Design and Implementation of Rental Information Analysis System

Yingying Chen[1,a], Taizhi Lv[1,b]

[1]College of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing, 211170, China.

[a]760846936@qq.com, [b]lvtaizhi@163.com

Abstract

In this paper, Scrapy crawler is used to obtain the information of house source. A visualization system of house source data of shell is designed based on several factors that are most concerned by users, such as city, rent, supporting facilities, area and city. The system is developed based on Python language, uses the Scrapy distributed crawler framework to obtain data, and pandas to realize data cleaning and analysis. Finally, pyecharts and flask framework are integrated to realize data visualization.

Keywords

Scrapy, Data visualization, Pandas, Data analysis.

1. Introduction

With the rapid development of big data and cloud computing technology, more and more people come into contact with big data. In the face of massive data in the era of explosive data, how to use data to directly reflect the problem has become one of the active research topics.

At present, most areas have entered the process of urbanization. With the increase of urban population, housing has become a problem that has to be considered. Therefore, how to find suitable housing has become an unavoidable livelihood problem. According to China daily.com, in 2017, the floating population in China was 250 million, accounting for nearly one fifth of the total population, of which 67.3% chose to rent. According to the Ministry of education, the number of fresh graduates will reach 8.74 million in 2020. They are very sensitive to the change of their first choice from renting a house to renting a house for three years, so they need a lot of experience to enter the society. According to the survey data, 85% of the freshmen regard rent as the primary consideration when renting a house.

According to the release of China social science network, since March 2020, housing prices in Shenzhen, Dongguan and other cities have risen, while prices in other core cities have not changed significantly. From a national perspective, the market has basically recovered, and the housing turnover in core cities has generally recovered. With the rapid development of the Internet, people can get the house source data from all over the country and even the whole world without leaving home, which provides convenience for house buyers to see and select houses. The Internet solves the problem of unequal information between tenants and landlords, and many platforms also support online viewing. From the perspective of supply, China's housing stock is nearly 100 million units, of which 25% are idle, and a considerable number of idle housing resources in many first tier cities have not entered the rental market. However, in this context, there are few big data analysis on housing.

2. Data acquisition

Scrapy is an application framework for crawling web data or extracting structured data based on python. It can be widely used in data processing, data mining and monitoring and other fields [1].This system uses scrapy as a data acquisition tool mainly because it can collect web data with multi-layer structure, which is more suitable for shell house searching.

Step 1: Make sure the name of the crawler is rent, set the range of pages allowed to crawl, and make sure that the initial URL is the city list page.

Step 2: Set random cookies to simulate user login.

Step 3: Analyze the city list structure.

Step 4: Locate the city page according to the obtained city page URL, and obtain the URL addresses of new house and rental house respectively. In this paper, the crawler of new house and rent house is written into two spiders. Because there is no difference between the two except the page number, taking the crawling of the Rental page as an example, the URL address of the rental house source page is obtained by copying the XPath of the rental house according to the source code, and the URL address is assigned to requests through the callback function callback.

Step 5: Enter the *review element* of the listing page, and analyze the web page information that the system needs to obtain, such as the city and the source address field. The city field is located at the top left of the page. You can get the text directly according to the XPath expression, and then use the regular expression to extract the city.

Step 6: In the source address field, you need to find the div of each source information, put all divs into a list, and then traverse the corresponding source address information under all divs and the link of the next level detail page, and pass in the item.

Step 7: Turn the other page. In order to crawl all the housing information of the city, the necessary page turning operation is needed. The next page list can be obtained in the format of current page URL address /PG/ #contentlist.

Step 8: Select the required data crawling on the detail page, review the page elements to obtain the XPath expression, use regular expression to extract the data needed by the system, and finally save all the values into item.

3. Data processing

Pandas is a tool based on numpy, which is created to solve data analysis tasks [3]. This paper uses pandas to complete the data exploration, cleaning and analysis.

3.1 Data exploring

Before data exploration, the database must be read through pymysql library. Pymysql is a pure Python implementation of MySQL client library, supporting Python 3 compatible, used to replace MySQL dB [2]. Explore the data of each field in the database, judge the data type, analyze the fields that need to be cleaned, such as rent, area, city, etc., and eliminate the useless data such as empty data and undetermined price in these fields.

Finally, determine the data to be analyzed:

(1) These data can reflect the development level of a city.

(2) The average price of housing resources in each city, the economic level of a city can be judged by the house price. The more developed the city is, the higher the house price is.

(3) The comparison of the highest price and the lowest price of housing supply in each city can often show the gap between the rich and the poor in this city.

3.2 Data cleaning

Analysis of rental data collection results shows that there are no useless characters in the rent field and the data is complete, while the data required for the area field is a specific number, and what is crawled down is a number plus m^2 or there is no data for the moment. All the fields without data need to be eliminated, and the useless character "M^2" that affects the total area of calculation should be removed.

It is necessary to eliminate the data whose price is to be determined, instead of changing it to 0, because these data are OK, but leaving the calculated average value will affect the accuracy of the final result. Then the data format is "average price x yuan / average", which can be used the str.extract method to extract the number of the string, and then found that the resulting array is object. Convert the data type to float type to calculate the mean value.

3.3 Data analyzing

This system classifies the data according to the city. It first calculates the number of houses in the city, takes the field city, and then uses the value_counts method which can get all the values of the city column and the corresponding frequency (that is, the number of houses). In order to facilitate the subsequent visual calls, it is directly written into a CSV file and placed in the root directory of the project.

The mean method can be used to get the average value of the house price. The max method takes the maximum value, the Min method takes the minimum value, and then uses the round method to get two decimal places.

4. Data Visualization

Echarts is an open source data visualization tool developed by Baidu. It is developed by pure JavaScript. Pyecharts is a combination of Python and Echarts [4]. Flask is a lightweight web framework, written in Python language, with strong customization [5]. This paper integrates Flask and Pyecharts framework to provide visual application.

The first step is to implement the page layout. A parent div is defined, and the other 7 Divs are all subclasses. The position of the parent div is set to relative positioning, and the layout of the subclass div is absolute positioning and fixed position. The space between divs is calculated to determine the location of div by the position to the left and top of the page.

The second step is to implement the timing switching display operation. After the div can be displayed in the center, you need to write a function to switch the display of div1 periodically, so that all charts can be displayed in turn. Write an if condition to judge the currently displayed Div1 attribute. If it is displayed, modify the display of Div1 on the same page to none to display Div2. Otherwise, display div1 and hide Div2. The function of switching the display property is set to 25 seconds.

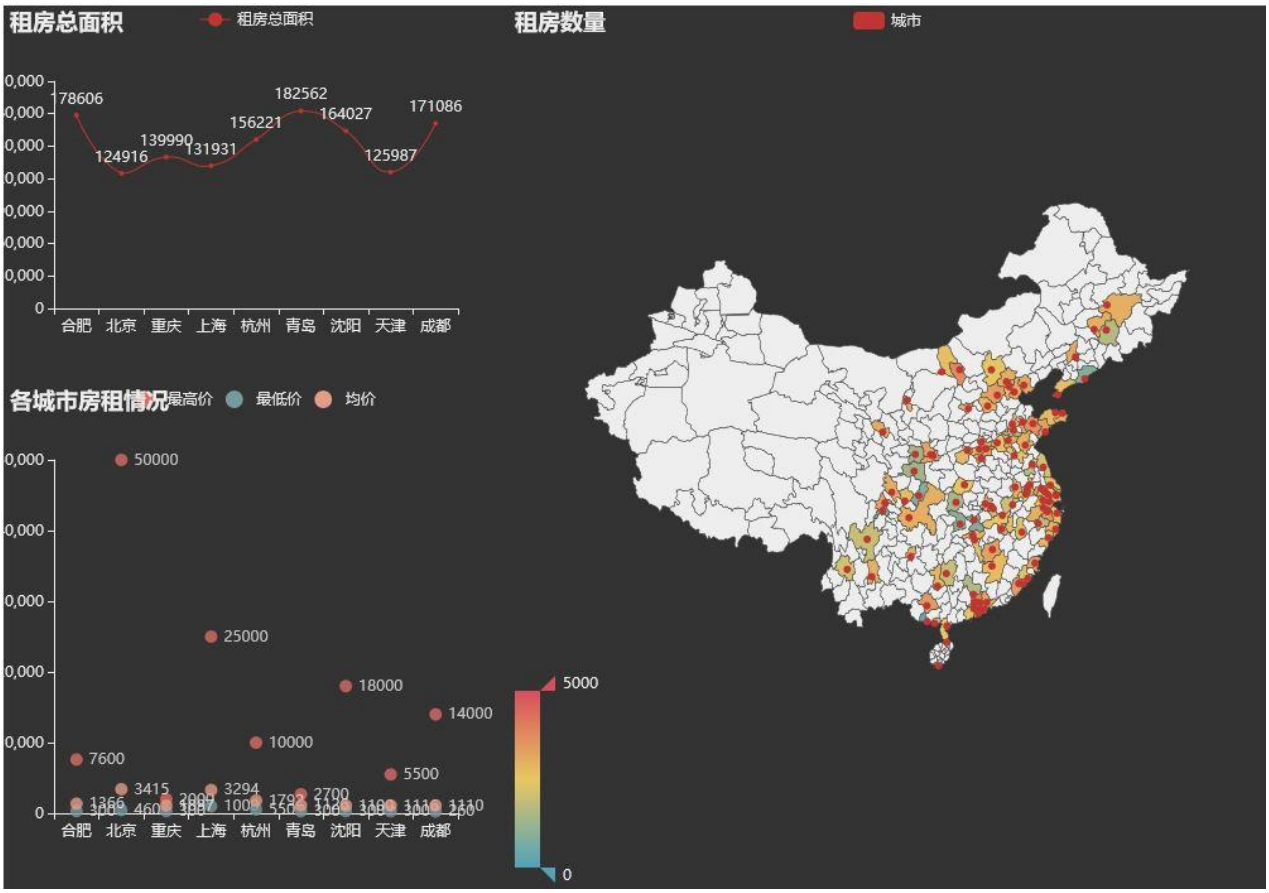

Figure 1. Data visualization of the rental information

5. Conclusion

Based on the shell house searching platform, this paper mainly completes the functions of shell house searching data collection, data cleaning, data analysis and data visualization. The useful data are statistically and visually analyzed, and valuable information is analyzed and mined out to make full use of the potential value of big data. In order to improve the significance of the system research, we select the problems that users are concerned about for data analysis.

(1) The data of the website is collected and saved to the database. The use of scrapy framework makes data collection much easier. Setting crawling delay avoids the problem of IP being temporarily blocked because of crawling too fast. Setting user agent to deal with the anti-crawling mechanism of website. In order to transfer the storage of data, the database is written into the database, which also facilitates the regularization and differentiation of fields.

(2) The function of cleaning redundant data and analyzing useful data is realized. The data analysis is realized by using regular expression and Pandas data processing tools to eliminate the unnecessary statistical fields and meaningless values. In addition to calculating the total rental area of the city, the number of housing resources, the housing price of each city, we can understand the gap between the rich and the poor between cities.

(3) Data visualization is realized. The results of data analysis are vividly represented by line chart, bar chart and map, and the charts are displayed on the web by integrating flask with pyecharts.

Acknowledgments

This work was financially supported by the funding of Qianfan project of Jiangsu Maritime Institute (Big data analysis and application research team), Young academic leaders of Jiangsu Colleges and Universities QingLan Project.

References

[1] Wang J , Guo Y . Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao. 2012: 44-52

[2] Benymol Jose, Sajimon Abraham. Performance analysis of NoSQL and relational databases with MongoDB and MySQL. 2020, 24(Pt 3):2036-2043.

[3] Radulescu-Banu P. Personal finance with Python: using pandas, Requests, and Recurrent. Computing reviews, 2019, 60(12):447-448.

[4] Deqing Li, Honghui Mei, Yi Shen, et al. ECharts: A declarative framework for rapid construction of web-based visualization. 2018, 2(2):136-146.

[5] Grinberg M . Flask Web Development: Developing Web Applications with Python[M]. O'Reilly Media, Inc. 2014.