https://doi.org/10.15302/J-LAF-1-020083 Received Date | 2023-02-13 Revised Date | 2023-05-23 Accepted Date | 2023-10-20

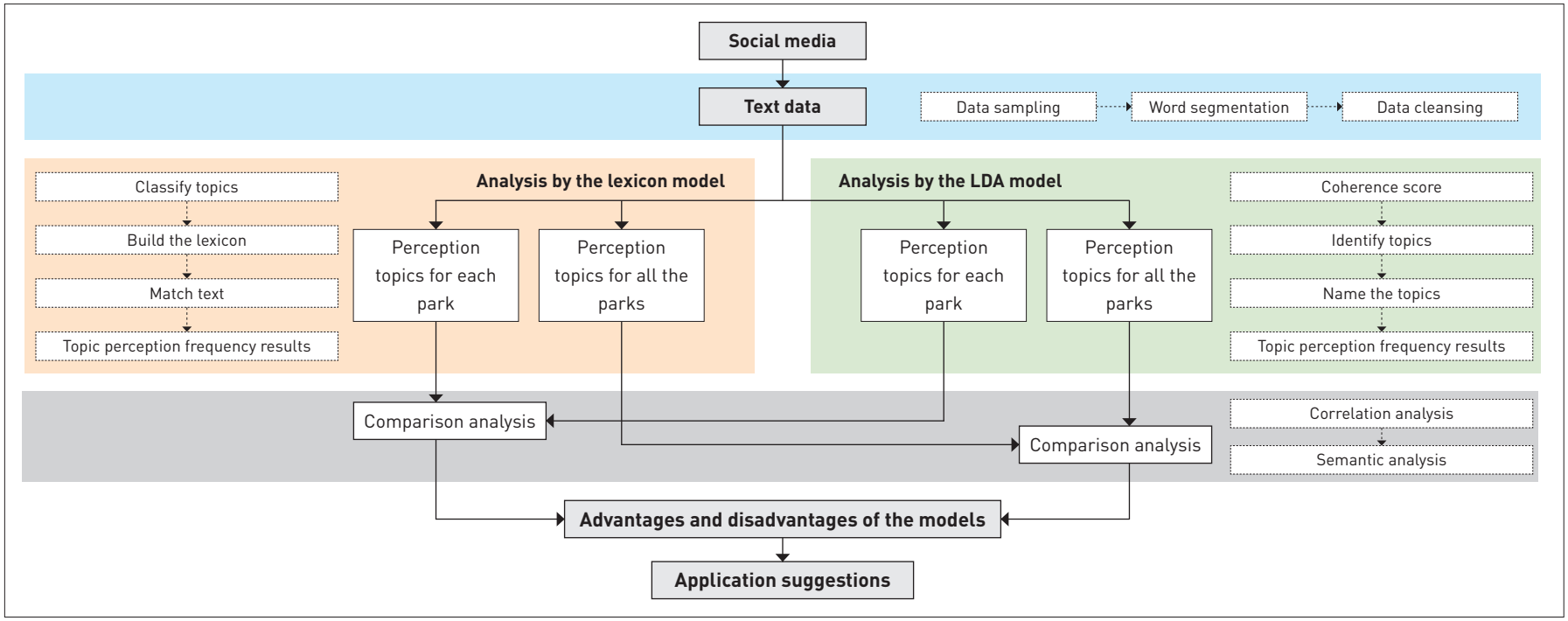
Comparison and Applicability Study of Analysis Methods for Social Media Text Data: Taking Perception of Urban Parks in Beijing as an Example

Zhenyu SHANG, Kexin CHENG, Yuqing JIAN, Zhifang WANG*

College of Architecture and Landscape, Peking University, Beijing 100080, China

*CORRESPONDING AUTHOR
Address: No. 3, North Street, Yandongyuan, Haidian District, Beijing 100080, China
Email: zhifangw@pku.edu.cn

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Exploring the advantages, disadvantages, and applicability of two text analysis models
- The lexicon model is more suitable for parallel comparison between perceived objects by users
- The Latent Dirichlet Allocation (LDA) model can better capture the characteristics of each individual perceived object
- Taking advantage of the two models' strengths is vital for optimizing landscape perception assessment

KEYWORDS

Social Sensing;
Text Analysis; Lexicon;
Latent Dirichlet Allocation (LDA) ;
Urban Park;
Landscape Perception

ABSTRACT

The booming Internet technology and media have generated large sets of social media data, with which the social sensing analyses based on users' reviews have become a research hotspot and have been increasingly applied in the study of urban park usage and perception. However, most existing studies adopt a single model for text data processing. To fill this gap, this study aims to compare social media text data analysis methods and assess their advantages, disadvantages and applicability in park perception research. The Lexicon-based classification analysis model (lexicon model) and LDA (Latent Dirichlet Allocation) model widely used in relevant research were selected. Based on text data obtained from public reviews of 10 urban parks in Beijing on Dianping, this study explored the perception topic distribution of each park and all parks in general, and compared the classification results of perception topics between these two models. Results show that the lexicon model is conducive to the parallel comparison of perception frequency between parks, while the LDA model can directly reflect each park's characteristics and visitors' perception preferences; the combined use of the two models can optimize park perception assessment. Results from the two methods reveal that visitors to urban parks in Beijing focused more on their social recreation needs and visual aesthetics brought by the natural landscape, as well as conditions of the transportation facilities and the consumption in the parks. This research can provide optimization suggestions for the selection and use of social media text analysis methods, and a basis and guidance for park construction and management improvement.

EDITED BY
Ying WANG, Jiayi ZHOU
TRANSLATED BY
Ying WANG, Zhenyu SHANG, Jiayi ZHOU

1 Introduction

With the rapid development of Internet technology, people socialize more and more frequently through online media, while generating a huge amount of information which provides the data base for the study of social sensing[1][2]. This type of research focuses on analyzing people's perceptions of spaces, as well as human mobility patterns and social relations between individuals by mining information on human behavioral characteristics contained in big data[3][4]. With the growth of social media data in recent years, there has been a gradual increase in research on analyzing geospatial sentiment perceptions[5][6]. The social media data can be mainly divided into three categories: check-in data, image data with geolocation, and text data; and the methods commonly used in early research were the analysis of arrival rate and motivational preference identified by the check-in data[7]~[10], as well as the perceived sentiment analysis based on image data contents and their geolocation[11]~[15]. In recent years, as the intuitive expression of sentiment by text data has been gradually recognized, there has been increasing research on perception analysis through text data mining[16]~[19]. For instance, in social perception analysis research, objects mainly include public sentiment on hot topics and response to risks and disasters[20]~[22], as well as the perception of using public facilities, especially the post-occupancy perception of scenic areas[23]~[26] and urban green spaces[27]~[29]. As a type of important public open space, urban parks provide residents with services such as access to nature, recreation, relaxation, and leisure[30][31]. Thus, researchers are attaching more attention to studying park perception through text data, i.e., analyzing the post-occupancy perception of visitors to guide and lay the groundwork for urban park construction and renewal.

Methods to analyze social media text data typically include word frequency analysis and semantic analysis[32]~[34]. The advancement of text mining technology has made it possible to build text analysis models to explore the internal laws and topics in text data, and topic models have become the basis for perception analysis and satisfaction evaluation. Commonly used topic models for text analysis include the Lexicon-based classification analysis model ("lexicon model" hereafter)[35], K-means model[36][37], Latent Dirichlet Allocation (LDA) model[5], Naive Bayes model[38][39], Linear Regression and Logistic Regression models[40], Random Forest and Decision Tree models[41], etc. Most existing studies tend to adopt a

single model for text data processing in perception analysis without exploring the advantages and disadvantages of different models and their applicability.

This study aims to compare social media text data analysis methods and reveal their applicability in park perception research. Since the lexicon model and LDA model are widely used in research on the perception of scenic spots and urban parks, they were chosen for comparison. Most lexicon models first semantically analyze high-frequency words in the text obtained, then establish a corresponding lexicon according to an existing standard system to classify different words and expand the lexicon to make it more complete, and finally further classify and analyze the text content according to the optimized lexicon[42]~[44]. The LDA model is a machine learning-based model, mainly used for topic extraction and classification in text analysis[45]~[49]. This study focuses on following questions: when analyzing social media texts related to park perception, what are the differences between the process and analysis results of lexicon model and LDA model? What are the advantages and disadvantages of the two models? On this basis, we further explore approaches to utilizing both models to provide guidance for urban park planning to summarize the applicability of text analysis methods in park perception research.

2 Data Processing and Research Methods

2.1 Study Area and Data Sources

Covering an area of approximately 16,410 km^2, Beijing has a permanent resident population of 21.89 million by 2020, 1,050 parks of various types, and a total park green space area of 357.2 km^2.[50] As a super first-tier city with rapidly advancing Internet technology, its residents frequently use social media, providing mass data for this study.

Dianping was chosen as the source of text data. It is one of the highly influential social review platforms in China with a large number of reviews, and the number of active users is increasing year by year. Meanwhile, the growing active participation of users enhances the accuracy of the review data[51]. This study used the Request module in Python to obtain all the text review data and reviewer information from April 2006 to September 2020 in the catalogue of Beijing parks on Dianping and selected the top 10 urban parks, ranked by the number of reviews, as the objects of study (Table 1).

To ensure the accuracy of the model analysis, the study pre-processed the contents of the acquired text data. Firstly, delete short reviews[51] with less than 50 characters. After that, Beijing Garden Expo Park had the least reviews (6,531 pieces), based on which we randomly selected the same number of reviews for the other 9 parks using the SPSS software and finally obtained a total of 65,310 reviews.

In this study, jieba (a Chinese word segmentation tool) in Python was used to process data. Compared with other similar tools, jieba has the advantage of generating a customized lexicon, resulting in more accurate and effective word segmentation, and the adaptation of language environment[51]. In the next step, we cleansed the data by filtering out meaningless symbols and words[48], utilizing a lexicon of Chinese stop words that combined and deduplicated words from several lists including the HIT (Harbin Institute of Technology) Stop Word List. Considering the semantically similar Chinese words, the HIT-CIR Tongyici Cilin (Expanded) was used to substitute synonyms in the segmentation results to improve the accuracy and processing efficiency of the model[46]. Finally, a manual screening was conducted to adjust the segmentation and synonym replacement results based on the actual use of the park and the perception contents. In this study,

Table 1: Overview of the selected parks

No.	Name	Area (hm^2)	Number of reviews
1	Yuanmingyuan Park	350.0	17,805
2	Yuyuantan Park	129.4	17,698
3	Fragrant Hills Park	188.0	13,825
4	Jingshan Park	23.0	13,628
5	Beijing Shiyuan Park	503.0	11,923
6	Chaoyang Park	288.7	11,750
7	Beijing World Park	53.3	11,338
8	Olympic Forest Park	680.0	10,673
9	Badachu Park	332.0	9,889
10	Beijing Garden Expo Park	513.0	8,736

there were instances of inappropriate synonym substitution, such as replacing “cherry,” “daffodil,” or “begonia” with “chamomile tablet,” and “Haidian District” or “Chaoyang District” with “Baiyun District.” In this case, we kept the original words and deleted the substitutions.

2.2 Research Methods

2.2.1 Lexicon-based Perception Topic Classification Model

This study adopted the model for classifying and evaluating landscape service-based urban park perception topics proposed by Zhifang Wang et al. in 2021, as its validity has been proven and its overall performance was excellent[42]. Landscape service research focuses on the comprehensive effects of landscape patterns and functions, and the spatial process and relationship between service providers and demanders. Thus, considering that parks are a type of important urban green space landscape, this model can effectively reflect tourists’ perceptions and evaluations of the parks[42]. In this study, we conducted a structured processing of the pre-processed text data with Python and extracted high-frequency words, manually classified these words to build a Chinese lexicon for landscape service perceptions, expanded the lexicon both using the Word2vec word embedding model and manual additions[52], and finally classified each word into different perception topics. According to existing literature[53]~[57], 9 topic classifications of urban park perception categorized by landscape services were identified (Table 2)[42].

The next step was to match the obtained text data of reviews related to park perception with the lexicon to identify words used in the reviews, then extract the perception topics covered in each review, and finally calculate the perception frequency of each topic. The frequency was determined by the ratio of the number of reviews related to a perception topic to the total number of reviews for a park[42]:

$$F_i = N_i / M, \tag{1}$$

where F_i is the perception frequency of topic i, N_i is the number of reviews for a park mentioning contents of topic i, and M is the total number of reviews of the park.

2.2.2 LDA-based Perception Topic Classification Model

As an unsupervised language processing model that automatically analyzes texts, LDA quickly extracts topics from unstructured texts (i.e., documents) to realize the dimensionality

Table 2: Topic classifications and example words from the lexicon of urban park perception based on landscape services

Topic	Contents	Example words from the lexicon
Environmental improvement	Air quality improvement, microclimate regulation, noise	Humidity, exposure, freshness, wind and sunshine
Biodiversity	Animals, plants	Swan, holly, dead wood, birdsong and floral fragrance
History and culture	Cultural and historical values, cultural heritage, historical sites	Qing Dynasty, relics, art, Dragon Boat Festival
Aesthetic appreciation	Scenery, beauty, inspiration	Flowers, photography, attractive, unpleasant
Education	Popularization of science, education	Knowledge, learned, knowledgeable, ignorant
Religion	Religious worship, belief, refuge	Rituals, Buddha beads, Taoism, enlightenment, marriage seeking
Physical and mental recovery	Relaxation, stress release, mind restoration	Soothing, beautiful, downcast, cheerful
Recreational activities	Outdoor activities, sports	Walking, boating, hiking, ball games
Social interaction	Social integration, interaction between individuals	Mom, dad, friends and relatives, gatherings

reduction of documents[58]. In this model, a document consists of multiple topics in a certain probability distribution. Similarly, each topic consists of multiple words in a certain probability distribution. The larger the probability value is, the more closely related the set and its components are[59][60]. The LDA model calculates probability distributions of both "document–topic" and "topic–word," so as to classify document topics and corresponding words (keywords).

This study utilized the gensim toolkit of Python to invoke the LDA model for topic analysis of the text data. Determination of the number of topics (K value) needs to consider the granularity of the topic, the interpretability of the topic content, as well as whether it is convenient for comparative analysis. In this study, we first calculated the Coherence score of different numbers of topics, which can effectively represent the degree of similarity between keywords in a topic—a higher Coherence score indicates that the model is more effective in analyzing this number of topics[61]~[63]. Then, topics with high coherence scores were manually selected to determine the appropriate number of topics conducive to desired modeling results. After this, the actual weight (i.e., the perception frequency) of each topic was calculated as follows.

1) Determine the number of topics as K, and the total number of reviews as N;

2) Calculate the expected probability F_0 of K topics in each review, i.e., $F_0 = 1/K$;

3) Obtain the actual probability of the jth topic in each review as F_j by the LDA model (j = 1, 2, 3, ..., K), and compare the values of F_j and F_0;

4) Count the number of reviews in which $F_j > F_0$ as A_j;

And 5) obtain the actual weight of the jth topic as $Q_j = A_j/N$.

Based on the analysis results, the topics of each park were named by three researchers specializing in landscape architecture, considering both the keywords and the corresponding high-weight review text. Meanwhile, "noisy" topics were removed due to their low weight and weak correlation of content.

2.2.3 Correlation Analysis of Topic Distribution

The study conducted a correlation analysis on the distribution of varied perception topics obtained from the two models. The distribution of these topics in each review text is a dichotomous variable, with results of "yes" ("1") or "no" ("0"). Thus, we calculated the Phi coefficient in SPSS for the correlation test, mainly utilizing the 2-by-2 contingency table of binary variable values. As shown in Figure 1, when the values are mostly distributed on the main diagonal, it means that the correlation between different variable distributions is high and the coefficient can be calculated by equation (2):

X \ Y	Y=1	Y=0
X=1	N_{11}	N_{10}
X=0	N_{01}	N_{00}

1. Phi coefficient contingency table

$$\Phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{10}+N_{00})(N_{01}+N_{00})}}, \quad (2)$$

where X and Y denote values of the two variables ("1"/"0"), N_{11}, N_{10}, N_{01}, and N_{00} for numbers counted for different values of the variables, and Φ for the correlation coefficient of the two variables' distribution. A module related to the Phi coefficient in SPSS was used for data analysis. When the significance level is less than 0.05 and the Φ value approaches 1, it indicates a stronger correlation between the two topics.

2.2.4 Semantic Analysis of Topic Contents

This study utilized Python for word frequency analysis of the review texts and illustrated the high-frequency words via word clouds, where the size of the words indicates their frequency. These illustrations can effectively visualize the main contents of the selected review texts, while analysis of word frequency for review texts from each park can help reveal corresponding perception topics.

2.3 Technical Route

Based on the review text data of 10 parks in Beijing from Dianping, this study conducted text analysis with two types of models to explore the perception topic distribution of each park and all parks in general and compared the classification results of perception topics between these two models. The specific technical route is shown in Figure 2.

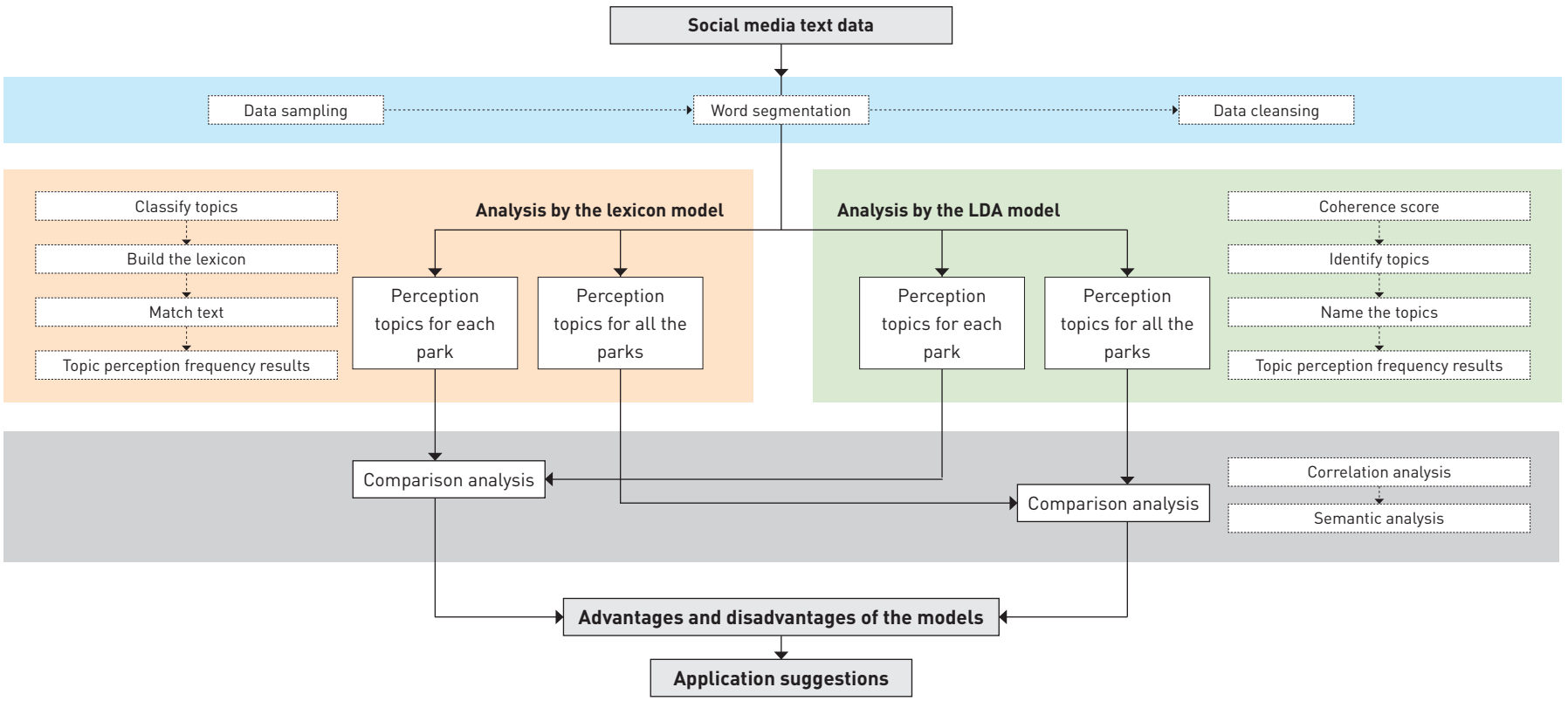
3 Research Results and Analyses

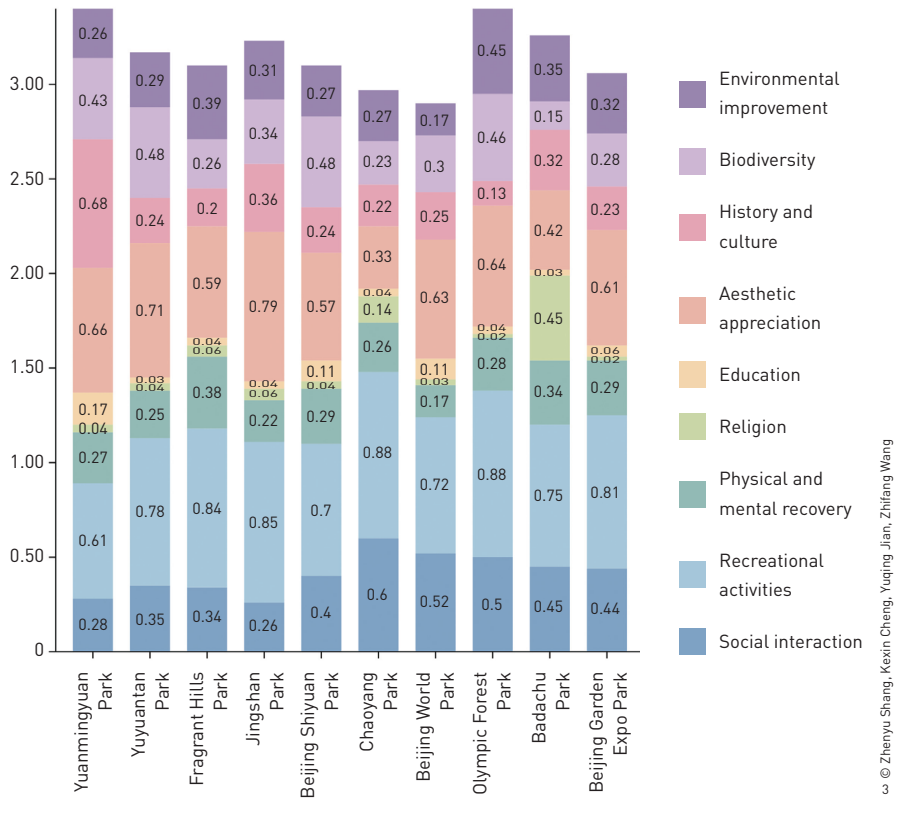
3.1 The Lexicon Model Facilitating Parallel Comparison Between Parks

Benefitting from manual presets, the Lexicon model covers relatively comprehensive and well-defined topics. Results of each

park's perception analysis were confined to the lexicon contents, which is conducive to further interpretation of the results and the parallel comparison of perception frequency and differences between parks.

Statistical results of the topic classification using the lexicon model show significant differences between parks on visitors' perception frequency of different topics (Fig. 3). The total perception frequencies of Yuanmingyuan Park (3.40) and Olympic Forest Park (3.40) were relatively the highest, while those of Beijing World Park (2.90) and Chaoyang Park (2.97) were the lowest. The largest difference between varied topics' perception frequencies existed in the Olympic Forest Park—0.88 for recreational activities and 0.02 for religion. In addition, among the perception topics of all the parks, recreational activities and aesthetic appreciation were comparatively more frequently perceived by visitors, while education and religion were less frequently perceived. Yuanmingyuan Park, Badachu Park, Jingshan Park, and Chaoyang Park showed a higher perception frequency in history and culture (0.68), religion (0.45), aesthetic appreciation (0.79), and social interaction (0.60) than other parks, respectively. Moreover, the topic of education in Yuyuantan Park and Badachu Park was less frequently perceived than in other parks.





2. Technical route of the research flow chart
3. Differences between parks on visitors' perception frequency of different topics shown in the lexicon model

3.2 The LDA Model Highlighting Each Park's Characteristics

The perception analysis results from the LDA model show significant differences between perception types of the 10 parks and reviews from social media highly reflected each park's landscape characteristics and visitors' perception preferences. After the review text data were processed by the LDA model, the appropriate number of topics for each park was determined according to their Coherence score (Table 3), after which the topics were named considering their interpretability and "noisy" topics were removed. For instance, in the results of Beijing Garden Expo Park, we identified a topic as a noisy one as its keywords include "arrive soon," "check," "sun umbrella," "turn left," "kite festival," "excellent," "wait for the bus," "Gate 3," "fully," "department," etc., which contained very vague perception contents, and had a relatively low weight of 0.016. The final distribution of topics varied from park to park (Table 4).

Table 4 shows that there were mainly 8 or 9 topics perceived by visitors in each park. Yuanmingyuan Park, Yuyuantan Park, and Olympic Forest Park had the most topics, while Beijing World Park had the least. In terms of the perception contents, although variation existed among all the parks, some topics

Table 3: Topic analysis results for each park based on the LDA model

Park	Number of topics	Coherence score
Yuanmingyuan Park	10	0.5611
Yuyuantan Park	10	0.5216
Fragrant Hills Park	9	0.6049
Jingshan Park	9	0.6334
Beijing Shiyuan Park	9	0.5829
Chaoyang Park	10	0.5845
Beijing World Park	6	0.6577
Olympic Forest Park	10	0.5046
Badachu Park	9	0.6385
Beijing Garden Expo Park	8	0.5411

Table 4: Perception topics for each park based on LDA model

	Yuanmingyuan Park	Yuyuantan Park	Fragrant Hills Park	Jingshan Park	Beijing Shiyuan Park
Topic 1	Transportation and tickets	Epidemic	Hiking activities	Cultural heritage	Natural landscape
Topic 2	Cultural relics	Cherry blossom festival	Park Introduction	Historical change	Service facilities
Topic 3	Garden landscape	Lupine view	Transportation and tickets	Park introduction	Social activities
Topic 4	Natural landscape	Cherry blossom view	Experiential perception	Featured flowers	Pavilion experience
Topic 5	Historical perception	Tickets and consumption	Cultural landscape	Featured constructions	Park introduction
Topic 6	Patriotic education	Natural landscape	Education and learning	Surrounding landscape	Music festival
Topic 7	Park introduction	Transportation facilities	Natural landscape	Epidemic	Service experience
Topic 8	Lotus flowers	Park introduction		The Forbidden City vision	Transportation and tickets
Topic 9	Featured ice-cream	Leisure and entertainment			
Topic 10					

Continued

Table 4: Perception topics for each park based on LDA model

	Chaoyang Park	Beijing World Park	Olympic Forest Park	Badachu Park	Beijing Garden Expo Park
Topic 1	Transportation and tickets	Transportation and tickets	Leisure sports	Worship activities	Featured gardens
Topic 2	Service experience	Collective memory and perception	Transportation facilities	Leisure facilities	Aesthetic experience
Topic 3	Parent-child activities	Park introduction	Social activities	Buddhist landscape	Activity experience
Topic 4	Leisure and entertainment	Performances	Night show	Gatherings	Transportation facilities
Topic 5	Book market	Featured landscapes	Sporting activities	Hiking activities	Cultural activities
Topic 6	Park introduction		Summer and Autumn view	Park introduction	Leisure activities
Topic 7	Temple fair		Spring view	Collective memory and perception	Park introduction
Topic 8	Spring activities		Park introduction	Transportation facilities	
Topic 9			Epidemic		
Topic 10			Gatherings		

like transportation (including information on buses, subways, parking lots, etc.) can be found in most parks. In addition, together with the word frequency analysis (Figs. 4, 5), it can be seen that some perception topics were expressed differently depending on each park's characteristics, such as the spring cherry blossom landscape in Yuyuantan Park versus the autumn foliage landscape in Fragrant Hills Park. Moreover, topics related to festivals reflected characteristic perception results, such as the Spring Festival temple fair in Chaoyang Park, the band performance in Olympic Forest Park, and other types of gatherings.

An overall perception analysis of the review text data from all the 10 parks by the LDA model identified 10 topics. Among these topics, the perception frequencies of transportation and tickets (0.60), spring view (0.53), collective memory and perception (0.52), and social activities (0.48) were higher than that of hiking activities (0.30), cultural history (0.29), gatherings and performances (0.26), autumn view (0.20), religious culture (0.14), and featured constructions (0.11). This implies that visitors to urban parks in Beijing prioritized their social interaction needs and visual aesthetics brought by the natural landscape, as well as conditions of the transportation facilities and the consumption in the parks.

3.3 Similarities and Differences of the Lexicon Model and LDA Model

3.3.1 Similarities in Perception Topics Between the Two Models

As can be seen from the results of the overall perception analyses of the 10 parks, topics of recreational activities and aesthetic appreciation from the lexicon model, as well as topics of transportation and tickets, spring view, collective memory and perception, and social activities from the LDA model were most frequently perceived. It shows that visitors to these parks paid more attention to whether their own social recreation needs and aesthetic needs (by enjoying the natural landscape) were satisfied. Meanwhile, they cared about the status of transportation facilities and consumption in parks.

Correlation can be found from the distribution of the 9 perception topics used in the lexicon model and the 10 topics generated from the LDA model. The results of the correlation

4. Word cloud for reviews related to Yuyuantan Park

5. Word cloud for reviews related to Fragrant Hills Park

analysis of these topics between the two models are shown in Table 5.

Among the perception topics obtained from the LDA model and the lexicon model, there were strong correlations between spring view and environmental improvement, biodiversity, recreational activities, and aesthetic appreciation; religious culture and history and culture and religion; hiking activities and religion; autumn view and aesthetic appreciation; social activities and recreational activities and social interactions; collective memory and perception and education; cultural history and history and culture, aesthetic appreciation, and education. In addition, the topic of physical and mental recovery in the lexicon model, as well as topics of transportation and tickets, featured constructions, and gatherings and performances from the LDA model had weak correlations with other perception topics (Fig. 6).

Results from both models revealed visitors' special attention to the landscapes of nature and cultural history, as well as recreational activities. Besides, results from the LDA model reflected a comprehensive perception of different natural landscapes and sightseeing activities, such as seasonal landscape perceptions that include botanical landscapes, aesthetics, and excursion activities. Meanwhile, the LDA model classified recreational activities into more specific topics, such as gatherings and performances and hiking activities. Different from the clear classification of topics in the lexicon model, the LDA model generated topics with less distinct differences. For example, it might be difficult to effectively differentiate between visitors' leisure activities and appreciation activities.

3.3.2 Differences in Perceived Contents Across Topics Under the Two Models

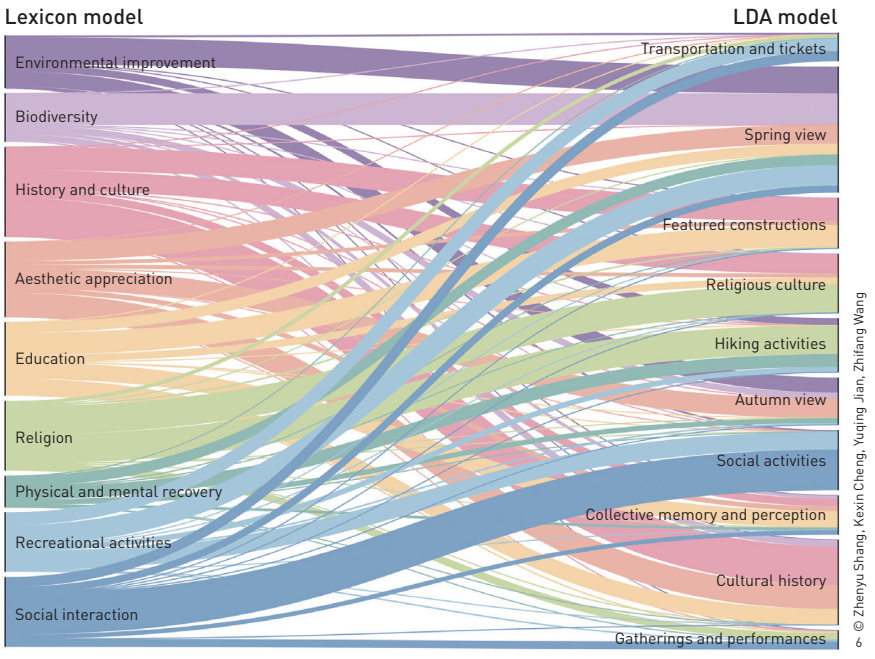
Various types of visitors' perception topics were presented significantly different across parks under the two models. In terms of the results of individual parks, it can be found from Table 2 and Table 4 that a significant difference existed between visitors' perceived topic types obtained by the LDA model and those related to urban parks concluded from the literature review. Firstly, the perception topics extracted by the LDA model were distinct in different parks. For example, worship activities and hiking activities were only presented in one or two parks. Secondly, topics obtained by the LDA model were mostly those with high perception frequency, excluding the low-frequency ones. Thirdly, these finely classified topics explicitly represented characteristics of each park and covered more contents than what were included in the lexicon model. One example was that in Yuyuantan Park, cherry blossom and related activities were frequently perceived by visitors—for the topic of the cherry blossom festival, there was a review that "it's so crowded, affecting the viewing experience," while for the topic of cherry blossom view, a visitor reviewed that "they are very beautiful; the wind blows and cherry blossoms fall." In contrast, the lexicon model was able to obtain all the given perception contents (Fig. 3), even when the topics were perceived less frequently. However, the perception topics and contents covered were limited by the range of the lexicon, resulting in an emphasis on visitors' perception of landscape services that were manually selected and neglect of their perception of the surrounding environment and landscape elements. For example, transportation conditions and

Table 5: Correlation analysis of the park perception topics between two models

Lexicon \ LDA	Transportation and tickets	Spring view	Featured constructions	Religious culture	Hiking activities	Autumn view	Social activities	Collective memory and perception	Cultural history	Gatherings and performances
Environmental improvement	0.012	0.172	−0.088	−0.019	0.045	0.095	0.011	−0.054	−0.030	−0.059
Biodiversity	—**	0.200	−0.049	—**	−0.126	0.039	−0.088	0.020	0.041	0.008
History and culture	−0.076	−0.066	0.148	0.126	−0.050	−0.053	−0.131	0.049	0.248	—**
Aesthetic appreciation	−0.063	0.121	0.024	0.024	−0.078	0.125	−0.123	0.031	0.153	−0.057
Education	−0.026	0.069	0.141	0.045	−0.067	−0.050	−0.044	0.104	0.101	0.009
Religion	0.023	−0.070	0.010	0.182	0.183	−0.020	−0.038	−0.049	−0.051	0.046
Physical and mental recovery	—**	0.067	−0.063	−0.036	0.082	0.032	−0.002	0.017	−0.024	−0.050
Recreational activities	0.082	0.130	−0.100	−0.030	0.034	0.010	0.113	−0.081	−0.062	0.009
Social interaction	0.063	0.045	−0.043	−0.073	−0.025	−0.092	0.259	0.028	−0.157	0.050

NOTE

** means $P > 0.05$, i.e., there is no significant correlation between the two topics of perception.



6. Sankey diagram showing the connection between different topics from the two models

ticket prices were frequently seen in the results of the LDA model but never in that of the lexicon model.

4 Discussion

4.1 Advantages and Disadvantages of the Two Models in Analyzing Perception Texts for Parks

The comparative analysis shows that there was a significant difference in the classification of perception topics between the lexicon model and the LDA model. Specific advantages and disadvantages of the two models can be summarized based on the

Table 6: Advantages and disadvantages of the lexicon model and LDA model applied to park perception analysis research

Research subject		Lexicon model	LDA model
Perception topic classification	Advantages	1) Clear topic classification to make the topics different from each other and with specific contents 2) Effective identification of perception contents with limited attention 3) Analysis results available for parallel comparison between parks	More comprehensive, real-time reflection of visitors' perception contents
	Disadvantages	Little consideration of the actual use of the park, resulting in a possible lack of perception contents	1) Lack of sensitivity to low-frequency perception topics due to failing to extract perception contents less frequently mentioned 2) Incapability of making parallel comparison between parks
Perception content identification	Advantages	Precise identification of the perception contents, covering topics with low perception frequency	1) Emphasis on the park characteristics, by a detailed classification of the perception topics with high awareness level 2) Clear perception contents to ensure quick identification of the topics
	Disadvantages	1) Words extracted basing on the lexicon, for which the analysis results rely heavily on the completeness of the lexicon 2) Further manual interpretation of specific perception contents required	Relatively blurred boundaries between perception topics
Scope of application	Advantages	Suitable for regional-scale, multi-park perception analysis and inter-park comparison, with a comprehensive lexicon that can be adapted to different research objects	More effective perception analysis for individual parks
	Disadvantages	Higher requirements for the lexicon to make it adjustable	Ambiguous results for regional-scale, multi-park perception analysis, making manual interpretation difficult

classification of park perception types, identification of perception contents, and the scope of model application (Table 6).

4.2 Application Suggestions for Combining the Two Models

Based on the research results, a possible optimizing strategy is to expand lexicon contents of the lexicon model, including the perception topics and words identified by the LDA model. For example, the LDA model identified high-frequency words depicting perception contents like transportation facilities outside parks and tickets. These words can be added to the lexicon by Word2vec.

To improve their applicability, we may combine the two models, allowing for their respective characteristics and advantages. When carrying out perception analysis of parks at the regional scale, the lexicon model can be used to analyze the current situation and provide a basis for the construction, management, and improvement of the parks; then based on these results, we can select perception topics that require further assessment by the LDA model. For the perception analysis of individual parks, start by using the LDA model to identify the park's characteristics and items that draw visitors' attention, then conduct a more comprehensive analysis with a lexicon model optimized based on these results to

identify any problems and propose corresponding suggestions for park improvement.

5 Conclusions

In recent years, research on social sensing analysis has paid more attention to the use of spontaneous reviews from big data, aiming to extract valuable information through semantic analysis. The accumulation of social media data and the continuous optimization of analysis methods have enriched the research contents of social sensing, better reflecting the sentimental and cognitive situation of users' interaction with space. Differing from the earlier studies that focused on scenic spots and tourist destinations[64][65], more and more current research surveys smaller-scale urban parks[66][67] utilizing a variety of methods for data analysis. However, there is a lack of comparative studies on these methods and exploration of their applicability. To fill the gap, this study employed two commonly used topic analysis models for text data, i.e. the lexicon model and the LDA model, to analyze the same research objects separately, explore the differences in the application of the two models in researching visitors' perception of urban parks, and finally clarify each model's strengths, weaknesses, and optimization paths. The results can not only guide the construction and management of urban parks but also provide a reference for relevant research on social perception through text analysis.

There are still some limitations in this study. In terms of data sources, the review text from Dianping provides little information for user profiles, making the analysis difficult to fully reflect visitors' perceptions of urban parks. In addition, the unsupervised LDA model cannot control the classification results. In response to this problem, there have been improved semi-supervised and supervised machine learning topic classification models[68][69], which need to be further explored. Finally, in addition to the two models studied in this research, there are many other text classification models based on big data and different algorithms, each with its advantages and disadvantages. Future research also needs to probe into these characteristics.

RESEARCH FUNDS

- Research on the Formation Mechanism of Recreational Service Values of Urban Green Space System Based on Multi-source Data, the National Natural Science Foundation of China (No. 42271300)
- Mechanism of Social Emotion's Spatial Heterogeneity in complex Urban System, the National Natural Science Foundation of China (No. 41871153)

REFERENCES

[1] Ferreira, A. P., Silva, T. H., & Loureiro, A. A. (2020). Uncovering spatiotemporal and semantic aspects of tourists mobility using social sensing. *Computer Communications*, (160), 240–252.

[2] Li, Y., Guo, J., & Chen, Y. (2022). A new approach for tourists' visual behavior patterns and perception evaluation based on multi-source data. *Journal of Geo-information Science*, 24(10), 2004–2020.

[3] Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., & Shi, L. (2015). Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*, 105(3), 512–530.

[4] Liu, Y. (2016). Revisiting several basic geographical concepts: A social sensing perspective. *Acta Geographica Sinica*, 71(4), 564–575.

[5] Mao, T., Wu, Y., & Huang, W. (2023). Content mining and sentiment analysis of online comments for ethnic museums in autonomous regions. *Economic Geography*, 43(8), 229–236.

[6] He, X. (2019). *Research on Social Sensing and Spatiotemporal Pattern of Xiong'an New District Based on Weibo Data* [Master's thesis]. Hebei Normal University.

[7] Zhang, S., & Zhou, W. (2018). Recreational visits to urban parks and factors affecting park visits: Evidence from geotagged social media data. *Landscape and Urban Planning*, (180), 27–35.

[8] Donahue, M. L., Keeler, B. L., Wood, S. A., Fisher, D. M., Hamstead, Z. A., & McPhearson, T. (2018). Using social media to understand drivers of urban park visitation in the Twin Cities, MN. *Landscape and Urban Planning*, (175), 1–10.

[9] Li, F., Li, F., Li, S., & Long, Y. (2019). Deciphering the recreational use of urban parks: Experiments using multi-source big data for all Chinese cities. *Science of the Total Environment*, (701), 134896.

[10] Liang, H., & Zhang, Q. (2021). Temporal and spatial assessment of urban park visits from multiple social media data sets: A case study of Shanghai, China. *Journal of Cleaner Production*, (297), 126682.

[11] Van Berkel, D. B., Tabrizian, P., Dorning, M. A., Smart, L., Newcomb, D., Mehaffey, M., Neale, A., & Meentemeyer, R. K. (2018). Quantifying the visual-sensory landscape qualities that contribute to cultural ecosystem services using social media and LiDAR. *Ecosystem services*, (31), 326–335.

[12] Oteros-Rozas, E., Martín-López, B., Fagerholm, N., Bieling, C., & Plieninger, T. (2018). Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. *Ecological Indicators*, (94), 74–86.

[13] Richards, D. R., & Friess, D. A. (2015). A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. *Ecological Indicators*, (53), 187–195.

[14] Pan, Y., & Li, J. (2021). Landscape preference based on user-

generated photograph metadata: The case of Xixi National Wetland Park. *Natural Protected Areas*, (1), 100–108.

[15] Zhu, X., Gao, M., Zhang, R., & Zhang, B. (2021). Quantifying emotional differences in urban green spaces extracted from photos on social networking sites: A study of 34 parks in three cities in northern China. *Urban Forestry & Urban Greening*, (62), 127133.

[16] Wartmann, F. M., Acheson, E., & Purves, R. S. (2018). Describing and comparing landscapes using tags, texts, and free lists: An interdisciplinary approach. *International Journal of Geographical Information Science*, 32(8), 1572–1592.

[17] Yan, Y., Chen, J., & Wang, Z. (2020). Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations. *Applied Geography*, (123), 102306.

[18] Marcotte, C., & Stokowski, P. A. (2021). Place meanings and national parks: A rhetorical analysis of social media texts. *Journal of Outdoor Recreation and Tourism*, (35), 100383.

[19] Bai, H., Song, Z., Liang, S., Zhang, P., & Zhang, G. (2023). Imagery perception analysis and comprehensive attraction evaluation of tourism destinations based on Internet text data—Taking Nanjing City as example. *Areal Research and Development*, 42(4), 89–94.

[20] Zhao, Y., Pang, S., & Wu, Z. (2021). Research on geographic semantic ontology model based on social sensing data for emergency management of events. *Information Science*, (2), 44–53.

[21] Chen, Y., Gong, C., Fan, Y., Li, X., Liang, Y., & Hu, M. (2022). Spatio-temporal variation assessment of urban waterlogging in Zhengzhou using social media data. *Journal of China Hydrology*, 42(3), 26, 48–52.

[22] Li, S., Zhao, F., Zhou, Y., Tian, X., & Huang, H. (2022). Analysis of public opinion and disaster loss estimates from typhoons based on Microblog data. *Journal of Tsinghua University (Science and Technology)*, 62(1), 43–51.

[23] Yang, B., & Zhang, J. (2017). Research on tourism image and perception of Tianmu Mountain based on network text analysis—Based on travel notes and comments of Ctrip. *Journal of Fujian Forestry Science and Technology*, 44(4), 118–125.

[24] Wang, X., & Xia, M. (2018). Research on tourist preference and satisfaction in Huangshan Scenic Spot based on network review data. *Tourism Overview*, (18), 59–60.

[25] Wight, A. C. (2020). Visitor perceptions of European Holocaust Heritage: A social media analysis. *Tourism Management*, (81), 104142.

[26] Xu, Z., Dong, J., Chen, Z., Fu, W., Wang, M., & Dong, J. (2021). Image Perception of the historical ancient town scenic spot of Yunshuiyao. *Journal of Chinese Urban Forestry*, 19(2), 115–120.

[27] Park, S. B., Kim, J., Lee, Y. K., & Ok, C. M. (2020). Visualizing theme park visitors' emotions using social media analytics and geospatial analytics. *Tourism Management*, (80), 104127.

[28] Widmar, N. O., Bir, C., Clifford, M., & Slipchenko, N. (2020). Social media sentimentas an additional performance measure? Examples from iconic theme park destinations. *Journal of Retailing and Consumer Services*, (56), 102157.

[29] Wan, C., Shen, G. Q., & Choi, S. (2021). Eliciting users' preferences and values in urban parks: Evidence from analyzing social media data from Hong Kong. *Urban Forestry & Urban Greening*, (62), 127172.

[30] Li, L., Zhang, C., Han, L., Qing, L., & Ji, H. (2021). Research on multi-scale evaluation system of parks based on comment text—Taking Chengdu parks as an example. *Intelligent City*, (2), 3–6.

[31] Jiang, Q., Wang, G., Liang, X., & Liu, N. (2022). Research on the perception of cultural ecosystem services in urban parks via analyses of online comment data. *Landscape Architecture Frontiers*, 10(5), 32–51.

[32] Jing, F., Sun, H., & Long, D. (2017). Tourist experience elements structure characteristics analysis of Xixi National Wetland Park based on web text. J*ournal of Zhejiang University (Science Edition)*, 44(5), 623–630.

[33] Wang, X., & Li, X. (2017). Research on the analysis of social services value of forest park in Beijing based on network big data. *Chinese Landscape Architecture*, (10), 14–18.

[34] Zhao, S., & Liu, B. (2019). Research on visitor perception of urban parks based on analysis of network text data—Take the main urban area of Nanjing as an example. *2019 Urban Development and Planning Proceedings* (pp. 263–272). Chinese Society for Urban Studies.

[35] Gao, X., Jin, Y., Wang, X., & Hao, J. (2021). Research on product perceptual evaluation method based on online review mining. *Modern Manufacturing Engineering*, (12), 13–20.

[36] Lu, X. (2014). Research on text clustering algorithm based on K-means. *Computer Programming Skills & Maintenance*, (24), 33–35.

[37] Wang, D., Li, J., & Shi, Y. (2020). Methods of government document clustering based on K-means algorithm. *Software Guide*, 19(6), 201–204.

[38] Ma, W., Chen, G., Li, X., Su, W., Chai, Y., Pu, Y., Zeng, J., & Liu, X. (2021). Chinese comment classification based on Naive Bayesian algorithm. *Journal of Computer Applications*, 41(S2), 31–35.

[39] Permana, F. C., Rosmansyah, Y., & Abdullah, A. S. (2017). Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter social media. *Journal of Physics: Conference Series*, (893), 012051.

[40] Han, X., & Li, Y. (2022). Research on the influencing factors of social media rumor-refuting information dissemination effect in emergencies. *Information Studies: Theory & Application,* 45(8), 97–103.

[41] Zeng, Y., Li, Z., & Zhou, Y. (2020). Article feature extraction and flow control based on text mining. *Electronic Technology & Software Engineering*, (2), 176–177.

[42] Wang, Z., Miao, Y., Xu, M., Zhu, Z., Qureshi, S., & Chang, Q. (2021). Revealing the differences of urban parks' services to human

wellbeing based upon social media data. *Urban Forestry & Urban Greening*, (63), 127233.
[43] Wang, Z., Zhu, Z., Xu, M., & Qureshi, S. (2021). Fine-grained assessment of greenspace satisfaction at regional scale using content analysis of social media and machine learning. *Science of the Total Environment*, (776), 145908.
[44] Zheng, T., Yan, Y., Zhang, W., Zhu, J., Wang, C., Rong, Y., & Lu, H. (2022). Landscape assessment on urban parks using social media data. *Acta Ecologica Sinica*, 42(2), 561–568.
[45] Taecharungroj, V., & Mathayomchan, B. (2019). Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management*, (75), 550–568.
[46] Dong, S., & Wang, Q. (2019). LDA-based tourist perception dimension recognition: Research framework and empirical research—Taking the National Mine Park as an example. *Journal of Beijing Union University (Humanities and Social Sciences)*, 17(2), 42–49.
[47] Liang, C., & Li, R. (2020). Tourism destination image perception analysis based on the Latent Dirichlet Allocation model and dominant semantic dimensions: A case of the Old Town of Lijiang. *Progress in Geography*, 39(4), 614–626.
[48] Song, Y., Wang, R., Fernandez, J., & Li, D. (2021). Investigating sense of place of the Las Vegas Strip using online reviews and machine learning approaches. *Landscape and Urban Planning*, (205), 103956.
[49] Zhou, W. (2021) *Research on Tourism Destination Evaluation Based on Improved AHP of LDA: A Case Study of 5A Scenic Spots in Jiangxi Province* [Master's thesis]. Jiangxi University of Finance and Economics.
[50] Beijing Statistics Bureau. (2021). *Beijing statistics yearbook*. China Statistics Press.
[51] Zhu, Z. (2020). *An Assessment Framework of Green Space Satisfaction Using Social Media Data: Content Analysis with Machine Learning* [Master's thesis]. Peking University.
[52] Wang, Z., Miao, Y., Zhu, Z., Zhou, J., & Wang, S. (2020). *A method for landscape service identification of parks* (No. CN111310444A). China National Intellectual Property Administration.
[53] Buchel, S., & Frantzeskaki, N. (2015). Citizens' voice: A case study about perceived ecosystem services by urban park users in Rotterdam, the Netherlands. *Ecosystem Services*, (12), 169–177.
[54] Huang, S., Pearce, J., Wen, J., Dowling, R. K., & Smith, A. J. (2020). Segmenting Western Australian national park visitors by perceived benefits: A factor-item mixed approach. *International Journal of Tourism Research*, 22(6), 814–824.
[55] Willemen, L., Verburg, P. H., Hein, L., & van Mensvoort, M. E. (2008). Spatial characterization of landscape functions. *Landscape and Urban Planning*, 88(1), 34–43.
[56] Sun, R., Li, F., & Chen, L. (2019). A demand index for recreational ecosystem services associated with urban parks in Beijing, China. *Journal of Environmental Management*, (251), 109612.
[57] van Riper, C. J., Kyle, G. T., Sutton, S. G., Barnes, M., & Sherrouse, B. C. (2012). Mapping outdoor recreationists' perceived social values for ecosystem services at Hinchinbrook Island National Park, Australia. *Applied Geography*, 35(1–2), 164–173.
[58] Wang, J., Wang, M., & Du, B. (2019). A study of the change trend of social concern in the field of consumption in China—The LDA Model analysis based on the text of Daily Economic News List in People's Daily Online (2007—2017). *Journal of Baoding University*, 32(2), 41–49.
[59] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, (3), 993–1022.
[60] Brandt, T., Bendler, J., & Neumann, D. (2017). Social media analytics and value creation in urban smart tourism ecosystems. *Information & Management*, 54(6), 703–713.
[61] Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). Association for Computing Machinery.
[62] Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 952–961). Association for Computational Linguistics.
[63] Syed, S., & Weber, C. T. (2018). Using machine learning to uncover latent research topics in fishery models. *Reviews in Fisheries Science & Aquaculture*, 26(3), 319–336.
[64] Chen, Y., Zhu, Y., & Fu, G. (2022). Visitor perception toward outstanding universal value of Xinjiang Tianshan—Based on web text analysis. *Special Zone Economy*, 398(3), 124–128.
[65] Liu, Q., Wang, X., & Liu, J. (2022). Study on relationship among tourist perceived value, satisfaction and environmental responsibility behavior in forest park. *Ecological Economy*, 38(2), 137–141.
[66] Cao, K., & Chen, Y. (2021). Service evaluation of Shenzhen parks based on social data. *Special Zone Economy*, (4), 127–129.
[67] Ye, Y., & Qiu, H. (2022). Urban park image perception based on network text analysis. *Journal of Chinese Urban Forestry*, 20(1), 90–95.
[68] Han, D., Wang, C., & Xiao, M. (2018). Text categorization scheme based on semi-supervised learning and Latent Dirichlet allocation model. *Computer Engineering and Design*, 39(10), 3265–3271.
[69] Guo, X., Ding, J., Jiang, H., & Chen, Z. (2020). ZeroNet text content analysis based on semi-supervised LDA topic model. *Information Technology*, (3), 32–38.

https://doi.org/10.15302/J-LAF-1-020083 收稿日期 | 2023-02-13 修改日期 | 2023-05-23 录用日期 | 2023-10-20 中图分类号 | TU986, TP391.1 文献标识码 | A

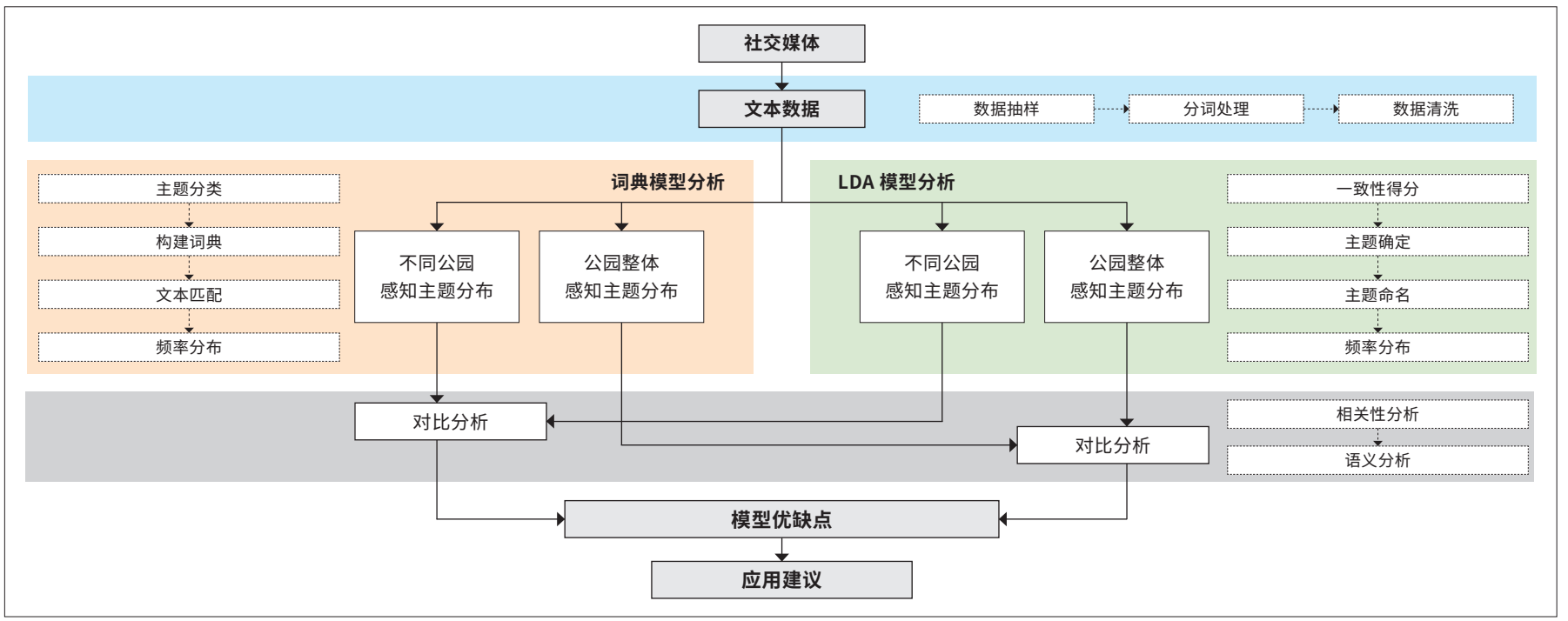
社交媒体文本数据分析方法对比与适用性研究：以北京市城市公园感知为例

尚珍宇，程可欣，简钰清，王志芳*

北京大学建筑与景观设计学院，北京 100080

*通讯作者
地址：北京市海淀区燕东园北小街3号
邮编：100080
邮箱：zhifangw@pku.edu.cn

图文摘要



文章亮点

- 探讨了两种文本分析模型各自的优劣及专业适用性
- 词典模型更有利于感知对象的横向对比研究
- LDA模型更能反映单个感知对象的特色
- 综合两种模型的优势对优化景观感知评估有重要意义

摘要

互联网科技和媒体的蓬勃发展产生了大量社交媒体数据，基于自发评论文本进行社会感知分析逐渐成为研究热点，也开始被应用于城市公园使用情况和情感认知的研究中。鉴于现有研究大都应用单一的文本分析方法，本研究尝试开展社交媒体文本数据分析方法的对比研究，并揭示其在公园感知研究中的优缺点和适用性。

关键词

社会感知；
文本分析；
词典；
隐含狄利克雷分布（LDA）；
城市公园；
景观感知

研究选择在相关领域广泛应用的词典模型和LDA模型，以大众点评网站上北京10座城市公园的点评文本为研究数据，分别从单个公园和公园整体使用感知两个层面进行文本分析，并对比分析感知主题的分类结果。结果表明：词典模型更有利于在公园间进行横向对比分析；LDA模型则可以直观显示公园特色和游客感知偏好；综合运用两种模型可优化公园感知评估。两种方法揭示了北京城市公园游客对公园的关注主要集中于社交活动的需求、自然景观带来的视觉审美需求，以及交通设施状况和城市公园消费情况。本研究既可为社交媒体文本分析方法的选择和使用提供优化建议，又可为公园建设与管理改进提供依据与指导。

编辑 王颖，周佳怡
翻译 王颖，尚珍宇，周佳怡

1 引言

随着互联网科技的高速发展，人们越发频繁地通过网络媒体进行社交活动，由此生成的海量信息为社会感知的研究提供了数据基础[1][2]。这类研究主要通过挖掘大数据中所包含的人类行为特征信息，分析人对于空间场所的认知感受、空间流动模式和个体间的社交关系[3][4]；随着近年来社交媒体数据量的增长，对于地理空间的情感认知分析研究逐渐增多[5][6]。社交媒体的数据类型主要包括签到数据、带有地理位置的图片数据和文本数据三大类，研究方法早期多集中在通过签到数据识别到访率和动机偏好分析[7]~[10]，以及结合照片图像内容及其地理位置进行的感知情绪分析[11]~[15]。近年来，随着文本数据的直观性逐渐受到认可，通过文本数据挖掘进行感知分析的研究也开始起步并日渐增多[16]~[19]。利用社交媒体文本数据进行社会感知分析的研究对象主要包括热点话题舆情和公共事件风险灾害反应[20]~[22]，以及公共设施现状的使用感知，后者的相关研究大多关注风景旅游区[23]~[26]和城市绿地[27]~[29]的使用后感知。城市公园作为重要的公共开放空间，为居民提供了亲近自然、休闲娱乐、放松游憩等服务[30][31]。通过文本数据进行公园感知研究——即对公园游客的使用后评价进行感知分析，以此作为城市公园建设和存量更新的方向与依据——正逐渐受到学者们的关注。

社交媒体文本数据的分析方法主要包括词频分析和语义分析[32]~[34]。随着文本挖掘技术的发展，目前已经可以通过建立文本分析模型来挖掘文本所呈现的内在规律及主题，主题模型的运用开始成为感知分析和满意度评价的基础。常见的文本分析主题模型包括基于词典规则的分类分析模型（下文简称“词典模型”）[35]、K-means模型[36][37]、隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）模型[5]、朴素贝叶斯模型[38][39]、线性和逻辑回归模型[40]、随机森林与决策树模型[41]等。已有研究在进行感知分析时，通常单独采用其中一种模型进行文本数据处理，鲜少探讨不同模型之间的优劣及专业适用性。

本研究尝试开展社交媒体文本数据分析方法的对比研究，并揭示其在公园感知研究中的适用性。由于词典模型和LDA模型在风景名胜区和城市公园感知研究中应用广泛，本研究针对二者展开对比分析。词典模型大多先对获取的文本高频词进行语义分析，而后通过已有的研究标准体系建立相应的语义词典对不同词语进行分类，通过内容扩充使词典更加完善，最后依据词典对文本内容进行进一步分类分析[42]~[44]。LDA模型是一种基于机器学习的主题提取模型，主要在文本分析中用于主题提取和分类[45]~[49]。本研究聚焦于以下问题：在对基于公园感知的社交媒体文本进行分析时，词典模型和LDA模型的感知研究过程与分析结果存在怎样的差异？两种模型的优劣是什么？此基础上，研究团队进一步探究如何利用两种模型的优势为城市公园规划提供指导，并总结文本分析方法在公园感知研究中的适用价值。

2 数据处理与研究方法

2.1 研究区概况与数据来源

北京市市域面积约为16 410km^2，截至2020年常住人口达2 189万，拥有各类公园1 050个，公园绿地面积累计达357.2km^2。[50]作为互联网技术发展较为迅速的超一线城市，居民的社交媒体使用频率相对较高，为本研究提供了大量的数据基础。

本研究选择大众点评网作为文本数据来源。大众点评网是中国拥有海量点评、极具影响力的社交点评平台之一，活跃用户数量逐年增加，用户评价主动参与度不断提升，因此其点评数据具有较高的准确率[51]。研究使用Python软件中的Request模块获取大众点评网北京市公园目录下

自2006年4月至2020年9月的所有文字点评数据和点评者信息，最终选取点评数量排名前10位的城市公园作为研究对象（表1）。

为保证模型分析的准确性，研究对获取的文本数据内容进行了预处理。首先剔除文字量较少的评价文本[51]，仅保留字符数大于50的文本数据。筛选后评价数量最少的公园为北京园博园，共计6 531条，以此为标准使用SPSS软件分别对其他各个公园的评价数据进行相应的完全随机抽样，最终获得65 310条点评文本数据。

研究选用Python语言工具jieba分词对数据进行分词。相比于其他中文分词工具，jieba分词具有自定义词典功能，能够更加准确有效地进行文本数据分词处理，在语言环境的适配上也更有优势[51]。将《哈工大停用词表》等多个停用词表合并去重后建立中文停用词典，以此筛选出文本数据中无意义的符号和文字语言，对文本数据进行清洗[48]。考虑到不同的中文用词可能语义相近，为提高模型的准确性和处理效率，利用以《哈工大信息检索研究中心同义词词林（扩展版）》为基础的同义词词典对分词结果进行同义词替换[46]。在此基础上，根据实际使用情况，针对城市公园感知内容，人工筛查及调整分词和同义词替换结果，还原本研究中不恰当的同义词替换内容，例如将“樱花”“水仙”“海棠花”等替换为“香菊片”，将“海淀区”“朝阳区”等替换为“白云区”等。

2.2 研究方法

2.2.1 基于词典的感知主题分类模型

词典模型采用王志芳等人于2021年提出的基于景观服务的城市公园感知主题分类评估模型，该模型经过词典有效性检验，整体性能测试结果优良[42]。景观服务研究重视景观格局和功能的综合作用，聚焦于服务供给者和需求者之间的空间过程和关系；而公园是重要的城市绿地景观，选用该模型能够有效反映游客对公园使用的感知评价情况[42]。在本研究中，模型运用Python编程对预处理后的评价文本数据进行结构化处理并提取高频词；之后进行人工分类，构建中文景观服务感知词典；继而利用Word2vec词向量模型和人工添加的方式扩建词典内容[52]，并划分到不同的感知主题类别中。根据已有的文献研究[53]~[57]，共划分出9类含义不同的公园景观服务感知主题（表2）[42]。

将获取的公园感知评价文本数据与词典进行匹配，以此识别评价数

表 1：选取公园情况

序号	名称	面积（hm^2）	评价数量（条）
1	圆明园遗址公园	350.0	17 805
2	玉渊潭公园	129.4	17 698
3	香山公园	188.0	13 825
4	景山公园	23.0	13 628
5	北京世园公园	503.0	11 923
6	朝阳公园	288.7	11 750
7	北京世界公园	53.3	11 338
8	奥林匹克森林公园	680.0	10 673
9	八大处公园	332.0	9 889
10	北京园博园	513.0	8 736

表 2：基于景观服务的城市公园感知主题分类与词语示例

主题分类	含义	词语示例
环境改善	空气质量改善、微气候调节、噪声	湿度、暴晒、清新、风和日丽
生物多样性	动物、植物	天鹅、冬青、枯木、鸟语花香
历史文化	文史价值、文化遗产、历史景点	清朝、遗迹、艺术、端午
美学欣赏	风景、美、提供灵感	赏花、摄影、好看、大煞风景
教育价值	科普活动、教育	知识、学问、渊博、孤陋寡闻
宗教信仰	宗教祭拜、信仰、避难	祭祀、佛珠、道教、开光、求姻缘
身心修复	放松、压力释放、心灵修复	舒心、美好、垂头丧气、手舞足蹈
娱乐活动	户外活动、体育运动	散步、划船、爬山、打打球
社会交往	社会融合、个体间交流	爸妈、同事、亲戚朋友、聚会

据中的用词，进而提取出单条评价中所涉及的感知主题，用于计算各类主题的感知频率。将涉及某项感知主题的评论数量与总评论数量的比值作为相应景观服务主题在该公园的感知频率[42]，即：

$$F_i = N_i / M \,, \qquad (1)$$

式中，F_i为第i项主题的感知频率，N_i为提及第i项主题内容的评价数量，M为该公园的评价总数量。

2.2.2 基于LDA的感知主题分类模型

LDA是一种通过计算机来自动分析文本的语言处理模型，能够快速从非结构化文本（即文档）中提炼出主题，实现对文档的降维[58]。该模型认为，一篇文档由若干主题以一定概率分布构成，而每个主题又由若干词语以一定概率分布构成，概率值越大说明二者关系越紧密[59][60]。LDA模型可以计算“文档—主题”和“主题—词语”两类概率分布，从而实现对文档主题和对应词语（关键词）的分类。

本研究使用Python软件的gensim工具包调用LDA模型，实现文本数据主题分析。主题数量K值的选取需要综合考虑主题颗粒度大小和主题内容的可解释性，以及是否便于进行对比分析。本研究中的主题数量主要通过计算主题一致性得分来确定，这一指标能够有效计算各主题所含关键词之间的相似度，越高的一致性得分说明模型在该数量主题下的分析效果越好[61]~[63]，最后结合人工对一致性得分较高的主题进行筛选，确定合适的主题数量以获得理想的模型运算结果。获得结果后，对于每个主题的实际权重（即不同主题的感知频率）进行计算，具体方式如下。

1）确定主题数量为K，总评价数量为N；

2）求取K个主题在每条评价中的预期概率F_0，即$F_0=1/K$；

3）通过LDA模型获得第j个主题在每一条评价中的实际概率为F_j（j=1，2，3，……，K），对比F_j与F_0的值；

4）统计$F_j > F_0$的评价数量为A_j；

5）获取第j个主题的实际权重，为$Q_j=A_j/N$。

针对每个公园数据各自的分析结果，分别进行主题命名——由三位景观设计专业的研究人员共同探讨，结合关键词内容和相应的高权重评价文本进行不同的主题命名，同时去除权重较低且感知内容相关性较弱的主题，即“噪声”主题。

2.2.3 主题分布相关性分析

对两种模型得到的不同感知主题的分布进行相关性分析。不同感知主题在每条评价文本中的分布情况为二分类变量，结果为“是”/“否”（分别记为“1”/“0”）两项，因此在SPSS软件中计算Phi系数，进行相关性检验。Phi系数主要通过2×2的二元变量数值列联表来计算相关性。如图1所示，当数值主要集中在主对角线上时，说明不同变量间的分布相关性较高，系数计算如公式（2）所示：

$$\Phi = \frac{N_{11}N_{00} - N_{10}N_{01}}{\sqrt{(N_{11}+N_{10})(N_{11}+N_{01})(N_{10}+N_{00})(N_{01}+N_{00})}} \,, \qquad (2)$$

其中，X、Y分别表示两类变量的值（“1”/“0”），N_{11}、N_{10}、N_{01}、N_{00}表示不同变量值下的统计数量，Φ表示两个变量分布的相关性系数。运用SPSS软件中的Phi系数相关模块进行数据分析，当显著性小于0.05时，Φ值越接近1，两者分布相关性越高。

2.2.4 主题内容语义分析

本研究使用Python语言对评价文本进行词频分析，通过词云图表达不同文本数据中被使用者提及频次较高的词语内容，出现频率越高的词语显示越突出，高频词汇的展示可以有效表达所选评价文本的主要内容，对不同公园所涉及评价文本的词频分析可以帮助获取各公园的感知主题内容。

2.3 技术路线

本研究基于北京市10座城市公园的大众点评评价文本数据，利用两种模型分别从单个公园和公园整体使用感知两个层面进行文本分析，并对比分析感知主题的分类结果，具体的技术路线如图2。

3 研究结果与分析

3.1 词典模型更适用于公园间横向对比

词典模型因采取了人工预设而具有主题类型覆盖度相对全面、类别界限明确的特点，对于所有公园的感知分析结果都被限定在词典划定的标准内，使解读更加清晰，有利于对公园间的感知频率与感知差异进行横向对比分析。

利用词典模型对北京市10座城市公园的感知主题进行分类统计，结果显示，游客对各公园不同主题的感知频率存在明显差异（图3）：圆明园遗址公园（3.40）和奥林匹克森林公园（3.40）的感知总频率最高，北京世界公园与朝阳公园的评价中提到的各项感知总频率相对较低（前者为2.90，后者为2.97）；奥林匹克森林公园不同类型的主题感知频率差异最大（娱乐活动为0.88，宗教信仰为0.02）。此外，在不同公园的各项感知主题中，娱乐活动和美学欣赏均表现出较高的游客感知频率，教育价值和宗教信仰的感知频率普遍较低。圆明园遗址公园在历史文化方面的感知频率（0.68）及八大处公园在宗教信仰方面的感知频率（0.45）最高，而景山公园的美学欣赏感知频率最为突出（0.79），朝阳公园的社

表 3：基于 LDA 模型确定的不同公园的主题情况

公园名称	主题数量	主题一致性得分
圆明园遗址公园	10	0.5611
玉渊潭公园	10	0.5216
香山公园	9	0.6049
景山公园	9	0.6334
北京世园公园	9	0.5829
朝阳公园	10	0.5845
北京世界公园	6	0.6577
奥林匹克森林公园	10	0.5046
八大处公园	9	0.6385
北京园博园	8	0.5411

会交往感知频率明显高于其他公园（0.60）。除此之外，玉渊潭公园和八大处公园的教育价值感知关注度相较于其他公园有所不足。

3.2 LDA模型突出公园自身特色

由LDA模型下的感知分析结果可知，北京市10座城市公园的感知类型差异明显，社交媒体评价突出体现了公园自身的景观特色和游客感知偏好。研究分别对每个公园的评价文本进行LDA主题模型处理，根据主题一致性得分确定每个公园合适的主题数量（表3），随后结合文本可解释性进行主题命名并去除“噪声”主题。例如，在北京园博园的分析结果中，其中一个主题中的关键词主要包括“快到、检查、太阳伞、左转、风筝节、极好的、等车、3号门、十足、部门”等，感知内容不明显，且权重较低（仅为0.016），因此判断其为“噪声”主题并予以去除。最终不同的公园获得的主题分布各有不同（表4）。

通过表4可以看出，不同公园游客感知的主题数量普遍被分为8或9项，其中圆明园遗址公园、玉渊潭公园和奥林匹克森林公园的感知主题较多，北京世界公园的感知主题最少。在感知内容上公园间存在差异，但部分感知主题在多数公园中均有体现，如与交通相关的感知（包括公交、地铁、停车场等信息）在除景山公园以外的9所公园的结果中均有

表 4：基于 LDA 模型的不同公园感知主题

	圆明园遗址公园	玉渊潭公园	香山公园	景山公园	北京世园公园	朝阳公园	北京世界公园	奥林匹克森林公园	八大处公园	北京园博园
主题 1	交通门票	疫情影响	登山活动	文化遗迹	自然环境	交通门票	交通门票	休闲运动	祭拜活动	特色展园
主题 2	遗址景观	樱花节庆	公园介绍	历史变迁	服务设施	服务体验	记忆感知	交通设施	娱乐设施	审美体验
主题 3	园林景观	鲁冰花景观	交通门票	公园介绍	社交活动	亲子活动	公园介绍	社交活动	佛教景观	活动体验
主题 4	自然景观	樱花景观	体验感知	特色花卉	展馆体验	休闲娱乐	演出活动	夜景表演	集会活动	交通设施
主题 5	历史感知	门票消费	人文景观	特色建筑	公园介绍	书市活动	特色景观	体育活动	登山活动	文化活动
主题 6	爱国教育	自然景观	教育学习	周边景观	音乐节庆	公园介绍		夏秋景观	公园介绍	休闲活动
主题 7	公园介绍	交通设施	自然景观	疫情影响	服务体验	庙会活动		春季景观	记忆感知	公园介绍
主题 8	荷花欣赏	公园介绍		鸟瞰故宫	交通门票	春季活动		公园介绍	交通设施	
主题 9	特色雪糕	休闲娱乐						疫情影响		
主题 10								集会活动		

表 5：不同主题分类模型下公园感知分布的相关性分析

词典 \ LDA	交通门票	春季景观	特色建筑	宗教文化	登山活动	秋季景观	社交活动	记忆感知	人文历史	集会表演
环境改善	0.012	0.172	− 0.088	− 0.019	0.045	0.095	0.011	− 0.054	− 0.030	− 0.059
生物多样性	—**	0.200	− 0.049	—**	− 0.126	0.039	− 0.088	0.020	0.041	0.008
历史文化	− 0.076	− 0.066	0.148	0.126	− 0.050	− 0.053	− 0.131	0.049	0.248	—**
美学欣赏	− 0.063	0.121	0.024	0.024	− 0.078	0.125	− 0.123	0.031	0.153	− 0.057
教育价值	− 0.026	0.069	0.141	0.045	− 0.067	− 0.050	− 0.044	0.104	0.101	0.009
宗教信仰	0.023	− 0.070	0.010	0.182	0.183	− 0.020	− 0.038	− 0.049	− 0.051	0.046
身心修复	—**	0.067	− 0.063	− 0.036	0.082	0.032	− 0.002	0.017	− 0.024	− 0.050
娱乐活动	0.082	0.130	− 0.100	− 0.030	0.034	0.010	0.113	− 0.081	− 0.062	0.009
社会交往	0.063	0.045	− 0.043	− 0.073	− 0.025	− 0.092	0.259	0.028	− 0.157	0.050

注

** 代表 $P > 0.05$，两者之间不存在显著相关性。

涉及。除此之外，结合词频分析（图4，5）可以看出，部分感知主题因公园自身的特色表现不同，如玉渊潭公园的春季樱花景观、香山公园的秋季红叶景观等。同时，节庆活动在不同公园中也会产生独特的游客感知，如朝阳公园的春节庙会活动、奥林匹克森林公园的乐队表演活动等各类主题不同的集会活动。

将10座公园的所有评论文本数据进行LDA模型分析，结果显示，感知主题可划分为10项，其中交通门票（0.60）、春季景观（0.53）、记忆感知（0.52）和社交活动（0.48）的感知频率高于其他主题，登山活动（0.30）、人文历史（0.29）、集会表演（0.26）、秋季景观（0.20）、宗教文化（0.14）、特色建筑（0.11）的感知频率相对较低。由此可见，北京城市公园游客对公园的关注主要集中于社交活动的需求、自然景观带来的视觉审美需求，以及交通设施状况和城市公园消费情况。

3.3 词典模型与LDA模型分析的共性与差异

3.3.1 两种模型下不同感知主题共性分析

综合对10座北京城市公园整体感知的分析结果——词典模型结果中娱乐活动和美学欣赏的高感知频率，以及LDA模型结果中交通门票、春季景观、记忆感知和社交活动的高感知频率——可以看出北京城市公园游客主要关注社交游憩需求和自然景观带来的视觉审美需求是否得到满足，同时对交通设施状况和城市公园消费情况较为敏感。

基于词典模型的9项感知主题和LDA模型的10项感知主题在评价中的分布具有一定的相关性。对两类模型下的不同感知主题分布进行相关性分析，结果如表5所示。

LDA模型与词典模型分析所得的感知主题中，主题分布相关性较强的有：春季景观与环境改善、生物多样性、娱乐活动和美学欣赏；宗教文化与历史文化、宗教信仰；登山活动与宗教信仰；秋季景观与美学欣赏；社交活动与娱乐活动、社会交往；记忆感知与教育价值；人文历史与历史文化、美学欣赏和教育价值。除此之外，词典模型下的身心修复主题，以及LDA模型下的交通门票、特色建筑及集会表演主题与其他感知主题的分布相关性都较弱（图6）。

两种模型感知内容分类的结果均表现出对自然景观、人文历史景观和娱乐活动的关注。此外，LDA模型的分类结果更侧重于对不同自然景

观和游览活动的综合感知，如包含植物景观和审美游玩活动的季节性景观感知；同时，将娱乐活动的内容划分为了集会表演、登山活动等更具体的主题。相比于词典模型清晰的感知主题划分，LDA模型的分析结果界限相对模糊，如难以有效区分游客游览过程中出现的休闲活动与欣赏活动。

3.3.2 两种模型下不同主题感知内容差异分析

两种模型下的游客感知主题类型在不同公园的表现存在明显差异。在单个公园的分析中，对比表2和表4，可以发现经LDA模型运算获得的游客感知主题类型与基于文献研究获得的城市公园感知主题类型存在明显差异。LDA模型提炼出的感知主题在不同公园中体现的内容各有不同，例如祭拜活动和登山活动等主题仅在个别公园中有所呈现；模型所获感知主题几乎未能呈现低频感知的内容，而只呈现出感知频率较高的内容；主题更加突出公园自身的特色，类型更加细分，且存在部分词典模型未涉及的感知内容，比如玉渊潭公园的分析结果显示出游客对公园中有关樱花景观和活动的感知较为突出——包含樱花节庆主题，评价有“游客也很密集，影响观赏体验”，以及樱花景观主题，评价如“很漂亮，风吹过，樱花飘落”。相比之下，词典模型则能够捕捉到所有设定的感知内容（图3），即使感知频率较低的主题也会有所呈现。模型分析的感知主题与涵盖内容受现有词典的影响，分类分析结果更加注重游客对人工选定的不同景观服务内容的感知，识别到的对周边环境和景观要素的感知较少。例如，交通条件、门票价格等内容在LDA模型的分析结果中频繁出现，但在词典模型分析结果中均未涉及。

4 讨论

4.1 两种模型在公园感知文本分析中的优缺点

通过对比分析可以看出，基于词典模型和LDA模型的城市公园感知分析在主题类型划分上具有显著差异。可从公园感知类型划分、感知内容识别及模型适用范围梳理两种方法的具体优缺点（表6）。

表 6：词典模型与 LDA 模型应用于公园感知分析研究的优缺点

研究内容		词典模型	LDA 模型
感知类型划分	优点	1）划分的主题界限清晰、差异显著，涉及的内容范围明确 2）有效识别关注度较低的感知内容 3）分析结果支持在公园间进行有效的横向对比	全面、实时地反映游客感知内容
	缺点	缺乏针对公园实际使用现状的考量，容易缺失部分感知内容	1）对低频感知主题缺乏敏感性，难以有效提取被提及较少的感知内容 2）难以在不同公园间横向对比分析数据
感知内容识别	优点	精准识别感知内容，能够有效反映感知频率较低的主题	1）突出公园特色，详细划分游客关注度高的感知主题内容 2）感知内容表现清晰，可以快速识别感知主题内容
	缺点	1）词语的提取依托词典的内容，词典的完整性对分析结果影响较大 2）缺乏对于具体感知内容的解读，仍需进一步的人工解读	感知主题间的内容界限相对模糊
模型适用范围	优点	适合区域尺度的多公园感知分析及公园间对比，完善的词典可以适配不同研究对象	对单个公园的感知识别更有效
	缺点	对词典内容要求较高，需要根据实际情况不断调整	针对区域尺度的多公园感知分析结果不够清晰，人工解读难度高

4.2 两种模型相结合的应用建议

根据上述研究，在模型优化方面，可以基于LDA分析结果对词典模型的词典内容进行扩充，完善感知类型和词量。例如，针对公园外部交通设施及门票等相关感知内容，可提取文本分析中涉及这些内容的高频词汇，依托Word2vec向量模型对词典进行扩容。

在模型专业适用性方面，可以结合两种主题分析模型的特点和优势来判断两者结合应用的途径。进行区域尺度的公园感知分析时，可先利用词典模型进行现状分析，为公园的建设、管理和改进提供依据；再针对具体的分析结果选定需要深入挖掘的感知类型，通过LDA模型进行具体的文本分析，细化需要深入研究的公园感知内容。对于单个公园进行感知分析时，可以基于LDA模型的结果确定公园的特色和游客的关注内容，了解游客对公园使用后感知的现状，再据此优化词典模型并展开进一步分析，以期更加全面地发现问题，提出相应的改进建议。

5 结语

近年来，社会感知分析领域愈发关注对大数据中自发文本数据的运用，尝试通过语义分析来提取有价值的信息。随着社交媒体数据的积累及分析方法的不断优化，社会感知的研究内容越来越丰富，愈发深刻地反映出人与空间交互的情感和认知情况。早前的研究大多针对景区和旅游目的地等区域展开[64][65]，如今针对较小尺度城市公园的研究也逐渐增多[66][67]，且相关数据分析方法较为多样，只是尚且缺乏针对不同方法的对比研究及对各方法专业适用性的探究。本研究选择了两种最常用的文本主题分析模型——词典模型与LDA模型，对相同的研究对象进行分析，探讨两种模型的应用在城市公园感知研究中的差异，以明确其优缺点和优化途径。研究结果不仅对城市公园的建设和管理具有指导性价值，也有利于推进通过文本分析进行社会感知的相关研究发展。

本研究仍存在一定的局限性。在数据来源方面，来自大众点评网站的评价文本缺乏使用者的个人信息，无法进行有效的用户画像分析，分析结果难以全面体现城市公园游客感知情况。此外，LDA模型作为传统的无监督分类模型，无法把控分类结果。针对LDA模型问题目前已有改进的涉及半监督和有监督的机器学习主题分类模型[68][69]，有待进一步探究。最后，除了本文所探究的两种模型外，基于大数据的文本分类模型还有多种，不同的模型算法具有各自的优势和不足，后续研究也需要结合更多模型进行进一步的深化和验证。

基金项目

· 国家自然科学基金项目“基于多源数据的城市绿地系统游憩服务价值形成机制研究”（编号：42271300）
· 国家自然科学基金项目“城市复合系统中社会情绪感知的空间分异及驱动机制”（编号：41871153）