ISSN: 2688-8653

Design and Implementation of Movie Data Analysis System based on Douban

Taizhi Lv *, Yujuan Zou

College of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing, 211170, China

Abstract

Douban movie website is a movie recommendation and rating website which people like to use. Through crawling, analyzing and visualizing the movie data from Douban website, this paper can present the development trend of movies. Supported by Python technology, this paper crawls the movie data of Douban website by Requests and Beautifulsoup library, stores the data to MySQL database by PyMysql library, and uses Pandas, Numpy and other library to sort out and analyze the data, visualize by Matplotlib graphics library to show the results of the analysis. With the help of visual graphics to analyze the data of the film industry, the development trend of the film industry is to be understand.

Keywords

Data Acquisition; Python; Matplotlib; MySQL; Visualization.

1. Introduction

With the advent of the era of big data and artificial intelligence, people have gradually reached a consensus on the value of data. Big data can help an enterprise make decisions based on widely collected information and use it in many different ways. The business world uses big data sets to inform and guide business processes [1]. So the acquisition of data becomes very important. After getting the data, we can analyze and visualize the data, and finally draw a conclusion. From the perspective of the whole process, obtaining data is the first task of the whole process [2].

So how to get the data means is particularly important. In so many programming languages, data crawling through Python language is more simple and easy to operate. Because the python language is very simple, and it's perfect for human reading. Even if we don't understand the nature of language, we can solve the problem. Second, python has a powerful standard library [3]. Python also has a definable third-party library to use. It can help with a variety of tasks, including regular expressions, document generation, databases, and so on. So it is very easy to crawl movie data with Python language.

This paper uses Python to crawl Douban movies, and then visualizes the information to get the number of different types of films, the number of films in different countries and regions, and the number of films in each year. Finally, according to these data information, the collected film data are analyzed, the current situation of the film industry is analyzed, and the future development of the film industry is predicted, and the suggestions that can be used for reference are put forward.

2. Data Acquisition

2.1. Related Technologies

The requests library is a Python crawler library which is very easy to understand and easy to use. In the requests library, all methods are ultimately called by the request () underlying

ISSN: 2688-8653

method [1]. As a web page analysis library, Beautifulsoup library has the characteristics of convenience and flexibility [2]. In the aspect of web page parsing, the library has high processing efficiency and supports a variety of parsers to parse.

2.2. Crawl List Page

Get all the information of the list page by get method of requests library.

response= requests.get (url,headers=headers,cookies=cookies)

In the process of crawling, considering the speed of crawling is too fast and the amount of crawling data is too much, it may be detected by the anti crawler mechanism of the website itself. Here you add the headers and cookies parameters. Simulate the state of people's access by setting parameters. Some codes are as follows:

```
 headers = {
"User-Agent": "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/65.0.3314.0 Safari/537.36 SE 2.X MetaSr 1.0"
 }
```

The data obtained from the request website is a string in JSON format, which needs to be decoded and converted into Python objects. The load() method of JSON is required. The code is as follows:

json_ data= json.loads (response.text)

2.3. Parse Details Page

According to the obtained string information, the web of the secondary web page is obtained, and the requests library method is used to access the detailed information of the movie in the detail web page. The code is as follows:

res = requests.get (url2, headers=headers,cookies=cookies)

Get all the information of the detail page, and then analyze the detail page through the beautifulsoup library, and extract the required content in the form of tags.

3. Data Storage and Analysis

3.1. Related Technologies

Pandas is a python data analysis package based on Numpy. Pandas provides efficient tools and methods to make data analysis more rapid and convenient. Series and dataframe in pandas are effective tools for data analysis [3]. MySQL is one of the most popular relational database management systems. The language used by MySQL is the most commonly used standardized language SQL to access databases [4].

This paper mainly uses pymysql library method to store the data, through pandas to sort out and analyze the data.

3.2. Data Storage

Call PyMysql library to store the extracted data in MySQL database. To connect to the database, the code is as follows:

ISSN: 2688-8653

```
conn = pymysql.connect ('localhost', 'root', '123456', 'big_ data')
cursor = conn.cursor ()
```

The data is inserted by SQL statement, and the database is closed finally. In order to prevent data crawling interruption phenomenon, automatic connection processing is executed. Mainly through the calculation of the number of database data, the number of pages crawled is known. The current page crawled data is stop, and then crawled again, which greatly reduces the crawling of duplicate data, and also avoids artificial pause, and continues crawling to do unnecessary calculation.

After the flight data is crawled, the obtained air_ info.csv The file is saved in JSON format and added to the project folder for development and call. In fact, the JSON file is used as a local server to transmit data to the node server. All data are analyzed and filtered by componentdidmount() function.

3.3. Data Cleanning

Data cleaning is mainly to process or delete dirty data. For records with missing values, you can use delete operation. You can first delete all lines with null field values. The code is as follows:

```
df.dropna (axis=0,how='all')
```

where axis = 0 represents the x-axis, and axis = 1 represents the y-axis. Similarly, you can delete any row with a null field value or any column with a null field value.

In addition to deleting missing values, we can also insert and supplement missing values. Inserting and supplementing missing values can reduce the loss of other attribute values. We usually use the insertion and supplement of mean value, or the insertion and supplement of intermediate value. This has the least impact on the data results. The Nan attribute in numpy can perfectly solve the insertion and supplement operation of missing value of numerical type. np.isnan () can be used to judge the missing of numerical type. The missing value can be directly assigned to mean value by mean method. For non numeric data, pandas isnull and notnull methods can be used. Fillna (A. mean ()) can be used to fill in the data after judgment, which also means to fill the average value in the position of null value. In the process of data collation, it is not necessary to fill in the values, but the number of types.

3.4. Data Analysis

Taking the data analysis of film types as an example, other types of operations are similar and will not be repeated. The specific operations are as follows:

Read with pandas Library_ The SQL method selects the required data file. This part mainly uses the data of type column. The operation of selecting column can be realized by using database language. The code is as follows:

```
df = pd.read_ sql("select leixing from movie", con=conn)
```

Since there is more than one type of each movie in the type column, split is used to split and the tolist method is used to store the type in a list. The code is as follows:

```
leixing = df['leixing']. str.split ('/').tolist()
leixing_ list = list(set([i for j in leixing for i in j]))
```

ISSN: 2688-8653

In order to calculate the specific number of movies for each type of movie, we use the method of constructing an all-0 array to transform the grouping type into a datafram array with all zeros, traverse the assignment, change the existing data to 1, and finally calculate the number of 1 to get the number of movies of this type. You can use numpy here np.zeros The () method creates an all-0 array of movie types to calculate the number of movies of each type. df.shape [0] used to determine the movie classification information of dataframe, len (Leixing)_ List) to determine the number of rows in the list.

4. Data Visualization

4.1. Related Technologies

Matplotlib library is a two-dimensional drawing library method of Python. Most of the data can be visualized through a few lines of code. The library includes drawing histogram, drawing line chart, drawing pie chart, drawing column chart, drawing scatter plot and so on. And we can add x-axis and y-axis, X-axis label and y-axis label, header information and other information to these basic charts. Not only that, matploylib library can also change the size of the drawn table, the background of the table graph, the color of the lines in the table graph, etc.

4.2. Film Statistics by Country

To calculate the number of films in each country, it is similar to the type, mainly because the information extracted from the database is different. This part can be realized by SQL statement, and the code is as follows:

df = pd.read_ sql("select contory from movie", con=conn)

Because there are too many categories of countries and regions, it is not convenient to display them all on the graph. It is more meaningful to select the countries and regions with more than 30 movies in the top 30. The obtained count array is intercepted and sorted. Ascending = false means descending sort, and head (30) represents the countries and regions with the top 30 ranking. The code is as follows:

count = count.sort_ values(ascending=False).head(30)

The drawing operation of count array is basically the same.

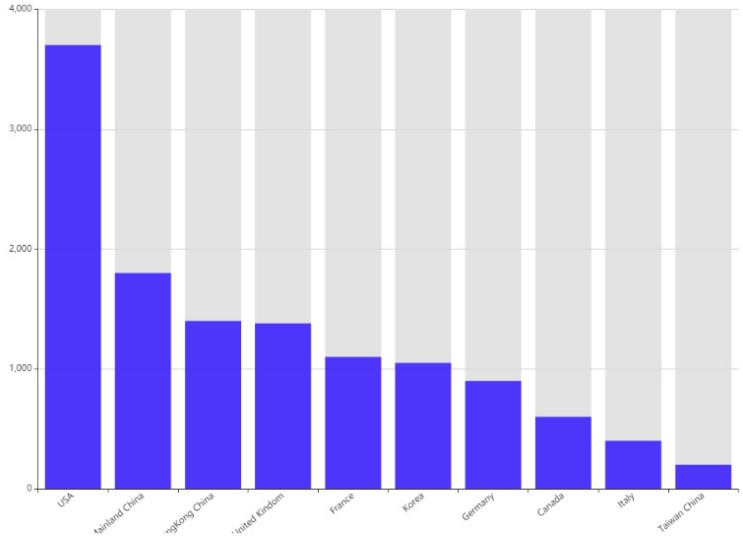

Figure 1. The bar-chart of film statistics by country (Top 10)

ISSN: 2688-8653

4.3. Proportion of Film Types

Because the pie chart cannot display too many data types, the top 10 movie types are selected. The code is as follows:

count = count.sort_ values(ascending=False).head(10)

Labels are the descriptive text to determine the outside of the pie chart. Size is to determine the proportion of the value in the pie chart. Expand is to determine the distance between each part of the pie chart and the center. In the case of more classification, the reasonable use of expand can avoid the overlapping of the words on the outside of the pie chart, so that the pie chart drawn can be better presented. Autopct can change the size of the scale font in the pie chart. The larger the number, the larger the font in the pie chart, and the smaller the number, the smaller the font in the pie chart. Shadow is used to determine whether a shadow is drawn below the pie chart. The default value is false, which means no shadow is drawn. Otherwise, shadow is drawn. Startangle represents the angle from which the pie chart is drawn. The default value is drawn counterclockwise in the positive direction of the x-axis. Radius is used to control the radius of the pie chart, so as to control the size of the pie chart. Its default value is 1, so as to avoid too many categories and unclear font graphics. You can adjust it by yourself. Labeldistance is the radius of annotation information.

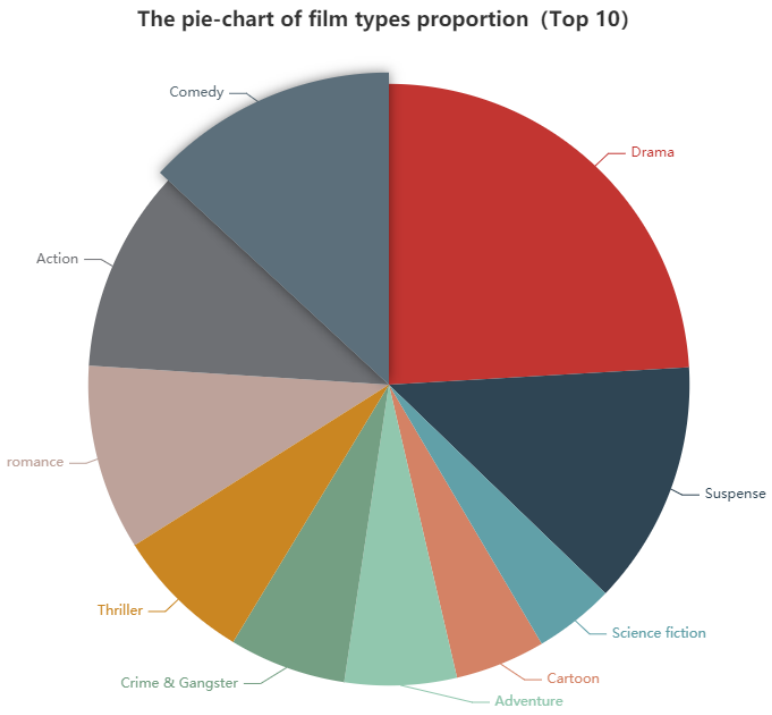

Figure 2. The pie-chart of film types proportion (Top 10)

5. Conclusion

This paper uses Python to realize the crawler of Douban movie website information. This crawler aims at the movie information on douban.com, and uses pandas and Matplotlib to analyze and visualize the information. In this paper design, we crawl the Douban web page in advance. The most important thing in this stage is to analyze the network source code information, store all the crawled data into MySQL database, and then use Python third-party library pandas to process and analyze the data. Pandas can handle all kinds of data, which is

ISSN: 2688-8653

very convenient and easy to understand. When the data is processed, the most important part of this paper, visual analysis. The data is presented by color, size and other specific representation, which provides the observer with clear information at a glance, so as to provide the observer with the basis for decision-making, and promote the enterprise to develop in a better direction.

Acknowledgements

This work was financially supported by the funding of the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province(2019SJA0650), 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] Haoshi Y U , Fangjun K . Crawler and anti-anti-crawler technology based on Python[J]. Intelligent Computer and Applications, 2018,04: 112-115.

[2] Chunmei, Zheng, Zuojie, et al. A Study of Web Information Extraction Technology Based on Beautiful Soup[J]. Journal of Computers, 2015:381-387.

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

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

[5] McClarren, Ryan G . Computational Nuclear Engineering and Radiological Science Using Python || NumPy and Matplotlib[J]. 2018:53-74.