Visualization Analysis of Tourism Data in Tai'an City with Python

Xue Tian(✉)

School of Information Science and Technology, Taishan University, Tai'an, Shangdong, China
tianx1122@163.com

Abstract. The functions of tourism websites are becoming more and more perfect, which provide a channel for tourists to freely share their travel experiences. These messages contain rich and valuable information. If we find a way to analyze these messages, we can identify problems in the development of tourism and find the direction for improvement. In this paper, we use the rich third-party library provided by python to extract, segment and display the original text on qunar.com, and finally obtain relevant data on the development of tourism in Tai'an. Through analysis, it is found that tourists have a higher awareness of Mount Tai than that of Tai'an. The brand effect of Tai'an as a tourist city is limited, and the overall competitive advantage is not obvious. Therefore, an integrated urban tourism strategy should be established to expand the popularity of Tai'an's tourism industry.

Keywords: Tai'an · tourist industry · web crawler

1 Introduction

Shandong is a big tourist province, famous for its culture, seaside and other landscapes. Tai'an, known as the Mount Tai, plays an important role in Shandong's tourism industry. In the "2022 Shandong Tourism Statistics Summary" released by Shandong Provincial Department of Culture and Tourism, many indicators of 2021 tourism across the province were collected. In 2021, Tai'an received 640.99 million domestic tourists, with tourism revenue of 65.5-billion-yuan, accounting for 8.8% and 7.86% of the provincial data respectively. Both indicators are in the fifth place, behind Qingdao, Jinan, Yantai and Weifang. According to the data of 2017–2020, except for the data of 2017, Tai'an surpassed Weifang by a slight margin, the rest of the years ranked the same as 2021. However, the statistical data can only show the objective number of tourists, and cannot understand the real feelings of tourists. Obtaining the real feelings and needs of tourists in proper way is more conducive to discovering existing problems and finding ideas for improvement.

After the development of the network, it has provided a platform for information interaction for all walks of life, including the tourism industry [1, 2]. Various tourism websites provide tourists with a variety of ways to share tourism information, such as ctrip.com, qunar.com, and mafengwo.cn [3]. In addition to providing tourists with services such as travel mode selection and scenic spot introduction, it also provides

J. Yen et al. (Eds.): ICBIS 2023, AHCS 14, pp. 1252–1257, 2023.
https://doi.org/10.2991/978-94-6463-198-2_129

tourists with channels to upload their tourism experience [4]. Many tourists publish travel strategies on the website and provide many opinions and suggestions for other tourists. If effective information can be extracted from these websites, it will better provide more effective countermeasures for the construction of tourist cities [5].

2 Data Sources and Methods

2.1 Data Sources

Qunar.com is one of the largest tourism search engines in China, providing relatively rich tourism resources and information. This study will take the tourism strategies on qunar.com as data sources, and select the tourism notes in Shandong and Tai'an regions for research. After confirming the areas, the number of strategies is 1000 and 359 respectively. We will collect the title, time, content and other data of these strategies.

2.2 Web Crawler Technology

Web crawler technology is to write programs to crawl web page data in the network according to certain rules and policies formulated. It can simulate the process of browser accessing the web page to extract the desired target data, and is currently widely used in data analysis, artificial intelligence and other fields. At present, the mainstream web crawlers are usually written in python language, and python plays a huge role through third-party libraries provided by developers. The third-party libraries commonly used in designing web crawlers include Requests, Beautiful Soup, Pandas, etc. Among them, the Requests library can simulate human access to the server to request web pages to achieve crawling. The Beautiful Soup library can parse the source code of the crawled web pages, The Pandas library can provide the data structure and methods used in processing crawling, making the data processing more convenient.

2.3 Data Analysis and Processing

2.3.1 Word Segmentation Technology

Capturing keywords from a text requires word segmentation technology, which can break sentences into words by specific algorithms, and then process them. It is the core technology of natural language processing. The process seems simple, but it is more complicated in practice because of the ambiguity of word segmentation. The commonly used word segmentation methods include regular word segmentation, statistical word segmentation and mixed word segmentation. Jieba word segmentation is one of the most commonly used word segmentation tools at present. It combines the advantages of regular word segmentation and statistical word segmentation and its workflow is shown in the Fig. 1. It can be implemented on many platforms and by programming languages, and is provided in Python as a third-party library. In this study, Jieba tool is selected to implement word segmentation processing.

In the process of word segmentation, in order to prevent some auxiliary words, prepositions or other words that have no meaning for the study from interfering with the

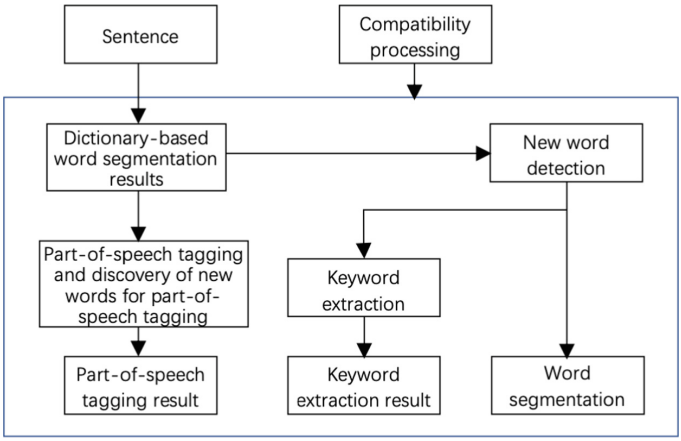

Fig. 1. Workflow of Jieba word segmentation

subsequent research results, these words need to be removed. Therefore, a stop-word list can be established. At present, many institutions have provided some general stoplists, such as Baidu stoplists, Harbin Institute of Technology stoplists, etc., which can be used directly or supplemented on this basis.

2.3.2 Chart Display and Word Cloud Display

In order to display the word frequency and hot words of the words after word segmentation more vividly and obtain more accurate conclusions, charts and word clouds can be created. The Python third-party libraries, Matplotlib and Wordcloud, provide these two functions respectively. Matplotlib library can draw a variety of static, dynamic and interactive charts, such as scatter chart, pie chart, histogram, 3D graphics, etc. Wordcloud library can generate a word cloud map, visually display the keywords with high frequency in the text, and get the main idea of the text at a glance.

3 Process and Results

3.1 Analysis of Tourism Data in Shandong Province

We crawled the tourism notes within Shandong Province, and obtained 837 records after deleting duplicate and invalid data. We obtained the frequency of 16 cities in Shandong province appearing in the notes. Use the Matplotlib library to generate pie charts, as shown in the Fig. 2. Tai'an ranked sixth, with the top six accounting for 94%.

WordCloud library is used to perform word cloud display on the word segmentation results. The hot words and word frequency are shown in the Table 1. It can be seen from the table that the popular words include Qingdao, Rizhao, Jinan, seaside, Mount Tai, etc.

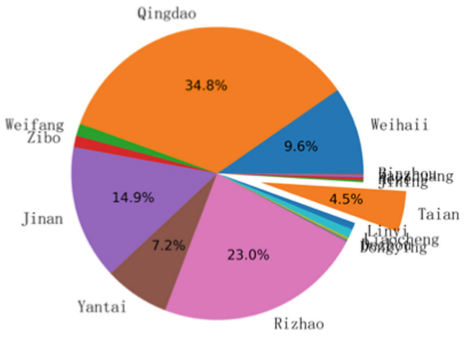

Fig. 2. Pie chart of heat statistics of 16 cities in Shandong Province

Table 1. The hot words and word frequency in Shandong Province

Word	Word frequency	Word	Word frequency
Qingdao	1517	Hotel	614
People	1149	Road	613
Rizhao	1002	Seafood	546
Eat	886	City	507
Good	827	Seaside	506
fisherman's family	676	Ticket	462
Mount Tai	654	Prepare	438
Jinan	647	Accommodation	435

3.2 Analysis of Tourism Data in Tai'an City

We crawled the tourism notes within Tai'an City, and obtained 312 records after deleting duplicate and invalid data. WordCloud library is used to perform word cloud display on the word segmentation results. The hot words and word frequency are shown in the Table 2. It can be seen from the table that the popular and important words include Mount Tai, Jinan, ticket, reservation, ticket price, Tai'an, Nantian gate, scenic spot, Spouting Spring, etc.

4 Discussion

The data analysis results show the following problems.

According to the results of statistics on the number of occurrences of 16 cities in Shandong Province, Tai'an City ranks sixth, which is one place away from the fifth place in "Shandong Tourism Statistics Summary". No matter from the government statistics or the data of network crawling, Qingdao's position is unshakable, and it is far away from the second place [6].

Table 2. The hot words and word frequency in Tai'an

Word	Word frequency	Word	Word frequency
Mount Tai	3917	Tai'an	828
Ticket	1964	Sunrise	697
People	1328	Nantian gate	691
Jinan	1187	Scenic spot	689
Ticket price	1043	Spouting Spring	684
0538	909	Mountain climbing	665
Eat	850	Taishan District	641
Climb	865	Daming Lake	623

From the word cloud generated by the results of crawling in Shandong Province, "Qingdao" has the highest popularity, while the popularity of the word "Tai'an" is not obvious. However, the frequency of the word "Mount Tai" is relatively high, and the number of occurrences of the word " Mount Tai" is 654, which is significantly compared with the 197 occurrences of "Tai'an". It can be seen that the popularity of Mount Tai is far above that of Tai'an, and the urban brand effect of Tai'an is relatively weak. Although Qingdao is famous for the seaside, its perception of tourists is still very high, which means that Qingdao is not limited by a single scenic spot and has built its own city brand. Therefore, we can refer to the study of Qingdao's tourism city planning [7].

According to the results of web crawling in Tai'an City, the hot spots are Mount Tai, relevant scenic spots in Mount Tai and Dai Temple. But there are also other natural or cultural scenic spots in Tai'an City, such as Taohuayu, Yuquan Temple, Fengshan Festival, Tianyi Lake, Sun Tribe, West Lake of Mount Tai, which are mentioned less in the notes, further verifying that the tourism industry in Tai'an City is relatively weak and has not formed enough scale. Tourists stay in Tai'an for a short time, so that the consumption of catering and accommodation is restricted, which further affects the overall tourism income [8].

5 Conclusion

Compared with the coastal cities and the provincial capital Jinan, the overall competitiveness of tourism in Tai'an is not obvious. However, it has its own characteristics and has certain potential [9]. How to give full play to the advantages of Mount Tai and improve the attraction of other scenic spots is an important direction for the further development of tourism in Tai'an. This study is mainly carried out from the perspective of tourists, which has certain reference value, but the amount of data is limited and there are certain limitations, which will be further improved and deepened in future research [10].

Acknowledgement. Tai'an Science and Technology Innovation Development Project (2020ZC329).

References

1. Yueqin Guo, Han Li. An analysis of the hot spots and frontier visualization of interpretation teaching at home and abroad. Shandong Foreign Language Teaching (2022)42–53. https://doi.org/10.16482/j.sdwy37-1026.2022-06-005.
2. Junping Wang, Wensheng Zhang. Construction and inferential analysis of the cognitive map for big data. Science China (2020)988–1002. https://doi.org/10.1360/SSI-2019-0273.
3. Na Wang, Huanqing Dong. Research on the construction of online tourism website information ontology with user participation. Journal of Modern Information (2021) 64–75. https://doi.org/10.3969/j.issn.1008-0821.2021.06.006.
4. Le Zhang, Yifang Sun. Visual analysis of Yuncheng tourism data based on Python. Computer Era (2022)85–88. doi:https://doi.org/10.16644/j.cnki.cn33-1094/tp.2022.10.021.
5. Qiang Zhao. Data analysis and visualization of tourism websites based on Python crawler. Electronic Design Engineering (2022)152–155. https://doi.org/10.14022/j.issn1674-6236.2022.16.033.
6. Yan Zhang, Botao Ding. The Role and Position of Shanghai in the Integration of Tourism in the Yangtze River Delta. Yangtze River Delta Research (2022)33–38. https://doi.org/10.3969/j.issn.1674-7739.2021.04.006.
7. Song Huang, Yanlin Li, Pingjuan Dai. Evaluation on tourism competitiveness of smart tourism cities. Acta Geographica Sinica (2017)242–255. https://doi.org/10.11821/dlxb201702005.
8. Yifan Lu, Ying Qu. Identification and dimension construction of local attachment elements of tourists in Jiangnan ancient towns. Areal Research and Development (2021)94–98. https://doi.org/10.3969/j.issn.1003-2363.2021.03.016.
9. Jianjian Shen, Ning Li. Research on the promotion path of tourism in Tai'an under the background of We-media. National circulation economy (2020)96–98.
10. Xu Qin, Yufeng Wu. Visualization analysis of hot spots of eco-tourism research in our country. Journal of Smart Agriculture (2023)29–32. https://doi.org/10.20028/j.zhnydk.2023.04.008.