ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0006

Land Engineering Fund Project Text Mining Data Collection and Preprocessing Method

Ya Hu[1,2,3,4], Tingyu Zhang[1,2,3,4]

[1] Shaanxi Provincial Land Engineering Construction Group Co., Ltd, Xi'an 710075, China

[2] Institute of Land Engineering and Technology, Shaanxi Provincial Land Engineering Construction Group Co., Ltd, Xi'an 710075, China

[3] Key Laboratory of Degraded and Unused Land Consolidation Engineering, the Ministry of natural Resources, Xi'an 710075, China

[4] Shaanxi Provincial Land Consolidation Engineering Technology Research Center, Xi'an 710075, China

Abstract

The current land engineering fund projects are immature in terms of deep topic mining, and have not yet used text mining and machine learning algorithms to study the direction and evolution of the land engineering field. Based on the analysis of the text data of the National Natural Science Foundation of The data collection and text cleaning, text segmentation and other text data preprocessing methods with crawler technology as the core have laid a research foundation for the text mining of the National Natural Science Foundation of China for land engineering.

Keywords

Text Mining; Crawling; Preprocessing; Land Engineering.

1. Introduction

With the development of the era of big data, data mining technology is applied in many fields. Text mining is a kind of data mining that uses information retrieval, information extraction, computational language, natural language processing, data mining and other technologies to discover previously unknown, implicit and useful information based on the cross-industry data mining standard process (CRISP-DM).method [1]. Compared with traditional mining, text mining requires additional data selection processing procedures and complex feature extraction steps. Data mining deals with structured data, while text mining deals with semi-structured or unstructured data [2].

Text mining technology can analyze text information more deeply, reflect the development trajectory of technological innovation in time, and provide new methods for the identification and prediction of scientific research project paths. Over the years, various websites have accumulated a large amount of scientific and technological innovation information, and these unstructured text data hide the development trend and direction of scientific and technological innovation. In the face of these huge and complex text data, relying solely on manual reading to obtain valuable information, the workload can be imagined, the operation is time-consuming and labor-intensive, and the traditional methods of scientific and technological innovation data collection are too limited, making quantitative methods in social and economic sciences difficult. No, it is difficult to accurately find the objective laws of scientific and technological innovation [3]. Existing scholars have carried out qualitative research methods from the system construction of land scientific and technological innovation [4], the management system of land scientific and technological innovation [5], the development strategy of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0006

land scientific and technological innovation [6], and the scientific and technological innovation of land consolidation [7]. The analysis has important theoretical and practical significance for the policy development of land science and technology innovation and the construction of a collaborative platform, but few studies have explored the situation of the National Natural Science Foundation of China for land engineering from a quantitative perspective. In recent years, with the development of text mining technology, more and more domestic scholars have begun to explore text mining research [8-12], providing important methods and technical support for analyzing massive text data.

Based on the analysis of the text mining technology of the National Natural Science Foundation of China for land engineering, this paper outlines the data acquisition and text data preprocessing methods, which lays a research foundation for the text mining of the National Natural Science Foundation of China for land engineering.

2. Land Engineering Fund Project Text Mining Data Collection Method

Web page collection, also known as web crawler, is one of the core parts of search engines. In fact, it uses programs to automatically browse and collect web pages. Web crawling is a technology to ensure the source of web data in the web crawler system. At present, with the rapid development of web scraping technology, it has become a necessary tool and an important method for people to obtain and collect information on the Internet.

This crawler is based on the python programming language, and uses libraries such as requests and Beautiful Soup to improve development efficiency of the crawler. The program mainly includes two files, index.py and detail.py, which correspond to the function of crawling and retrieving content and the function of crawling the details page respectively. The python file index.py includes post_ url and get_ id methods. The function of the post_ url method is to obtain the search content list according to the search keyword, such as searching for "soil" or "land".

A total of 1144 items of content have been retrieved in the figure, and each retrieval content has a corresponding retrieval code. Through this retrieval code, the details page of the retrieval content can be further obtained. The code structure of the post_ url method is relatively simple. It mainly uses the post method of the requests library to obtain the page request result of http://output.nsfc.gov.cn/ and returns it. Before this, you need to set the request header parameters in detail, otherwise the other server do not respond.

The get_ id method is to obtain the retrieval code of the details page by retrieving the content list. The retrieval code of each retrieval content is included in the 'results Data' field, which needs to be traversed to obtain. Finally, write the obtained retrieval code into the test.txt file to prepare for the next step to obtain the details page information. The detail.py file includes post_ url, txt_ pro, excel_ init, and handle_ execl methods, respectively corresponding to requesting details page information, processing txt function, initializing excel function, and writing excel function. The txt_pro file is mainly to recombine the content of the test.Txt file is generated after the execution of index.py is completed into web page parameters that are convenient for request, and return it to post_url for use. The post_ url method constructs the URL used by the web page request according to the parameters parsed by txt_ pro, and returns the request result. The excel_ init method is used to initialize the excel sheet in preparation for writing the result of the save request. The handle_ excel method is used to save the request result to excel. Using the above method, you can quickly and easily obtain low-demand results.

3. Construction of Land Engineering Science and Technology Text Data Preprocessing Method

Preprocessing mainly removes punctuation marks, numbers, stop words and other texts that are not meaningful for subsequent analysis, reducing the impact on the word segmentation results. The preprocessing process is completed in the R software environment. First, set the environment, load the required R packages such as pipeline setting packages, data processing packages, stuttering word

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0006

segmentation, etc., and then perform text cleaning and repetition for first-level punctuation removal and second-level content removal. The value is deleted, next, the Jieba tool is used to segment the Chinese text, mark the part of speech, and remove the stop words.

3.1 Text Cleaning

Although it is possible to break through some defense lines when collecting data, due to the heavy collection workload and frequent operations, it is inevitable that problems such as repeated returns, errors, and invalid data will occur during the collection process. In order to ensure the accuracy of follow-up work, Firstly, the collected text data is manually cleaned, that is, some data that have low utilization rate in the process of text analysis or will cause large errors in the results are deleted, including deletion and deduplication, and elimination of symbols. Deduplication is to remove repeated data in the collected text corpus, such as system default or repeated collection, compression and word removal is to delete words that are repeatedly expressed in sentences, deletion of short sentences is to remove some meaningless or meaningless words. Clear sentences, on the basis of the first two steps, need to further delete short sentences.

In order to make the mining results more accurate, it is necessary to extract keywords from the title and content of scientific and technological innovation text data, including stop word punctuation, numbers and part-of-speech filtering, etc., to provide structured normative data for the next mining.

3.2 Text Segmentation

The word segmentation method without a dictionary is a statistical method based on word frequency. This method counts the frequency of occurrence of two adjacent words in the document as a word. When the threshold is set, it is indexed as a word, but this kind of word segmentation also has defects, and some words with high frequency but no actual meaning is extracted, which reduces the accuracy of text word segmentation. At present, there are many open source word segmentation tools. The most popular is the stuttering word segmentation algorithm, which has an interface in Python and R. The jieba word segmentation is a method based on a statistical dictionary. The input text is preliminarily constructed by constructing a prefix dictionary. Divide, and then construct a directed acyclic graph according to all the possible positions of the segmentation, and calculate the maximum probability path through the dynamic programming algorithm. The HMM-related model is trained, and then the Viterbi algorithm is used to analyze and solve to obtain an optimal state sequence. Finally, the word segmentation result is output according to the state sequence.

The principle of the stuttering word segmentation algorithm is mainly: ① Implement efficient word graph scanning based on the prefix dictionary, and generate a directed acyclic graph (DAG) composed of all possible word formations of Chinese characters in a sentence. ② Use dynamic programming to find the maximum probability path, find The maximum segmentation combination based on word frequency is obtained. ③ For unregistered words, the HMM model based on the ability of Chinese characters to form words is adopted, and the Viterbi algorithm is used.

Use the R software to call the jieba package for word segmentation. Compared with the Rwordseg package in the R software, the jieba package does not need to configure the Java environment and can be downloaded and called directly. In addition, the jieba package supports four word segmentation modes such as the maximum probability method and the index model. , also has functions such as part-of-speech tagging, Chinese name recognition, etc., and its word segmentation speed is fast. Just import the text data into R, and then load the worker function in the jieba package to perform preliminary word segmentation on the text data.

3.3 Remove Stop Words

After the text is divided into words, there will be some virtual words or transition words, which will increase the dimension of the text after the whole word segmentation, so it is necessary to eliminate such words on the basis of word segmentation. Or words that are no longer used are called stop words (Stop Words). Stop words are roughly divided into the following two categories: one is some words

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0006

that are widely used and frequently used, such as "is", "just", "a", etc. , these words appear in almost every document: the other category is meaningless words, mainly including modal particles "ah", "ya", etc., conjunctions such as "also", "but", etc., prepositions, adverbs, etc. , only when placed in the context of a specific sentence has a certain effect. Deleting such words can reduce the dimension of the text and improve the accuracy of text analysis. In addition, non-stop words that are not frequently used in the text can also be filtered. Such words are often of no value for the representation of text features. Generally, the processing of such words can be filtered according to the length of the word or the frequency of occurrence.

4. Conclusion

The discipline of land engineering is developing very rapidly, and there will be more and more problems in the application of scientific research projects. Scientific research management and scientific research workers must do differentiated analysis, and fully understand the nature of land engineering through data mining, text mining, etc. It is of great value to the development of the discipline of land engineering to understand the research rules of funded projects and master the research direction of land engineering. This paper summarizes the data collection and text cleaning, text segmentation and other text data preprocessing methods with crawler technology as the core, which lays a research foundation for the text mining of the National Natural Science Foundation of China for land engineering.

Acknowledgments

This work was Supported by the Scientific Research Item of Shaanxi Provincial Land Engineering Construction Group (DJNY-2021-28).

References

[1] Fytilakos Ioannis. Text mining in fisheries scientific literature: A term coding approach[J]. Ecological Informatics, 2021,61.

[2] Yao Tianfang, Cheng Xiwen, Xu Feiyu, et al. Review of Text Opinion Mining [J]. Journal of Chinese Information, 2008, 22(3):71-80.

[3] Gianpiero Bianchi, Renato Bruni, Cinzia Daraio, et al. Exploring the Potentialities of Automatic Extraction of University Webometric Information[J].Journal of Data and Information Science,2020,5(04): 43-55.

[4] Du Yamin, Gao Shichang, Miao Limei. Thoughts on improving the management system of land science and technology innovation [J]. China Land, 2018, 394(11): 29-31.

[5] Hao Han, Yang Jinyu. Research on urban agricultural land use problems and countermeasures based on rural revitalization strategy [J]. China Agricultural Resources and Zoning, 2020, 41(09): 80-84.

[6] Chen Cheng, Zhong Jixiang, Zhang Danfeng, et al. Research on scientific and technological innovation of land remediation from the perspective of ecological civilization-based on the registration and award-winning achievements in the fields related to land remediation of the former Ministry of Land and Resources [J]. China Land Science, 2018, 32 (4): 82-88.

[7] Li Shanghao, Chao Lemen. A review of the application of text mining in Chinese information analysis [J]. Information Science, 2016, 34(08): 153-159.

[8] Zhang Zhenhua, Xu Baiming. Construction and Application of Business Competitive Intelligence Analysis Model Based on Online Review Text Mining [J]. Information Science, 2019, 37(2):149-153.

[9] Nie Xiuping, Xie Nengfu, Hao Xinning, et al. Analysis of hot topics in foreign agricultural scientific research projects based on text mining [J]. Jiangxi Agricultural Journal, 2018, 30(7):102-106.

[10] Wang Ge, Zhang Anlu, Yang Fan, et al. Research on the cooperation network and hotspot evolution of land science and technology innovation--Taking the Land and Resources Science and Technology Award as an example [J]. China Land Science, 2019, 33(6):104-112.

[11] Wang Lingyan, Fang Shu, Ji Peipei. Research on the Technical Framework of Using Patent Documents to Identify Emerging Technology Topics [J]. Library and Information Service, 2011, 55(18):74-23.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202206_8(6).0006

[12] Li Mei. Research on some key technologies in text mining [D]. Northwest A&F University, 2020.