mathematics

MDPI

Article

An Exploration of Architectural Design Factors with a Consideration of Natural Aspects Based on Web Crawling and Text Mining

Dongmiao Zhao [1,2], Yufeng Liu [3], Boyi Pei [1], Xingtian Wang [1], Sheng Miao [3,*] and Weijun Gao [1,4]

[1] Innovation Institute for Sustainable Maritime Architecture Research and Technology, Qingdao University of Technology, Qingdao 266520, China
[2] School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao 266520, China
[3] School of Information and Control Engineering, Qingdao University of Technology, Qingdao 266520, China
[4] Faculty of Environmental Engineering, The University of Kitakyushu, Kitakyushu 808-0135, Japan
* Correspondence: smiao@qut.edu.cn

Abstract: Architectural construction is responsible for the consumption of large amounts of resources, so the optimization of architectural design and evaluation is significant for sustainable global development. Most architectural assessments focus on energy conservation, novel materials and eco-friendly strategies, but without agreed indicators and criteria. Since the consideration of natural aspects is somewhat fuzzy and vague, this study utilized data mining technology to explore the major factors related to relationships between buildings and nature. By employing the popular technique of web crawling, this study collected 38,320 architectural descriptions from the "Archdaily", including descriptions of 11 types of buildings, four of which were taken as typical research representatives. The 100 most frequent words were used to create a word cloud. Using Python script, all of the text was refined and processed with the word2vec model, thereby allowing to conduct Agglomerative Hierarchical Clustering (AHC). The frequency of words related to natural aspects were analyzed within 15 architectural design elements. Different building types in different areas have obvious similarities in terms of design elements, so it is feasible to adopt the same evaluation factors for the building evaluation systems of different regions. This paper mainly focuses on improving the accuracy and validity of assessment by providing basic evaluation indicators that could enhance connections between design and evaluation progress, stimulating the improvement of building environmental performance.

Keywords: text mining; building design evaluation; agglomerative hierarchical clustering; natural language processing

MSC: 68U15

Citation: Zhao, D.; Liu, Y.; Pei, B.; Wang, X.; Miao, S.; Gao, W. An Exploration of Architectural Design Factors with a Consideration of Natural Aspects Based on Web Crawling and Text Mining. *Mathematics* **2022**, *10*, 4407. https://doi.org/10.3390/math10234407

Academic Editor: Victor Mitrana

Received: 20 October 2022
Accepted: 21 November 2022
Published: 22 November 2022

1. Introduction

The United Nations has predicted that urban areas will house 66% of the world's population by 2050 [1]. Population expansion and increasing basic needs for survival force cities to develop in unappropriated ways. Buildings are responsible for the consumption of large amounts of resources [2,3], demanding nearly 36% of all energy usage and discharging 37% of global greenhouse gas emissions in 2020 [4]. With the proposal of sustainability goals, more and more design theories and new techniques are being tried in practice and green building evaluation systems are being explored. Green building rating tools (GBRTs) have been critically discussed in recent years [5]. With the very complex interactions between subjectivity and social backgrounds, the evaluation criteria for green building rating tools differ from nation to nation and the effectiveness of GBRTs is doubted as there are no widely adopted criteria. Since they mainly focus on advanced technological

promotion [6], green building evaluation factors have an obvious separation from design processes. Thus, it is necessary to identify the most appropriate indicators for both design and evaluation systems in order to adapt to sustainability requirements.

The accumulation of data has grown rapidly with the fast development of information technology. Data mining technology has been developed and is used to process large amounts of disordered data in many fields [7,8]. As a branch of data mining [9,10], text mining can extract obscure content hidden from massive texts, which is difficult to achieve through manual analysis. Text mining has been widely used in finance [11], library management [12], education [13], information science [14] and other fields. This method can also be adapted to the analysis of interactions between buildings and nature using large amounts of architectural information.

This study intends to explore and identify the major design factors with a consideration of natural aspects by utilizing text mining technology. An essential task in text mining task is to establish a text corpus, which includes potential knowledge related to the research topic. In order to extract design philosophy and major factors, this study focuses on descriptions of real architectural projects by designers. "Archidaily.com" is the most popular website including news, products, events, and a particular description repository with more than forty-eight thousand records. The repository is contributed by architects all over the world under various cultural background, geographic areas, technologies applied, and materials utilized. It is comprehensive and abundant to provide subject matter knowledge and discover connotative design factors from experienced architects. For each description, the specifications of building project are presented, such as location, area, year, surroundings, and construction technology. Architects express their design thought in the text description, which contains the most important factors to be discovered. To collect the descriptions of the building project from the "Archdaily.com" repository, this study employs web crawling technology to retrieve raw data with respect of Internet access policy.

By applying text mining tools, the text corpus was refined and processed. As this study focused on the consideration of natural aspects, the six keywords that were selected were nature, environment, green building, sustainability, landscape and ecology. Since only high-frequency co-occurrence words in the text corpus could contribute to the exploration of critical architectural factors, 60 words were retained for four building types. Considering that plain text is difficult for machines to handle and process, all of the words needed to be processed via a vectorization operation that could convert text into computable word vectors. This operation was based on word2vec modeling and implemented a CBOW algorithm. Based on the word vectors, Agglomerative Hierarchical Clustering (AHC) was utilized in this study, which is an effective and popular algorithm in text mining. AHC is an unsupervised machine learning algorithm that can discover relationships between elements within a dataset based on distance metrics. This research focused on extracting architectural design factors from mass online texts; therefore, AHC could reveal hierarchical relationships between them and produce critical indicators for building evaluation systems. Due to the vectorization operation, the cosine distance was selected as the metric for the clustering algorithm. The primary contributions can be summarized as follows:

1. Combined with data science, the introduction and application of a new method for obtaining evaluation indicators within the architecture domain;

2. The use of architectural design descriptions as a text corpus to obtain evaluation indicators using text mining, thereby enhancing connections between design and evaluation progress;

3. The provision of basic evaluation indicators that could improve the accuracy and validity of assessments for the sustainable performance of buildings, which could make evaluation results more instructive.

The rest of this paper is organized as follows. Section 2 presents related works. Section 3 states the methodology used in this research, including data collection and text mining technologies. The result analysis and discussion are presented in Sections 4 and 5 concludes the paper.

2. Review of Related Studies

As well as subjective and judgmental [15], traditional review methods are also time-consuming and labor-intensive, as are manual inspections [16]. Text mining provides an effective way to obtain useful information from massive amounts of text. It is an advanced data analytics tool that originated from multiple technologies, including text classification [17], text clustering [18,19], and automatic text summarization [20]. The basic function of automatic abstracts is to transform unstructured texts into organized formats in order to support further analysis [21]. The superiority of this method for literature reviews has been validated within many study fields. Huang et al. used it to analyze free-text medical records to obtain a better understanding of smoking cessation processes and helpful strategies for decision-making [22]. In another study, 201,141 pediatric papers were analyzed using text mining to explore the changing trends in epidemiological studies, which have been gaining more attention [23].

Today, more and more researchers are analyzing and extracting insights from the literature. Bibliometric reviews have been adopted in many building studies, including studies on green buildings [24], BIM [25], and building maintenance [26]. The application of text mining is not uncommon within the building domain. Xu et al. used test mining to extract the main safety risk factors from construction accident reports to improve safety management [27]. Abdelrahman et al. explored the relationship between data science and energy conservation in buildings by text mining about 30,000 pieces of scientific literature [28]. Combined with techniques of text mining and case-based reasoning, an integrated system was introduced by Shen et al. that could retrieve green building cases [29]. Ding et al. used text mining to automatically identify research trends within the domain of building energy management [30]. It has been observed that the usage of data science techniques has reached a saturation point within building operation and maintenance, but remains underappreciated within building commissioning and design [28].

Architectural design is a process with many intricacies. The complex social and natural backgrounds and personal experiences of architects have a great influence on design outputs. Even within the same concept of "environmental sustainability", each architect or researcher has their own design theory and philosophy. Biophilic, biomimetic, resilience, restorative, permaculture and regenerative [31], theories are all related to the concept of environmental architecture and focus on improving the long-term harmony between humans and nature [32,33]. Ecological design enhances the importance of the "biological" and "social" aspects of ecology [34]. Sustainability and energy efficiency requirements are more significant in architecture within the concept of environmentally sustainable design [35]. All of these design theories and concepts have their own priorities, even though they all have the same aim of environmental sustainability.

Currently, there are more than 10 influential green building assessment methods from regions all over the world, such as BREEAM (the UK), LEED (the USA), GB Tool (Canada), ESGB (China), CASBEE (Japan), ESCALE (France), NABERS (Australia), DGNB (Germany) and HK-BEAM (Hong Kong), etc. BREEAM (Building Research Establishment Environmental Assessment Method) was launched in 1990 and is regarded as the first evaluation system for green buildings [36], but it was replaced by the Code for Sustainable Homes (CSH) in 2008 [37]. LEED (Leadership in Energy and Environmental Design) is the most representative green building rating tool (GBRT), with projects in over 150 regions. In order to clarify the differences between these GBRTs, many researchers have studied and compared them. Energy category is the highest weighting in BREEAM, GSAS and Estidama systems, while the Indoor Environmental Quality category has more priorities in LEED [38]. Yurong et al. [37] compared the assessment methods and indicators in LEED, CSH and ESGB. Mattoni et al. compared and analyzed CASBEE, Green Star, BREEAM, LEED and ITACA and showed that their homogeneity was not equal in terms of quality and quantity [36]. The quantity and content of evaluation factors are not unified across all existing evaluation systems. In addition, there is no common understanding or concept of building performance, which impedes the development of the building domain [39].

At present, there is an obvious separation between building design processes and evaluation methods. Current green building rating tools evaluate building performance using checklists, but these contribute less to improving holistic [40] and people-centered design methods [41]. Design factors and evaluation indicators are different, both in terms of quantity and content. It is doubted that GBRTs can really reflect the sustainability performance of buildings. In addition, a "green" building in one country may be assessed as having a low score by another sustainability evaluation system [36]. It is understandable that different nations have their own social and environmental considerations that lead to differences in the weight distributions of evaluation indicators; however, these huge differences in indicators enhance the separation between GBRTs. Thus, it is necessary to establish a basic evaluation indicator system.

3. Research Strategy and Experimental Results

Text mining is a comprehensive technology that is closely related to natural language processing, pattern classification, relation graph extraction and machine learning, among others. The selection of the relevant technologies is based on research purposes. This study aimed to extract the main building design elements with natural aspects, so the chosen technique was Agglomerative Hierarchical Clustering (AHC). AHC is a popular unsupervised machine learning algorithm that can establish hierarchical trees and reveal relationships between data. This method is widely utilized within the text mining field because it requires less a priori knowledge. The research outline is shown in Figure 1 and included five key study phases: data collection and database establishment; data processing, i.e., extracting valuable words for text mining; word co-occurrence analysis; converting valuable words into vectors based on word2vec modeling; implementing the AHC algorithm.

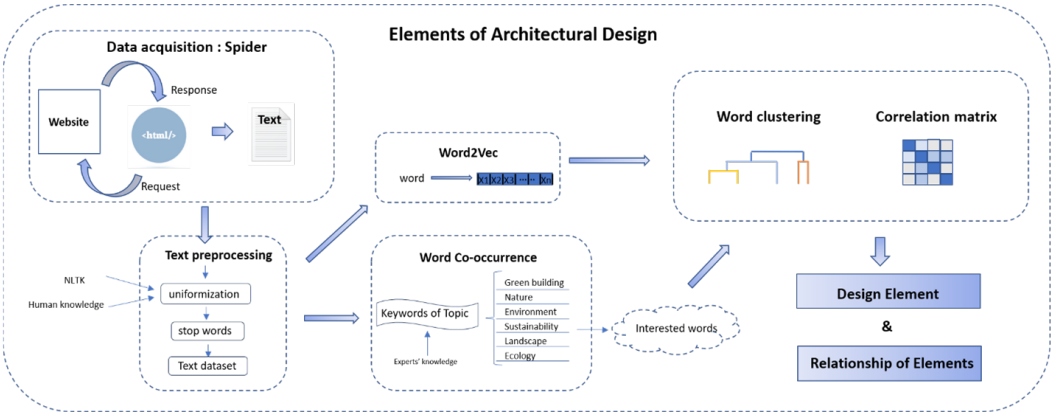


Figure 1. The research process.

3.1. Data Collection

For data collection, the appropriate data acquisition methods and channels had to be firstly identified. Disproportional data distributions could affect final analysis results. Existing research articles have tended to be more related to technical research and less related to architectural design. In addition, little research in articles has been applied to real construction projects and guidance for architectural design has been weak. Design descriptions from completed projects emphasize the key points and main design highlights that were considered during the architectural design process, which correlates more with the research topic of this study. The "Archdaily.com" website contains a large repository including the most project descriptions provided by architects all over the world. There are more than forty-eight thousand descriptions of real building projects that contain abundant subject matter knowledge to be discovered. Such knowledge indicates the major design factors of experienced architects in multitudinous scenarios, thus the repository can be used for text mining corpus.

In order to retrieve thousands of design descriptions from "Archdaily.com", this study implemented web crawler techniques to construct a text corpus for further analysis. Since the website does not use the robots exclusion protocol (which can indicate the availability of allowed content), the Selenium automated browser was employed as a source code extraction tool. Considering internet morality, this research tried to limit the influence of general users as much as possible. So, a crawler script based on Python was designed to extract website content during low traffic periods, i.e., midnight and the early hours of the morning. Moreover, the script could monitor connection parameters, such as latency, retransmission time, and IP address changes to predict server status, thereby decreasing the frequency of the script.

The next step was to analyze the HTML source code extracted from the "Archdaily.com" website. A general approach that is used in Python programming is the BeautifulSoup library, which can retrieve valuable data efficiently. In this study, the adoptive library could automatically recognize titles, architects, regions and project years within HTML source codes and then import them into a database for further processing. An attractive feature of the BeautifulSoup library is that it can reduce noise within websites, such as navigation, advertising and statements. Considering the strong relationships between instances extracted from the crawler script, this study applied a general MySQL database for data storage. This relational database can provide both reliability and scalability in the text mining procedure. There were four tables in the database: the text corpus, proper nouns, extracted features and clustering results. The full data collection procedures are illustrated in Figure 2.

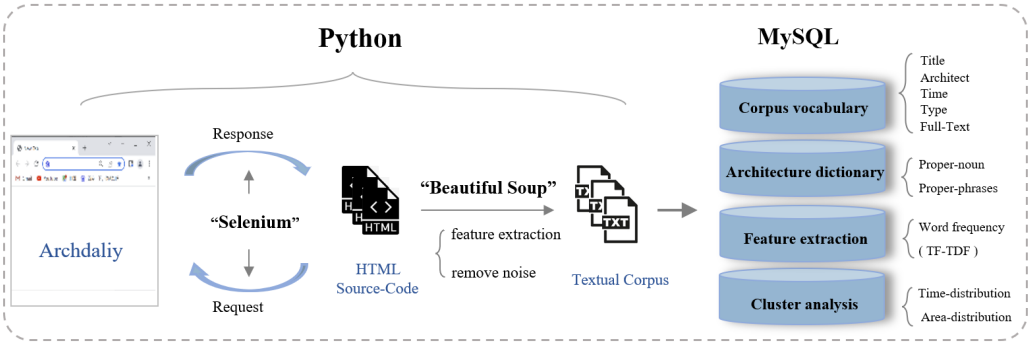


Figure 2. Data collection procedures.

In total, the test corpus contained 38,320 articles and 14,680,948 words. The numbers of articles and words in each category are summarized in Table 1. Among these types of buildings, the category of residential architecture had 15,734 articles containing 5,495,271 words and was the largest category. Religious architecture had 694 articles with 295,754 words and was the smallest category. The top four building types with the most articles were residential architecture, cultural architecture, commercial architecture and educational architecture.

Table 1. Text corpus information statistics.

Type	Articles	Words
Culture Architecture	4433	1,930,078
Commercial and Offices	4977	1,898,949
Education Architecture	3455	1,423,947
Healthcare Architecture	1034	388,799
Hospitality Architecture	3253	1,282,415
Industrial and Infrastructure	1406	556,029
Landscape and Urbanism	1390	594,095
Public Architecture article	1038	457,440
Religious Architecture	694	295,754
Residential Architecture	15,734	5,495,271
Sports Architecture	906	358,171

The quantity and descriptions of projects from the Archdaily website covered the majority of countries across the world. The distribution and number of project introductions can be seen in Figure 3. Each country had more than 100 project descriptions, except several countries in Africa. The United States had the largest number of projects, with more than 3000. There was no information about architectural projects in some countries and regions in western Asia, northern South America and Africa. The dataset had good validity in terms of quantity and coverage.

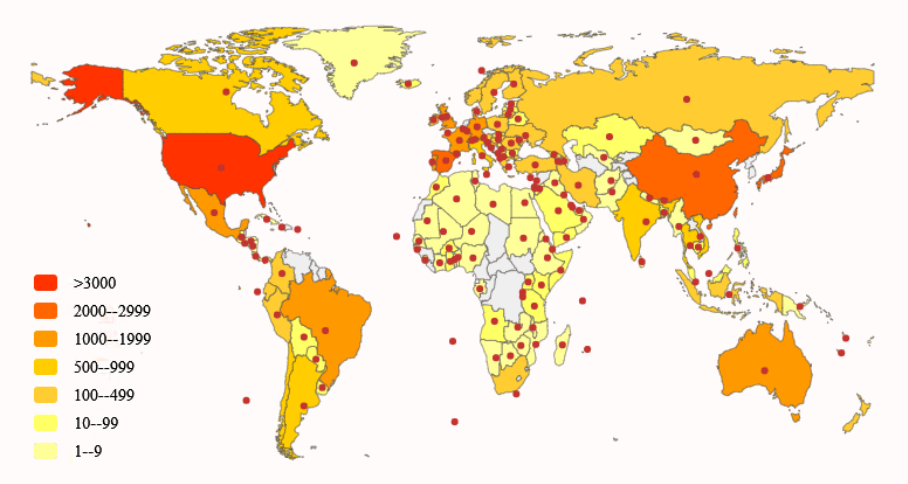


Figure 3. The distribution and number of project introductions.

The number of projects involving each building type published from 2007 to 2021 was counted and is shown in Figure 4. In terms of holistic trends, the number of projects increased year by year; however, the number started to decline from 2020. There was a significant decrease in 2021, partly because of the incomplete data acquisition (the data on Archdaily were crawled in October 2021). This drop was likely due to the reduction in new construction projects that was caused by the COVID-19 pandemic. Among the 11 architectural types, project descriptions of residential buildings were the most common each year, followed by commercial and office buildings, cultural buildings, educational buildings and hotel buildings.

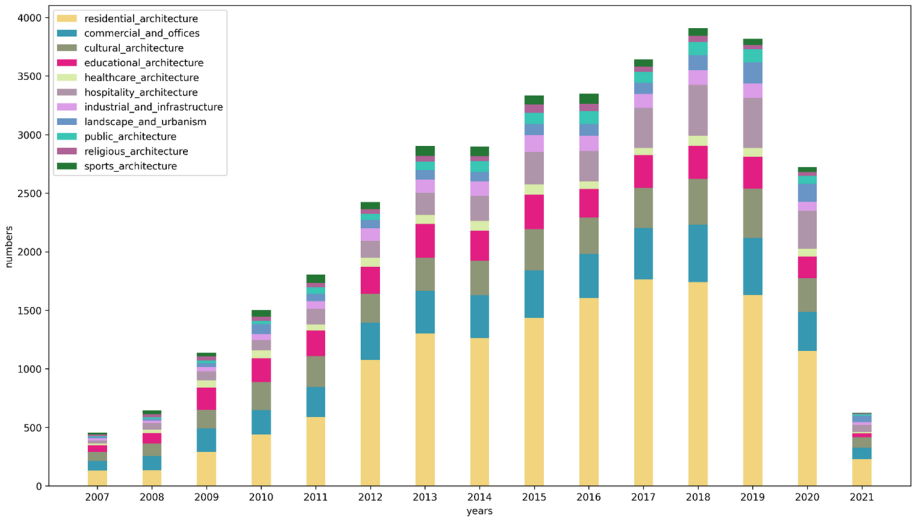


Figure 4. The number of projects involving each building type from 2007 to 2021.

3.2. Data Preprocessing

In the data mining process, data preprocessing is a critical procedure because collected data can include noise, dimension disasters and redundant instances. A very popular viewpoint is that data preprocessing should occupy more than half of the total workload in a data mining task. In this study, the data were unlikely to include noise since all of the text that was collected by the crawler techniques was generated by subject professionals. However, there were definite dimension disasters and redundant instances in the collected data, such as conjunctions, prepositions and numbers. These words were of extremely low value to the research purpose and increased the number of unnecessary calculations in the subsequent processing; therefore, they were deleted during the preprocessing. Another kind of word that had to be disposed of was common words. Common words were too general to offer valuable information from text mining and also increased the number of unnecessary computations. This study employed a Natural Language ToolKit module (NLTK), along with subject experience, to obtain stop words and remove redundant words from the text corpus. This operation could also decrease the dimensions of word vectors in the subsequent procedures.

In addition, word normalization is an essential preprocessing step in text mining. Since there are many different forms of single words expressing similar meanings within text corpora, they would be treated as different words without normalization. Therefore, this study unified lowercase, singular form and primary tense words within the text corpus. Moreover, a spell check and word root extraction were also implemented automatically by the NLTK. All of these preprocessing steps could decrease dimensions and redundant words with little information loss, thus increasing the value density of the text corpus for further mining processes.

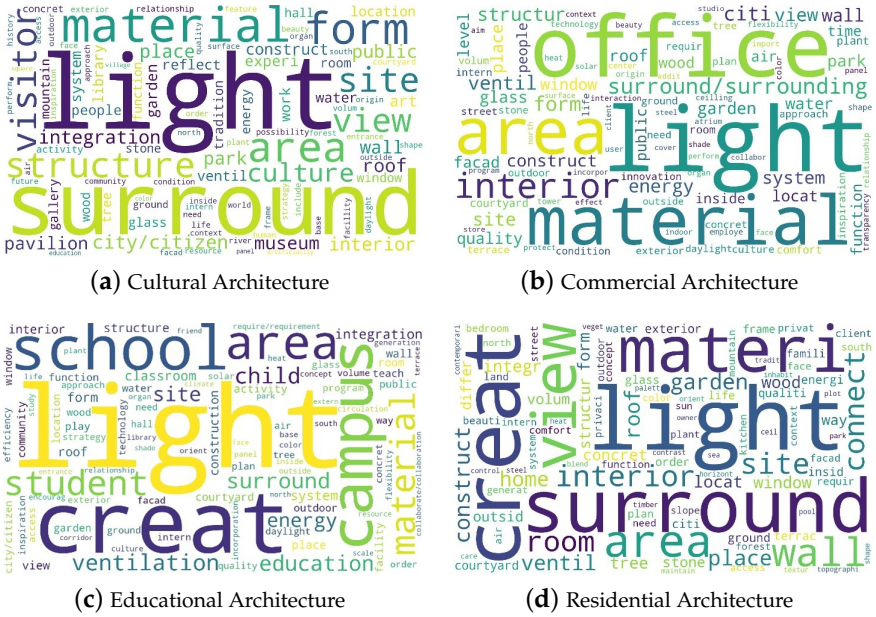
3.3. Word Co-Occurrence Analysis

In the linguistics domain, there are certain mutual relationships between the words in every sentence; thus, the degree of association between words can be captured by word co-occurrence relationship matrices. The more two words appear within the same natural statement, the closer their relationship. In this study, six words with high correlations to the research topic were chosen as keywords: nature, environment, green building, sustainability, landscape and ecology. The words were converted into their word roots for counting; thus, all of the different forms of each word were counted together. The words that co-occurred with the keywords were used for further analysis. Table 2 showed the results of word co-occurrence.

Table 2. The results of word co-occurrence.

Type	Number of Co-Occurrence Words with Nature Aspect	Average Number of Co-Occurrence Words with Nature Aspect per Article	The Average Percentage per Article
Culture architecture	260,417	58.75	13.49%
Commercial and offices	243,464	48.92	12.82%
Education architecture	220,044	63.69	15.45%
Healthcare architecture	37,978	58.63	15.59%
Hospitality architecture	197,206	60.62	15.38%
Industrial and infrastructure	82,814	58.90	14.89%
Landscape and urbanism	126,658	91.12	**21.32%**
Public architecture	63,474	61.15	13.88%
Religious architecture	34,318	49.45	**11.60%**
Residential architecture	802,824	51.02	14.61%
Sports architecture	51,709	57.07	14.44%

Word cloud maps are important data visualization tools that are widely applied in data mining. They emphasize the main points of articles by filtering out low-frequency and low-quality text information. The "keywords" that appear the most frequently in mass text content are visually highlighted in word cloud maps, i.e., the more frequently the "keywords" appear, the larger their font in the word cloud. As cultural architecture, commercial architecture, educational architecture and residential architecture had the largest numbers of article, these four building types were taken as the typical representatives for further research. For each of these four building types, the 100 most frequently used words were highlighted in word cloud maps, which are shown in Figure 5.



(**a**) Cultural Architecture (**b**) Commercial Architecture

(**c**) Educational Architecture (**d**) Residential Architecture

Figure 5. The word clouds for the four selected building types.

By comparing the word frequency lists for the four types of architecture, the most high-frequency words were found as light, landscape, material, surround, structure, facade, area, site, view and the main user of buildings. Most of the top 100 words could be categorized. For example, roof, wall and floor could be categorized as structure and plant, trees and gardens could be categorized as landscape. Thus, the design factors within the natural dimension could be summarized as area, light, landscape, structure, material, facade, site, water, wind and view. However, these classifications were only based on subject experience

and knowledge, so the results were too subjective. Therefore, it was necessary to further explore the relationships between the words.

3.4. Word Clustering

Considering text data are unstructured and difficult to calculate directly, the first step of clustering is the implementation of the word2vec tool, which is currently regarded as the most popular approach in text mining. Word2vec was developed by Google and works as a three-layer neural network to convert text into word vectors. Since words that are close to each other are semantically relative within a sentence, the algorithm can execute training in the neural network, thus establishing mapping from words to vectors. This procedure can retrieve relationships between words in text corpora and express them as vectors, which is convenient for machine calculations and processing. There are many methods for using word2vec; this study employed CBOW, which can predict central words based on context.

The CBOW model includes input layer, hidden layers, and output layer, which follows the neural network structure. Similarly, there are two weight matrices connecting the layers, named ω and φ. The predicted word is called center word, named c, while words in the context are called surrounding words named s. It should be noted that the number of surrounding words is also called window size. In this study, the total length of the text corpus is named V, window size is named w, and the dimension of word vector is N. The first procedure of CBOW modeling is to perform one-hot coding for surrounding works, so the s is converted to vectors. The next procedure is calculating the Hidden Layer vector (HL), and the equation is as follows:

$$HL = \frac{1}{n}\omega_1 \cdot \sum_{i=1}^{n} x_i \tag{1}$$

where: ω is input weight matrix, x is surround word vectors.

In this procedure, all the words in text corpus are mapping into weight matrix ω in the form of n-dimension vectors, hence ω is also called the matrix of center word vector. The third procedure of CBOW modeling is utilizing the matrix of surrounding word vector named φ to calculate output layer μ, which is expressed as follows:

$$\mu_j = \varphi_j \cdot HL \tag{2}$$

where: φ is output weight matrix, $varphi_j$ is the j column of matrix φ, HL is hidden layer vectors.

In order to output classification result, SoftMax function is employed to calculate posterior probability that indicates the probability of center word c given surrounding words s. The equation is expressed as follows:

$$p(c|s) = y_i = \frac{exp(\mu_i c)}{\sum_{j=1}^{V} exp(\mu_j s)} \tag{3}$$

where: μ_j is j column of matrix φ, y is the probability vector of output.

According to the Equation (3), the word with the highest probability is the predicted center word c given surrounding words s, and the CBOW model can establish the forward calculation. Then, the two weight matrices ω and φ can be trained by text corpus, and finalize the CBOW modeling. This training is performed by a very common error back propagation algorithm.

Since the parameter setup is important in the CBOW algorithm, the dimensions of word vectors and the widths of the surrounding word sequences can affect training accuracy. If the dimensions are set too small, it can cause redundant collisions, while dimensions that are too large can lead to huge computation consumption. In addition, the widths of the surrounding word sequences determine the dependencies between central words and define the context scope. This study set the dimensions to 300 and the width to 5. After 50 epochs of training, the proposed word2vec model achieved convergence. During the

experiment, 10 independent training runs could predict highly similar vectors for three common words in architectural design: window, glass and light. This proved that the word2vec model could convert words from the text corpus into vectors accurately and efficiency. The CBOW algorithm procedures are shown in Figure 6.

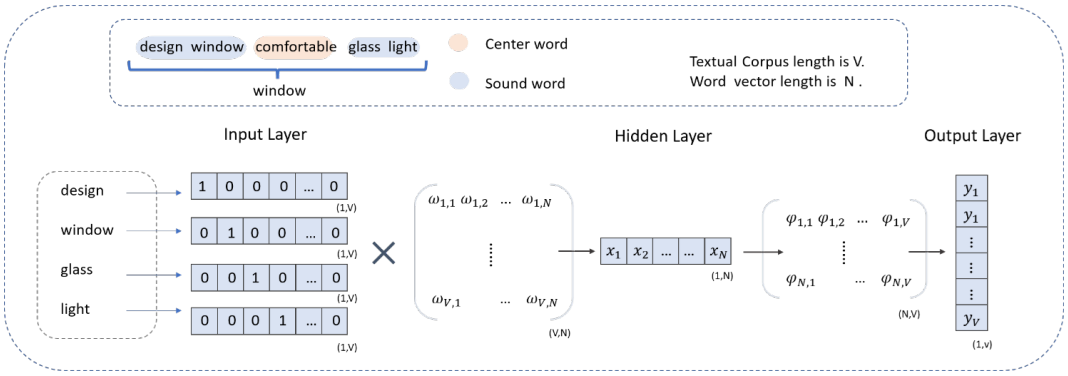


Figure 6. The word cloud of four building types.

For each building type, the number of co-occurring words was huge. For example, there were 27,462 words co-occurring with the keywords for residential architecture. Obviously, it was unnecessary to process all of these words during the text mining. This study conducted frequency filtering before the clustering analysis via term frequency–inverse document frequency (TF-IDF). The fundamental principle of TF-IDF is that a word is more important if it has a higher frequency in one type of sentence but a lower frequency in other types. TF-IDF can assess the real importance of words within text corpora and this study selected the top 60 words based on the TF-IDF ranking.

With word vectors, text clustering can be performed effectively using the following procedure. Clustering is an unsupervised learning method that is commonly applied in text mining. It can aggregate elements into multiple clusters based on similarity indicators, thus revealing potential relationships within datasets. In this study, since the words in the text corpus were converted into vectors, it was convenient to use the cosine distance as the similarity calculation, which was expressed as follows:

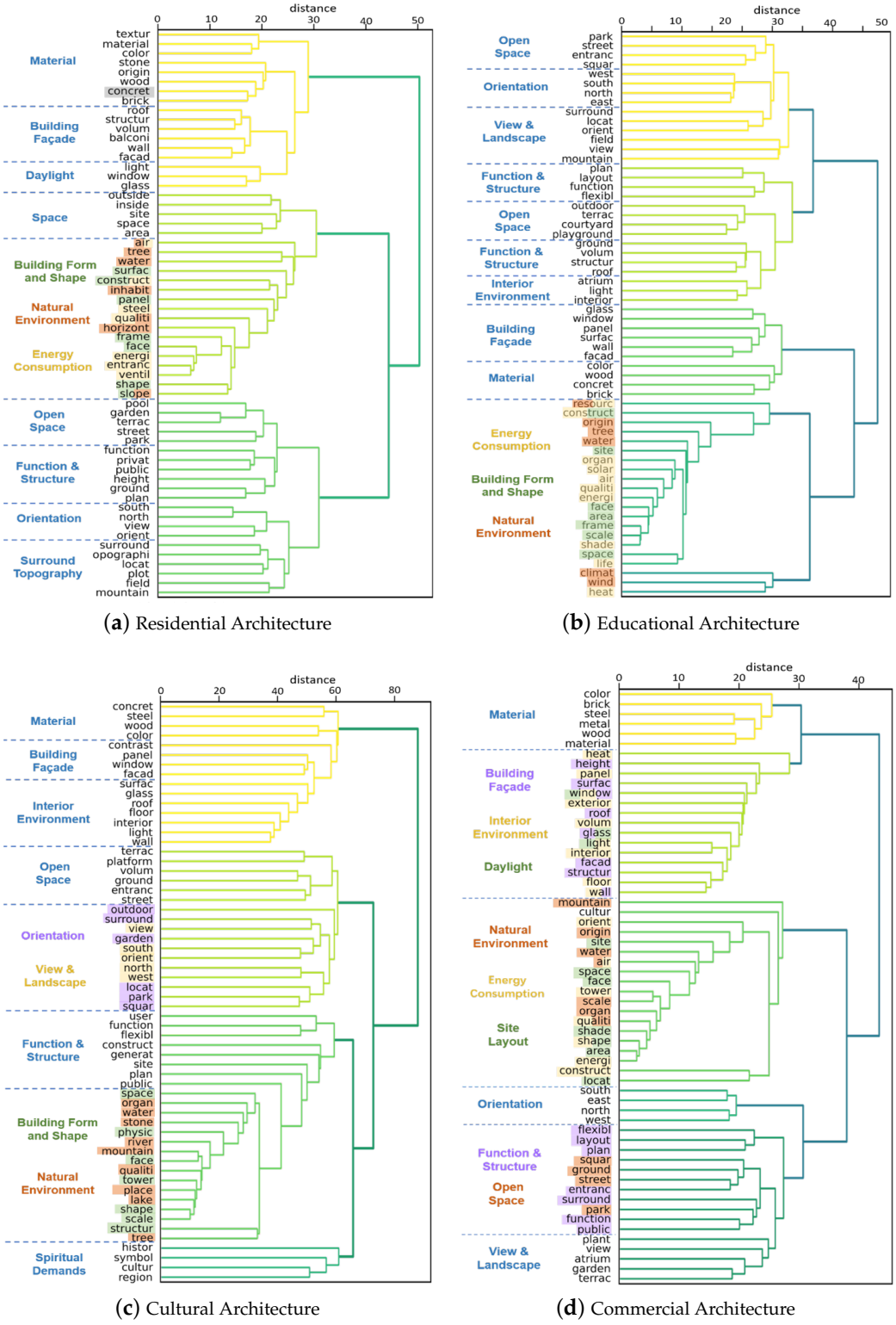
$$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}} \quad (4)$$

where:
X and Y are word vectors,
Theta is the angle of vectors.

Because the word vectors were generated by word2vec modeling and represented semantic relationships, the cosine distance of the word vectors could express their relevance within the text corpus. Therefore, this clustering method could aggregate highly correlated words together and identify major architectural design factors. Considering that a priori knowledge could not determine the number of clusters, agglomerative hierarchical clustering was applicable in this study and the algorithm was expressed as follows.

Compute the similarity matrix, if necessary.
Repeat
Merge the closest two clusters.
Refresh the similarity matrix to reflect the similarity among the new cluster and the initial clusters
Until only one cluster remains.

The distances between clusters were calculated using Ward link method, which is suitable for highly dimensional data. This algorithm could construct hierarchy trees to indicate similarity relationships between the words within the text corpus, which are illustrated in Figure 7.



(**a**) Residential Architecture (**b**) Educational Architecture

(**c**) Cultural Architecture (**d**) Commercial Architecture

Figure 7. Clustering result.

4. Results and Discussion

This section focuses on the analysis of the results, which identified architectural design elements and compared the differences and similarities between different building types, thereby providing guidance for future research.

4.1. The Consideration of Nature Was Far Below Expectations

According to the analysis results of text mining, as shown in Table 2, the proportion of co-occurring words related to natural aspects was highest in healthcare architecture articles (21.32%) and lowest in religious architecture and commercial architecture articles (11.60% and 12.82%, respectively). The proportions in the majority of the other articles were from 13% to 15%. Each article contained 50 to 60 co-occurring words that were related to natural aspects.

The importance of nature in building design processes was far below our expectations. Each article only contained 50 to 60 co-occurring words that were related to natural aspects, which meant that only five to eight sentences in each design description took natural considerations into account. It is normal and reasonable for architectural designs to compromise on many aspects, such as culture, economy and function. However, in the interest of sustainable development, the weights of natural considerations need to be increased.

4.2. Design Elements Clustering

According to the word clustering results, most of the words in the articles for each architecture type were clustered into different groups, as illustrated in Figure 7. Based on the means and similarities of the words in each group, architectural design elements could be summarized. In total, 11 different design elements were identified in both residential architecture and commercial architecture and 10 were identified in both educational architecture and cultural architecture. Overall, 15 design elements were identified by the word clustering: material (M), building form and shape (BF&S), building façade (BF), function and structure (B&S), site layout (SL), daylight (DL), energy consumption (EC), space (S), open space (OS), orientation (O), interior environment (IE), landscape and view (L&V), natural environment (NE), surround topography (ST) and spiritual demands (SD). The analysis results are shown in Table 3.

Among all of the identified design elements, only six were involved in all four selected building types: material (M), building façade (BF), function and structure (B&S), open space (OS), orientation (O) and natural environment (NE). Unlike the other three architecture types, cultural architecture took the design element of spiritual demands (SD) into consideration. Although the word roots histor, cultur and region were also identified in other architecture types by the word clustering, they were not clustered into the same word group. Residential architecture paid attention to the element of surround topography (ST), but it was not specifically mentioned in any other building types. The element of site layout (SL) was only clustered in commercial architecture articles.

For each type of building, just because design elements were not identified did not mean that they were not involved or considered in the design process. Many word roots did not belong to just one design element. For example, window and glass had strong connections with building façade (BF) and interior environment (IE) and they were also the main word roots in daylight (DL). Additionally, construct and entrance were both mentioned in function and structure (F&S) and energy consumption (EC). Thus, even elements that were not clustered still had a high probability of being considered in every building type. On the other hand, the clustered design elements appeared more frequently in textual expression and the words they contained had closer relationships within the text corpus. To some extent, this meant that these elements drew more attention or had higher priorities in design processes.

Table 3. Summary of design elements in the four building types.

	Residential Architecture	Education Architecture	Commercial and Offices	Culture Architecture
Material	brick, material, origin, wood, stone, color	color, wood, concret, birck	material, metal, wood, steel, brick, color	steel, wood, color, concret
Building Form and Shape	shape, frame, height, slope, face, surfac, panel	construct, site, face, area, frame, scale, space		space, face, tower, shape, scale, structur
Building Façade	roof, structur, volum, balconi, wall, façade	glass, window, panel, surfac, wall, façade	height, surfac, window, exterior, roof, volum, façade, structur, wall	contrast, window, façade, panel,
Function & Structure	function, privat, public, plan, ground	layout, function, plan, flexibl, ground, volum, structur, roof	flexibl, layout, sorround, plan, ground, entranc, function, public,	construct, site, flexibl, plan, public, function,
Site Layout			orient, site, space, face, tower, area, shape, locat	
Daylight	light, window, glass		light, window, glass	
Energy Consumption	construct, energy, qualiti, air, steel, ventil, entrance,	resourc, organ, climate, energi, wind, heat, solar, construct, north, shade, qualiti, air	orient, air, space, tower, scale, qualiti, shape, energy, construct	
Space	outside, inside, site, space, area			
Open Space	garden, terrac, street, pool, park	street, entranc, terrac, playground, squar, park, outdoor, courtyard	squar, street, public, ground, park	terrac, platform, ground, street
Orientation	south, north, view, orient,	south, north, west, east	south, north, east, west	south, north, east, west
Interior Environment		interior, light, atrium	heat, panel, window, glass, light, interior, floor, wall	surfac, roof, glass, wall light, interior, floor,
Landscape & View		mountain, locat, view, surround, field, orient	view, terrac, garden, plant, atrium	outdoor, surround, view, garden, locat, park, squar
Natural Environment	air, tree, water, inhabit, slope, qualiti, horizont	wind, origin, climate, water, resourc, tree,	mountain, origin, water, air, organ, quality	organ, water, stone, tree, river, mountain, quality, place, lake,
Surround Topography	surround, topographi, locat, field, mountain, plot			
Spiritual Demands				histor, symbol, cultur, region

4.3. Relationship between Elements

Heat maps are efficient for visualizing semantic relationships between words. The relationships between words of interest can be clearly observed using heat maps, which are also used to explore correlations between words and further explore semantic relationships and potential knowledge. The cosine similarity between words is used to represent the semantic distance between words. Each color square in a heat map represents the degree of cosine similarity between words, corresponding to the horizontal and vertical axes. Different correlation coefficients correspond to different color shades and the closer the semantic relationship, the darker the color of the squares between the words. In this study, all words were reranked into heat maps according to the design elements into which they were clustered; thus, the connections between design elements could be efficiently illustrated as well.

Figure 8 shows the correlations between the clustered words that were related to residential architecture. It is obvious from the figure that many design elements had connections with each other, as demonstrated by the red squares. Natural environment (NE) and energy consumption (EC) had no obvious correlations with the other elements, while apace (S) only had an obvious connection to building form and shape (BF&S). However, other design elements for residential architecture were closely related to each other. Building façade (BF) had high correlations with material (M), building form and shape (BF&S) and orientation (O), as well as strong connections with function and structure (F&S),

daylight (DL) and open space (OS). Function and structure (F&S) had close relationships with open space (OS) and orientation (O). Daylight (DL) was closely linked to material (M), open space (OS) and orientation (O). Open space (OS) had high correlations with orientation (O) and surround topography (ST). Surround topography (ST) also had a close relationship with orientation (O).

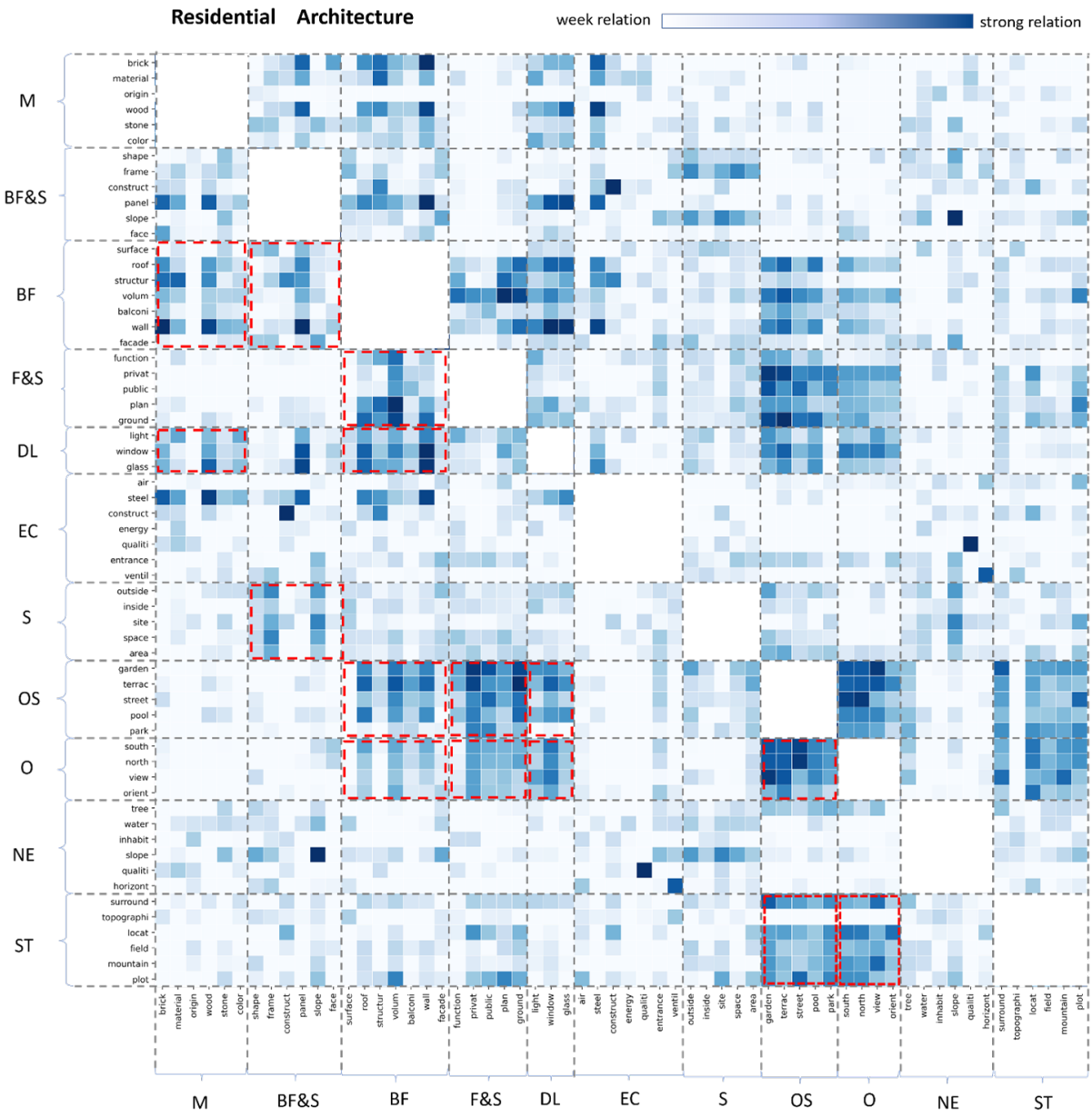


Figure 8. Heat map of clustered words related to residential architecture.

Figure 9 shows the correlations between the clustered words and design elements that were related to educational architecture. There were no very strong connections between the design elements for educational architecture, but the correlations were still obvious. To some extent, energy consumption (EC) had a connection with natural environment (NE) since some words that were clustered into these two elements correlated closely with each other. Building façade (BF) had correlations with material (M), function and structure (F&S), open space (OS) and interior environment (IE). Function and structure (F&S) also had relationships with open space (OS), orientation (O) and interior environment (IE). Open space (OS) was closely linked to orientation (O), interior environment (IE) and landscape

and view (L&V). Orientation (O) was also obviously connected to interior environment (IE) and landscape and view (L&V).

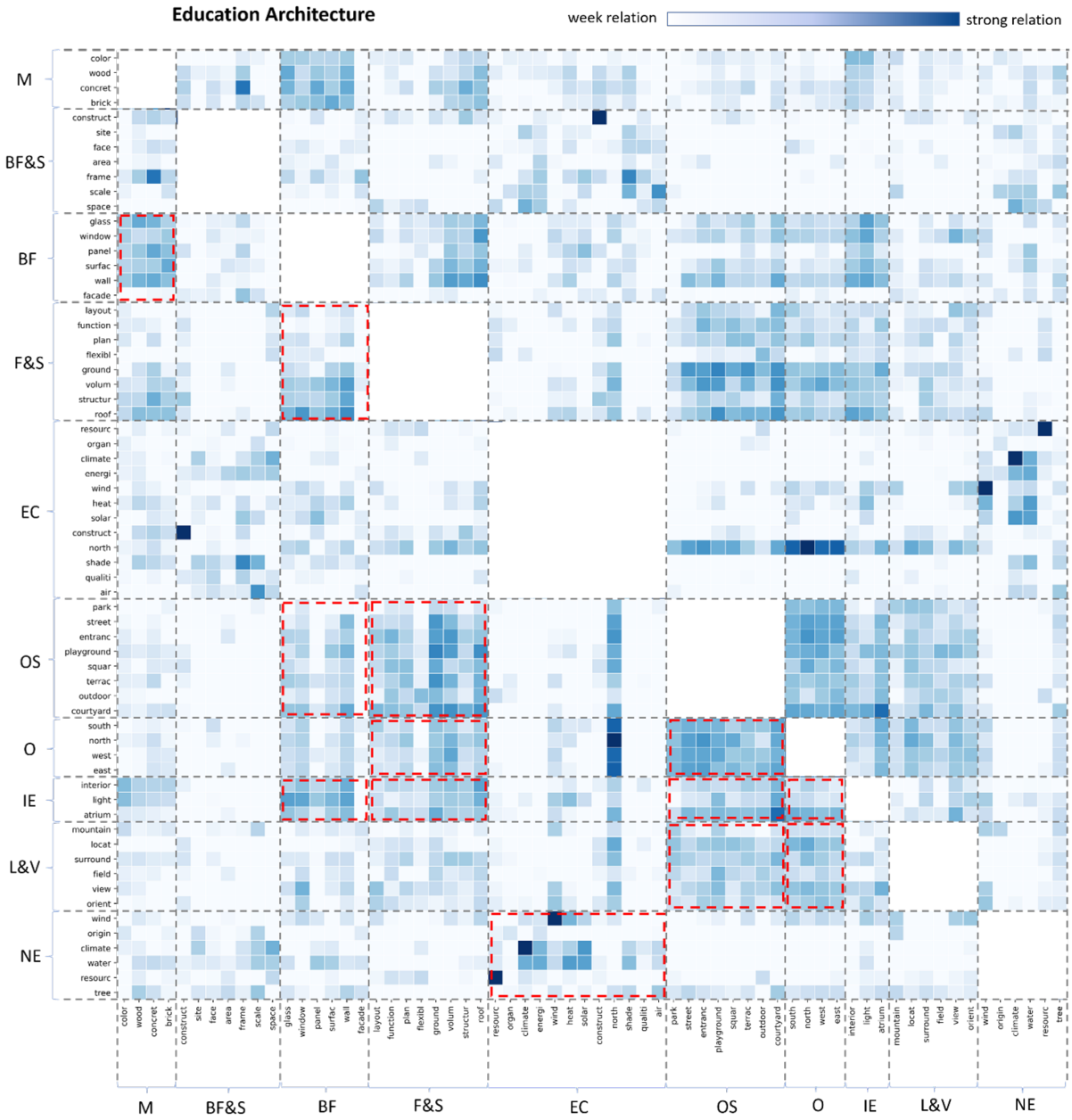


Figure 9. Heat map of clustered words related to education architecture.

For commercial architecture, the correlations between the clustered words and design elements were quite different. Site layout (SL), energy consumption (EC) and natural environment (NE) had no obvious connections with the other elements. However, the other eight design elements had stronger and more obvious connections with each other those for the other building types, as visualized in Figure 10. These eight design elements had close relationships with at least three elements. Interior environment (IE) and building façade (BF) both had relationships with seven elements and the correlation between these two elements was stronger than predicted.

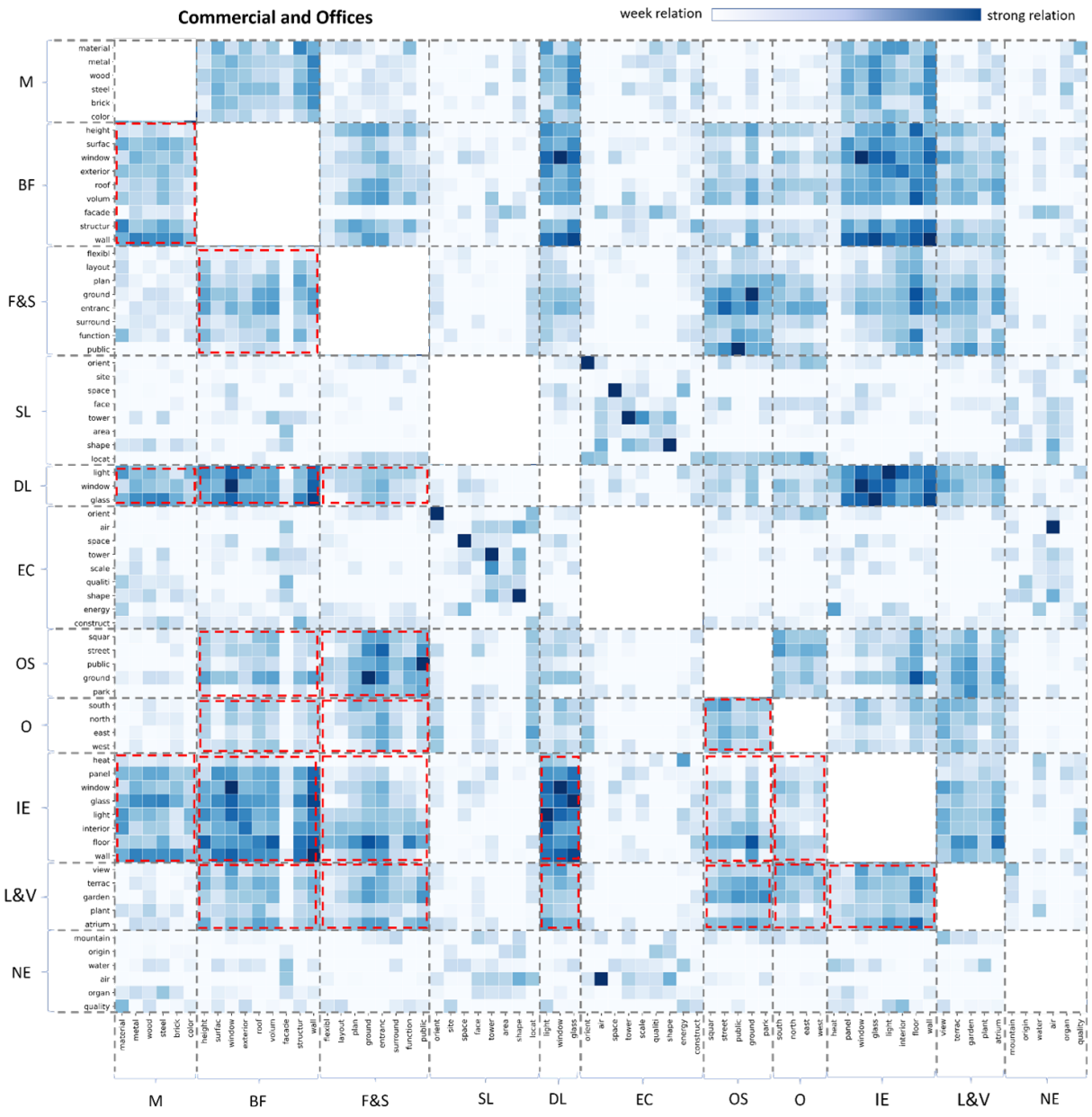


Figure 10. Heat map of clustered words related to commercial and offices.

Compared to the heat maps of the other types of architecture, the color in Figure 11 is quite light. This meant that the correlations between the clustered words and design elements were weaker than those for the other building types. Only six correlations could be recognized between four elements. This result could have been because cultural architecture is more influenced by region and local culture. Compared to the other building types, it also has wider variations in design and text expression.

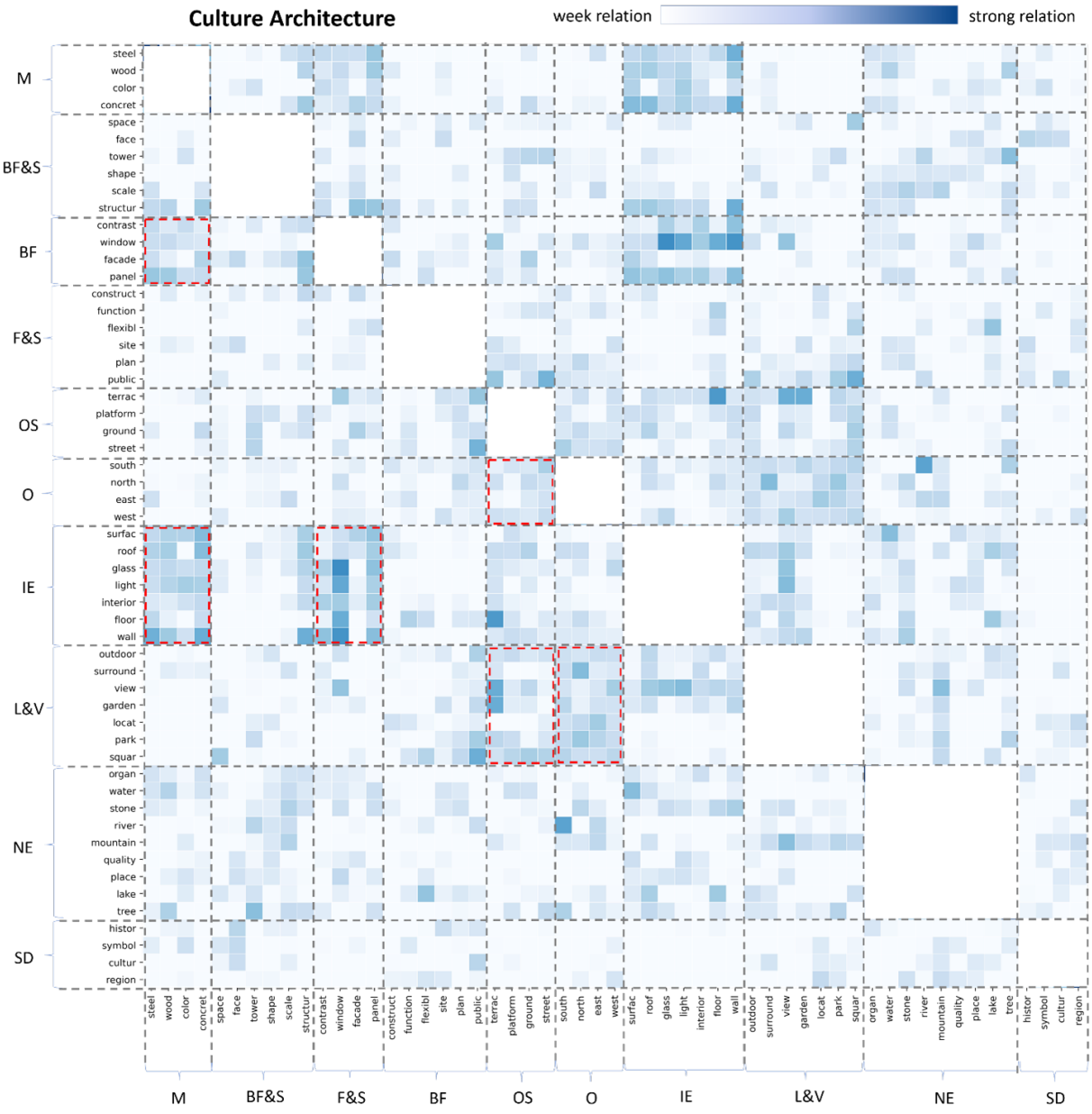


Figure 11. Heat map of clustered words related to culture architecture.

4.4. Potential Evaluation Indicators

To select the evaluation elements, two perspectives were mainly considered. The first was mutual relationships between buildings and the natural environment, focusing on the protection and utilization of natural resources. From this perspective, the effectiveness of design schemes was mainly considered to resist or improve harsh natural environments, including how to use appropriate natural resources to improve interior environments and whether construction projects would have adverse effects on the surrounding environment. The second was the impact of architectural designs on landscapes, focusing on the evaluation of the harmony between a project and its surrounding landscape and considering the local characteristics and natural topography, as it is necessary to protect natural landscapes, city skylines and the public's view of landscapes. Combined with the word clustering results and heat maps, the 15 identified design elements were separated into eight evaluation features and each element had two to four indicators, which are shown as follows:

- Ecological Value: Green area; The variety of species; Preservation & Creation of natural environment.

- Land Utilization: Site selection; Appropriate site layout; The changes of natural topography.
- Light: Enough indoor daylight; Light pollution reduction; Daylight obstruction.
- Wind: Ventilation air & quality; Outdoor wind environment.
- Water: Water landscape variation; Water saving; Rainwater management.
- Architectural Space: Appropriate Building Form and Shape; Height of building; Function & structure.
- View & Landscape: View out; Enough open space; Appropriate orientation; View interference.
- Material and Resource: Renewable energy utilization; Recycling Material selection.

From the text mining results, the same types of buildings in different areas had obvious similarities in terms of design elements. Otherwise, there would have been no obvious word clustering results, if the text expression had been discrete. In addition, different building types also had a high similarity in terms of design elements. This indicated that it could be feasible to adopt the same evaluation factors in building evaluation systems for different regions. The weights of the evaluation factors just need to be adjusted according to the specific local conditions and different building types.

5. Conclusions and Limitations

By applying text mining technology, this paper analyzed 38,320 professional architectural design articles from the "Archdaily.com" website. The study used web crawling and text mining technologies to explore the major factors related to natural aspects within architecture. Using Python script, descriptions from the "Archdaily.com" website were collected efficiency and accuracy. These building descriptions included subject knowledge from designers and the clustering analysis discovered potential relationships between design factors and nature. In total, 15 design elements were identified by the word clustering: material, building form and shape, building façade, function and structure, site layout, daylight, energy consumption, space, open space, orientation, interior environment, landscape and view, natural environment, surround topography and spiritual demands. Different building types in different areas had obvious similarities in terms of design elements, so it could be feasible to adopt the same evaluation factors in building evaluation systems for different regions. This paper aimed to clarify the main building design factors with a consideration of natural aspects, thereby offering possible evaluation indicators and improving the environmental performance of buildings throughout the design processes. Through the process of text mining, the correlations between words and design elements were discovered, which could be used to stimulate the optimization of architectural design.

Even though the results showed reasonable and valuable outputs, there were some limitations in this research, and further study is required. First, all of the architectural design descriptions were collected from the "Archidaily.com" website, which is a single source. Even though it is the most popular website that provides mass data about architectural design, the data still contain bias to some extent. Including more data sources and constructing comprehensive text corpora in future works could increase the confidence of the research findings. Second, since this research focused on descriptions written in English, the findings from the text mining were only applicable in certain regions. Therefore, the methodology for this research could be improved to include multi-language text to produce complementary results. Furthermore, according to the text mining results, some specialized words were identified incorrectly, so creating terminology dictionaries based on subject knowledge could benefit text mining.

Author Contributions: Conceptualization, D.Z.; methodology, Y.L.; software, Y.L.; validation, B.P.; investigation, B.P. and X.W.; resources, X.W. and W.G.; writing—original draft, D.Z.; writing—review & editing, S.M. and W.G.; supervision, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2014 Revision: Highlights*; (ST/ESA/SER.A/366); United Nations, Department of Economic and Social Affairs, Population Division: New York, NY, USA, 2015.
2. Li, G.; Ma, X.; Song, Y. Greening Building Efficiency and Influencing Factors of Transportation Infrastructure in China: Based on Three-Stage Super-Efficiency SBM-DEA and Tobit Models. *Buildings* **2022**, *12*, 623. [CrossRef]
3. Duarte, R.; Sanchez-Choliz, J.; Sarasa, C. Consumer-side actions in a low-carbon economy: A dynamic CGE analysis for Spain. *Energy Policy* **2018**, *118*, 199–210. [CrossRef]
4. United Nations Environment Programme. *2021 Global Status Report for Buildings and Construction: Towards a Zero-emission, Efficient and Resilient Buildings and Construction Sector*; United Nations Environment Programme: Nairobi, Kenya, 2021 .
5. Kayıhan, K.S. Examination of Biophilia Phenomenon in the Context of Sustainable Architecture. In Proceedings of the 3rd International Sustainable Buildings Symposium (ISBS 2017), Dubai, United Arab Emirates, 15–17 March 2017; Fırat, S., Kinuthia, J., Abu-Tair, A., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 80–101.
6. Hanafi, M.; Naguib, M. Bio-regenerative rating technique: A critical review. *Ecosyst. Sustain. Dev.* **2013**, *175*, 233–246. [CrossRef]
7. Rahman, N. A Taxonomy of Data Mining Problems. *Int. J. Bus. Anal.* **2018**, *5*, 73–86. [CrossRef]
8. Czibula, G.; Czibula, I.G.; Miholca, D.L.; Crivei, L.M. A novel concurrent relational association rule mining approach. *Expert Syst. Appl.* **2019**, *125*, 142–156. [CrossRef]
9. Tandel, S.S.; Jamadar, A.; Dudugu, S. A Survey on Text mining techniques. In Proceedings of the International Conference on Advanced Computing & Communication Systems (ICACCS)-2019, Coimbatore, India, 15–16 March 2019.
10. Jung, H.; Lee, B.G. Research Trends in Text Mining: Semantic Network and Main Path Analysis of Selected Journals. *Expert Syst. Appl.* **2020**, *162*, 113851. [CrossRef]
11. Mohsen, A.M.; Idrees, A.M.; Hassan, H.A. Emotion Analysis for Opinion Mining From Text: A Comparative Study. *Int. J. e-Collab.* **2019**, *15*, 38–58. [CrossRef]
12. Anderson, C.B.; Craiglow, H.A. Text mining in business libraries. *J. Bus. Financ. Librariansh.* **2017**, *22*, 149–165. [CrossRef]
13. Ferreira-Mello, R.; Andre, M.; Pinheiro, A.; Costa, E.; Romero, C. Text mining in education. *Wiley Interdiscip. Rev.-Data Min. Knowl. Discov.* **2019**, *9*, e1332. [CrossRef]
14. Drury, B.M.; Roche, M. A survey of the applications of text mining for agriculture. *Comput. Electron. Agric.* **2019**, *163*, 104864. [CrossRef]
15. Cooper, H.M. Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowl. Technol. Policy* **1988**, *1*, 104–126. [CrossRef]
16. Choi, H.S.; Lee, W.S.; Sohn, S.Y. Analyzing research trends in personal information privacy using topic modeling. *Comput. Secur.* **2017**, *67*, 244–253. [CrossRef]
17. Dhar, A.; Mukherjee, H.; Dash, N.S.; Roy, K. Text categorization: Past and present. *Artif. Intell. Rev.* **2021**, *54*, 1–48. [CrossRef]
18. Min, K.; Yoon, M.; Furuya, K. A Comparison of a Smart City's Trends in Urban Planning before and after 2016 through Keyword Network Analysis. *Sustainability* **2019**, *11*, 3155. [CrossRef]
19. Kotlerman, L.; Dagan, I.; Kurland, O. Clustering small-sized collections of short texts. *Information Retrieval* **2017**, *21*, 273–306. [CrossRef]
20. Patel, S.M.; Dabhi, V.K.; Prajapati, H.B. Extractive Based Automatic Text Summarization. *J. Comput.* **2017**, *12*, 550. [CrossRef]
21. Lee, J.; Yi, J.S. Predicting Project's Uncertainty Risk in the Bidding Process by Integrating Unstructured Text Data and Structured Numerical Data Using Text Mining. *Appl. Sci.* **2017**, *7*, 1141. [CrossRef]
22. Huang, H.L.; Hong, S.H.; Tsai, Y.C. Approaches to text mining for analyzing treatment plan of quit smoking with free-text medical records: A PRISMA-compliant meta-analysis. *Medicine* **2020**, *99*, e20999. [CrossRef]
23. Levy Mendelovich, S.; Barbash, Y.; Budnik, I.; Erez, D.; Somech, R.; Soffer, S.; Furth, S.; Klang, E. Pediatric literature trends: High-level analysis using text-mining. *Pediatr. Res.* **2021**, *90*. [CrossRef]
24. Zhao, X.; Zuo, J.; Guangdong, W.; Huang, C. A bibliometric review of green building research 2000–2016. *Archit. Sci. Rev.* **2018**, *62*, 1–15. [CrossRef]
25. Saka, A.; Chan, D.D. A Scientometric Review and Metasynthesis of Building Information Modelling (BIM) Research in Africa. *Buildings* **2019**, *9*, 85. [CrossRef]
26. Rocha, P.; Rodrigues, R. Bibliometric Review of Improvements in Building Maintenance. *J. Qual. Maint. Eng.* **2017**, *23*, 437–456. [CrossRef]
27. Na, X.U.; Ling, M.A.; Liu, Q.; Wang, L.; Deng, Y. An improved text mining approach to extract safety risk factors from construction accident reports. *Saf. Sci.* **2021**, *138*, 105216.
28. Abdelrahman, M.M.; Zhan, S.; Miller, C.; Chong, A. Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. *Energy Build.* **2021**, *242*, 110885. [CrossRef]

29. Shen, L.; Yan, H.; Fan, H.; Wu, Y.; Zhang, Y. An integrated system of text mining technique and case-based reasoning (TM-CBR) for supporting green building design. *Build. Environ.* **2017**, *124*, 388–401. [CrossRef]
30. Ding, Z.; Rongsheng, L.; Li, Z.; Fan, C. A Thematic Network-Based Methodology for the Research Trend Identification in Building Energy Management. *Energies* **2020**, *13*, 4621. [CrossRef]
31. Istiadji, A.; Hardiman, G.; Satwiko, P. What is the sustainable method enough for our built environment? *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *213*, 012016. [CrossRef]
32. Ryan, C.; Browning, W.; Clancy, J.; Andrews, S.; Kallianpurkar, N. Biophilic design patterns: Emerging nature-based parameters for health and well-being in the built environment. *Archnet-IJAR* **2014**, *8*, 62–76. [CrossRef]
33. Gillis, K.; Gatersleben, B. A Review of Psychological Literature on the Health and Wellbeing Benefits of Biophilic Design. *Buildings* **2015**, *5*, 948–963. [CrossRef]
34. Pedersen Zari, M.; Connolly, P.; Southcombe, M. *Ecologies Design: Transforming Architecture, Landscape and Urbanism*; Routledge: Oxfordshire, UK, 2020.
35. Wijesooriya, N.; Brambilla, A. Bridging biophilic design and environmentally sustainable design: A critical review. *J. Clean. Prod.* **2021**, *283*, 124591. [CrossRef]
36. Mattoni, B.; Guattari, C.; Evangelisti, L.; Bisegna, F.; Gori, P.; Asdrubali, F. Critical review and methodological approach to evaluate the differences among international green building rating tools. *Renew. Sustain. Energy Rev.* **2018**, *82*, 950–960. [CrossRef]
37. Zhang, Y.; Wang, J.; Hu, F. Comparison of evaluation standards for green building in China, Britain, United States. *Renew. Sustain. Energy Rev.* **2016**, *68*, 262–271. [CrossRef]
38. Awadh, O. Sustainability and green building rating systems: LEED, BREEAM, GSAS and Estidama critical analysis. *J. Build. Eng.* **2017**, *11*, 25–29. [CrossRef]
39. Wilde, P. Ten questions concerning building performance analysis. *Build. Environ.* **2019**, *153*, 110–117. [CrossRef]
40. Gou, Z.; Xie, X. Evolving Green Building: Triple Bottom Line or Regenerative Design? *J. Clean. Prod.* **2016**, *30*, 600–607. [CrossRef]
41. Xue, F.; Lau, S.; Gou, Z.; Song, Y.; Jiang, B. Incorporating biophilia into green building rating tools for promoting health and wellbeing. *Environ. Impact Assess. Rev.* **2019**, *76*, 98–112. [CrossRef]