a strong classical flavor, and running sentences are more common. The average sentence length of "Palace Locked City" is 30.45, which is significantly higher than the average sentence length of "Brand of Plum Blossoms", which is 25.74. However, in terms of the trend, "Palace Locked City" has the lowest average sentence length among Yu Zheng's works, and "Brand of Plum Blossoms" has the high average sentence length among Qiong Yao's works, which are not very similar, but show a convergent and close tendency. In addition, the sentence length dispersion fluctuates between 15.64-22.42 in general, and the sentence length dispersion of Yu Zheng's works is generally higher than that of Qiong Yao's works, which indicates that Yu Zheng's sentence variations are richer than Qiong Yao's. In terms of sentence length matching, Yu Zheng's works have the lowest average sentence length among Qiong Yao's works. In terms of sentence length, Yu Zheng tends to use a combination of short and long sentences, with a weakened sense of rhyme, which is not as elegant and catchy as Qiong Yao's works, but the rhythm of expression is not confined to a single pattern, and is more rich and tumultuous, which attracts people's attention. On the other hand, Qiong Yao's works are more stable in the use of short sentences, and the language is rich in rhythmic beauty like poetry, but the narrative rhythm of the text is therefore fixed, and lacks the sense of twists and turns in the plot of novels.

2) Concentration of Text Topics

In discourse, paragraph length reflects the rhythm of expression of the text, the degree of thematic concentration reflects the degree of fit or overlap of thematic content, and the degree of textual similarity grasps the similarity of information features from the text as a whole. By examining the degree of thematic concentration to appreciate the author's different understanding and treatment of theme, rhythm and emotion, which is also an expansion and extension of the linguistic style

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

Ranking	Cold Night	Word frequency	Recreation Park	Word frequency
1	Say	1213	Say	735
2	No	859	No	563
3	Mother	684	Go	402
4	Go	565	Nothing	375
5	Nothing	527	Father	353
6	See	396	See	289
7	Think	381	Still	234
8	Still	349	Mother	208
9	Go	321	Man	192
10	Man	259	Go	183
11	Upper	231	Both	164
12	Speak	216	Kids	153
13	Good	202	Wife	146
14	Both	193	Good	127
15	Know	174	Brother	120
16	Very	161	Very	115
17	Head	148	Just	108
18	Feel	132	See	92
19	Wife	116	Face	79
20	Just	108	Think	63

Table 2: Frequency analysis of high-frequency words

No.	Works	Total	Total sentence	Average sentence	Length dispersion
1	GongSuoChenXiang	134518	4864	27.65	19.36
2	GongSuoLianCheng	463583	20221	22.92	18.83
3	GongSuoZhuLian	402967	12637	31.88	22.42
4	GongSuoXinYu	156876	6515	24.07	16.01
5	ZuiHouDeGeGe	231524	6754	34.27	19.23
6	MeiHuaLao	90153	3502	25.74	15.99
7	ShuiYunJian	93521	4534	20.62	17.75
8	GuiZhangFu	124735	4153	30.03	16.21
9	HuanZhuGeGe	153294	6159	24.88	15.64
10	GongSuoChenXiang	264174	14283	18.49	16.55

Table 3: Statistical results of average sentence

of comparative works, should be included in this study. This paper analyzes the thematic concentration of Yu Zheng and Qiong Yao's literary works in the corpus based on the LDA thematic text feature model given in the previous section, in order to analyze the similarities and differences and stability of the two writers' thematic expression techniques. Figure 4 shows the distribution of theme concentration of 10 literary works.

As a whole, the variation range of theme concentration of these 10 works is small, fluctuating between 0.09 and 0.18. Most of Qiong Yao's works have higher thematic concentration than Yu Zheng's works, with the maximum value being 0.71 for Branding Plum Blossoms, and the minimum value being 0.096 for The Raiders of Yanxi.In terms of the size of the values, Qiong Yao's works have a more concentrated degree of thematicity and greater story integrity. As for Yu Zheng's works, due to the generally more ambitious length, and similar to "Palace Locked Pearl Curtain" and "Palace Locked Heart Jade" belonging to the theme of modern crossing the ancient world, the main content of the work is more extensive, telling the story of characters' love and hate, but also more space is used to recount the "crossing of the book", and the sense of the main line needs to be further clarified.

In terms of fluctuation trends, Yu Zheng's works are more stable in their treatment and approach to theme expression, and Qiong Yao's works have relatively more significant changes in theme concentration. Under the premise that the overall difference is not significant, the difference in theme concentration between "Palace Locked City" and "Brand of Plum Blossoms" is more obvious, with its theme concentration of 0.118 and 0.171 respectively, with the former 30.99% lower than the latter. This is due to the fact that "Palace Locked City" sets up more characters and relationships than "Brand of Plum Blossoms", and also makes the subplots richer and fuller with more twists and complexities, thus lowering the thematic concentration of the work. In addition, it is the author's approach to the textual narrative that leads to the difference in the degree of thematic concentration. As a great writer of romance novels, Qiong Yao is very good at using the rhetorical device of repetition, which fully expresses the emotions and deepens the theme, highlights the thoughts of the characters, and deepens the rhythmic beauty of the text of the work. Qiong Yao's use of repetition is very brilliant, interpretation to strengthen the character's emotions, but also to strengthen and highlight the theme of the idea, thus forming its unique "Qiong Yao style" mood, but also to make the text of the theme of the invisible more compact, the text of the theme of the degree of concentration can be enhanced.

In addition, this paper also "Palace Locked City", "Plum Blossom Brand" text and other literary works of text similarity

ARCHIVES DES SCIENCES
A Multidisciplinary Journal

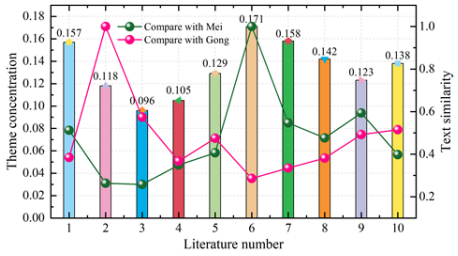

Figure 4: The topic concentration distribution of literary works

comparison, can be clear that Yu Zheng's works between the similarity is generally higher, and "Palace Locked City" compared to "Palace Locked Jade", the highest is the "Palace Locked Heart" to reach 0.574, this is related to the Yu Zheng's linguistic style and thematic subject matter of the more stable, and the lowest is the "Palace Locked Curtains" to reach 0.364. Between Qiong Yao's work The similarity between Qiong Yao's works is more complicated, with the highest textual similarity between Brand of Plum Blossoms and Outside the Window at 0.594, and the lowest at 0.398, which shows that Brand of Plum Blossoms, Between the Clouds of Water and The Ghost Husband belongs to the series of Plum Blossoms, and the serialization of the subject matter makes the textual similarity close to each other. On the whole, the cosine vector value of Branding Plum Blossoms and Palace Locked City is only 0.126, which indicates that the expression habits of the co-occurring words directly related to the two works are different, and the degree of text similarity is relatively low.

V. Conclusion

In this paper, taking the value of Chinese language and literature text as the entry point, using web crawler technology to obtain Chinese language and literature text data and establish a corpus, using Gibbs sampling algorithm to estimate the parameters of LDA topic model, and analyzing the linguistic features of Chinese language and literature text in terms of two dimensions: vocabulary and similarity. The conclusions are as follows:

(1) In Bajin's Cold Nights and Rest Garden, the vocabulary density of the text of Cold Nights is slightly higher than that of Rest Garden by 2.1 percentage points, and the verb vocabulary density of the two literary works is the highest and reaches 27.31% and 28.25% respectively. Both literary works use a large number of dialogues to describe the language of the characters, in which the frequency of "say" reaches 1,213 times and 735 times respectively. The use of text mining technology can effectively analyze the language and vocabulary characteristics of Chinese language literary texts, clarify the emotional expression of Chinese language literary texts, and provide a reference for understanding the author's writing purpose.

(2) In the textual similarity comparison of Chinese language literature between Yu Zheng and Qiong Yao, the average sentence lengths of Yu Zheng's five works are significantly larger than those of Qiong Yao's five works, and the average sentence lengths used by the two writers in their 10 works fluctuate between 18.49 and 34.27. The theme concentration of Qiong Yao's works is mostly higher than that of Yu Zheng's works, among which the difference between the theme concentration of "Palace Locked Liancheng" and "Plum Blossom Brand" is larger, with the former being 30.99% lower than the latter. Analyzing the linguistic thematic expressions of different literary works through the LDA theme model can help readers understand the similarities and differences between literary works and clarify the spiritual core of literary works.

References

[1] Chilman, N., Song, X., Roberts, A., Tolani, E., Stewart, R., Chui, Z., ... & Das-Munshi, J. (2021). Text mining occupations from the mental health electronic health record: a natural language processing approach using records from the Clinical Record Interactive Search (CRIS) platform in south London, UK. *BMJ Open, 11*(3), e042274.

[2] Spinellis, & Diomidis. (2017). A repository of unix history and evolution. *Empirical Software Engineering, 22*(3), 1372-1404.

[3] Wang, Z. (2019). Problems and analysis of improving the applicability of chinese language and literature major. *Basic & Clinical Pharmacology & Toxicology.(s2),* 125-131.

[4] Xie, Z. (2017). 77.research on the application of chinese paper-cut art in primary and secondary school art education. *Boletin Tecnico/technical Bulletin, 55*(20), 541-547.

[5] Tassone, V. C., Dik, G., & van Lingen, T. A. (2017). Empowerment for sustainability in higher education through the EYE learning tool. *International Journal of Sustainability in Higher Education, 18*(3), 341-358.

[6] Su, F. , Chang, J. , Li, X. , Zhou, D. , & Xue, B. (2021). Urban circular economy in china: a review based on chinese literature studies. *Complexity, 2021*(3), 1-10.

[7] Shuhui, W. (2018). Life education for college students from the perspective of chinese traditional culture. *Journal of Teacher Education.*

[8] Shen, D., Guo, H., Yu, L., Ying, J., Shen, J., Ying, S., ... & Wang, Y. (2022). Sound design of guqin culture: Interactive art promotes the sustainable development of traditional culture. *Sustainability, 14*(4), 2356.

[9] Weizhe, Y. (2017). The influence of ancient poetry on chinese traditional culture. *International Journal of Technology, Management,*(1), 24-25.

[10] Windle, J. (2017). The public positioning of refugees in the quasi-education market: Linking mediascapes and social geographies of schooling. *International Journal of Inclusive Education, 21*(11), 1128-1141.

[11] Li, X., Fan, M., Zhou, Y., Fu, J., Yuan, F., & Huang, L. (2020). Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining. *Nano Energy, 71*, 104636.

[12] Luo, H. , & Yang, C. (2018). Twenty years of telecollaborative practice: implications for teaching chinese as a foreign language. *Computer Assisted Language Learning, 31*(5-8), 546-571.

[13] Only, S. (2017). A comprehensive evaluation of new course reform of chinese language and literature by using online survey. *Boletin Tecnico/technical Bulletin, 55*(4), 726-731.

[14] Tian, L. (2017). Study on the promotion strategy of the effect of chinese teaching based on the fuzzy comprehensive evaluation model. *Revista de la Facultad de Ingenieria, 32*(14), 826-832.

[15] Guan, L. (2017). Research on the optimization of chinese language and literature teaching based on big data and modern educational thought. *Revista de la Facultad de Ingenieria, 32*(9), 323-328.

[16] Huang, Y. , & Miao, W. (2021). The internationalization of chinese english-language humanities and social science journals: their status and challenges. *Journal of Scholarly Publishing, 52*, 273-293.

[17] He, X., & Tian, S. (2022). Analysis of the communication method of national traditional sports culture based on deep learning. *Scientific Programming, 2022*, 1-8.