Exploration of the Development Course of Data Visualization Based on Python and Its Multimedia application

Junkai Zhu, Yujun Hu

School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu, Anhui 233030, China.

Abstract

In the 21st century, with the arrival of the big data era, Python has ushered in a golden period of decades of development, and its application advantages in big data are gradually highlighted. On the other hand, data visualization technology has won the opportunity of application oriented development on the premise of the current social needs and Python's powerful Numpy library and Matplotlib library. By introducing and studying Python's powerful data analysis and visualization functions, this paper conducts scientific research on current popular research applications, and then introduces data visualization technology into real applications.

Keywords

Python; Data visualization; Big data; Third party library.

1. Introduction

How to analyze data and obtain large, complex and multi-dimensional information? Research shows that the vast majority of information about the external world obtained by human beings is obtained through visual channels, so the answer is to provide intuitive, interactive and responsive visual environment like human eyes. It can be seen that the development of data visualization technology is of great significance. For the definition of data visualization, there are many research directions in the academic circle at present. Zhang Jinlei [1], Shen Enya [2], Zhang Hao [3] and others have published their views and researches in their respective academic journals. In general, the two main components of data visualization: statistical graphics and topic maps, mainly for the presentation of information.In terms of the components of visualization, Chen Jianjun [4] believes that visualization technology can be divided into four processes:pretreatment, papping, drawing, and display.Visualization of scientific computing transforms the digital information involved in and generated by computing into intuitive physical phenomena or physical quantities that are represented by images or graphical information and change with time and space [5].

Data visualization technology refers to the technology of presenting data to users by means of intuitive charts, which is widely used in people's life and production in the 21st century, big data analysis and investigation and other fields. Chen Jianjun [4] and others mentioned in their research on data visualization technology and its application:as a high-level programming language, Python is both process oriented and object-oriented. Its advanced data structures, dynamic types, and interpreted languages make it a programming language for scripting and rapid application development on most platforms. It is easy to learn, free of charge, open source, portable, and embeddable.At present, many colleges and universities across the country have opened Python programming courses.Among the famous programming languages TIOBE, Python has entered the top three[6]. It is one of the languages that must be learned by the data research community and the academic community of universities. At the same time, the language ranks first in AI, big data and other fields of programming.Therefore, Python, as the first data research language, is so popular for no reason. It also has a huge advantage that other languages do not have: tens of thousands of third-party libraries that are completely open source, such as Lodash, ECharts, Panda, Numpy, Matplotlib, Seaborn, etc.

With the arrival of the age of big data, data visualization is getting more and more attention. Data visualization plays an important role in people's daily life. Visualization technology is also becoming increasingly mature. More and more visualization images appear in people's lives, becoming a subject that people gradually pay attention to and research. With the demand of social development, data visualization still has many problems, and faces huge challenges and various difficulties. The main difficulties faced in the development of data visualization are: 1. Visual noise is in the large-scale concentration of data, and most data have strong relevance to a certain extent, so it can not be separated as a separate object to display. Relevant equipment will be used during data processing and operation, and different types of noise will appear, which is also an important factor affecting the further development of data visualization. 2. Information loss is a serious phenomenon. In the process of data processing and visualization, enterprises will lose information. In the process of data visualization transformation, it will lead to the loss of relevant key information, which is also one of the difficulties in the process of data visualization transformation. 3. The realization of data visualization in the aspect of large image perception is restricted by the length ratio of equipment, the resolution of equipment and the perception of the real world. In terms of the development and extension of data visualization, it is restricted and developed. At the same time, it is restricted and affected by reality, which affects the actual effect of large image perception and also affects the process of data center visualization.

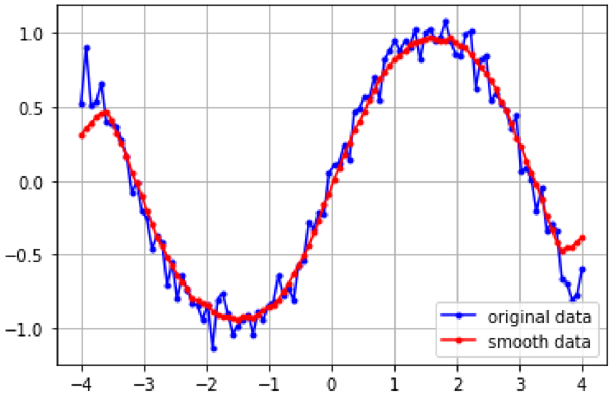


Figure 1 Data noise smoothing

As one of the third-party libraries in Python, Matplotlib is a multi platform data visualization library based on NumPy array, designed to work with a wider range of SciPy. It was conceived by John Hunter in 2002, and was originally used as a patch of IPython to achieve interactive MATLAB style drawing through gnuplot from the IPython command line. Fernando Perez, the founder of IPython, was completing his doctorate at that time, and John knew that he had no time to patch it for several months. John thinks this is a reminder of his own beginning. Matplotlib software package was born, and version 0.1 was released in 2003. When it was used as the drawing package selected by the Space Telescope Science Institute, it got an early promotion. The scientists behind the Hubble telescope financially supported the development of Matplotlib and greatly expanded its functions. Numpy is a third-party library for processing multidimensional array operations with the same kind of elements. Numpy library also includes trigonometric operation function, Fourier transform, random and probability distribution, basic numerical statistics, bit operation, matrix operation and other very rich functions. Seaborn is a Python data visualization library based on matplotlib. It provides an advanced interface for drawing attractive and informative statistical graphs.

2. Related Work

2.1. Real time data processing

Today, the society has entered the era of big data. The traditional data visualization analysis technology is not enough to express huge data information, and needs to rely on more effective data processing algorithms[7]. In the Internet era, data is always updated in real time, but the display of real-time data depends on data visualization. The value of real-time data can only be reflected by data visualization. Therefore, the focus of data visualization research today is how to establish a dynamic and interactive analysis method to express large-scale data. For example, in the annual "Double 11" Tmall Shopping Festival, Alibaba will display some key data on the network platform, and the constantly changing figures represent the real-time sales volume. The data jumping every moment is the result of data visualization.

When a large amount of data has been processed, it is usually necessary to conduct visual analysis on these data and show the information hidden in the data in the form of charts. The data visualization analysis can be implemented with Excel and Tableau. Excel has its own charts. Charts change with the change of data. They are presented in various forms and can generate various high-quality pictures. Tableau is a visual analysis tool, which can help us quickly make visual images. At the same time, its operation is simple, easy to learn, and the user experience is good.[8]But, Python is easy to operate and can provide a large number of third-party libraries and machine learning platforms, auxiliary data visualization is more convenient, big data is more accurate and visual with the support of the machine learning platform. There are many practical applications, most of which are using Python language. For example, research and visualization of Chinese word segmentation technology [9], visual analysis of film data [10], visual analysis of data based on tourist comments [11],etc.

2.2. Application of OpenGL 2.0

OpenGL2.0 is a specification that defines a cross programming language and cross platform programming interface. It is used to generate two-dimensional and three-dimensional images. This interface consists of nearly 350 different function calls, which are used to draw complex 3D scenes from simple primitives. Another programming interface system is Direct3D for Microsoft Windows only. OpenGL 2.0 is often used in CAD, virtual reality, scientific visualization programs and video game development.

The main drafter of the OpenGL 2.0 standard is not the original SGI, but 3Dlabs that gradually take the initiative in ARB. The first thing to do for version 2.0 is to have full compatibility with the old version, and to work with DirectX on vertex and pixel and memory management to maintain the balance of power. OpenGL 2.0 will consist of the existing features of OpenGL 1.3 plus fully compatible new features. With this, we can make a thorough reduction of various entangled extension instruction sets launched by various companies in the era of ARB stagnation. In addition, the implementation of hardware programmability also provides a better way to integrate existing extension instructions. With the continuous development and improvement of DirectX, the advantages of OpenGL are gradually lost. So far, although the version 2.0 advocated by 3Dlabs has come out, and many designs similar to the programmable units in DirectX have been added, the recognition level of manufacturers' users is not high, and the future development prospect of OpenGL is unclear.

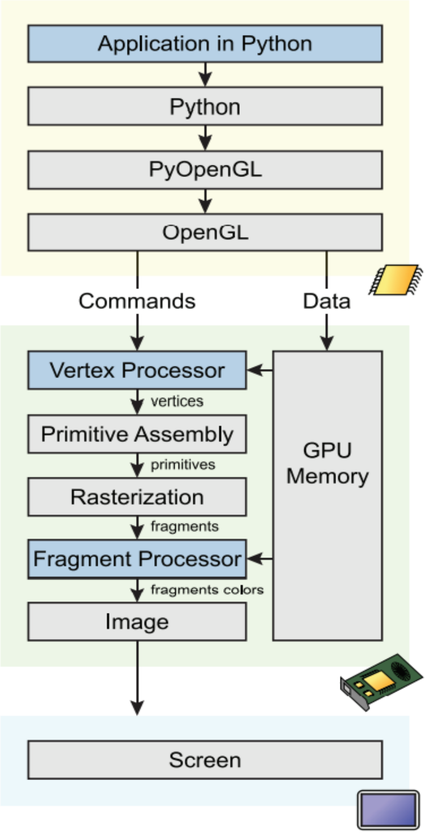


Figure 2 A diagram of the implementation process

2.3. Shader references in visualization

Shaders are powerful programs that were originally used to shade objects in 3D scenes. Today, shaders have many uses. Shader programs usually run on a computer's graphics processing unit (GPU), where they can run in parallel. Shader languages such as High Level Shading Language (HLSL) and OpenGL Shading Language (GLSL) are the most commonly used languages for programming GPU rendering pipelines. The syntax of these languages is similar to the C programming language. When you play games such as Minecraft, shaders are used to make the world look like 3D when viewing the world from a 2D screen (that is, your computer monitor or mobile phone screen). Shaders can also completely change the appearance of the game by adjusting the way light interacts with objects or the way objects render on the screen. You usually see that there are two forms of shaders.

Shaders are powerful programs that were originally used to shade objects in 3D scenes. Today, shaders have many uses. Shader programs usually run on a computer's graphics processing unit (GPU), where they can run in parallel. Shader languages such as High Level Shading Language (HLSL) and OpenGL Shading Language (GLSL) are the most commonly used languages for programming GPU rendering pipelines. The syntax of these languages is similar to the C programming language. When you play games such as Minecraft, shaders are used to make the world look like 3D when viewing the world from a 2D screen (that is, your computer monitor or mobile phone screen). Shaders can also completely change the appearance of the game by adjusting the way light interacts with objects or the way objects render on the screen. You usually see that there are two forms of shaders.

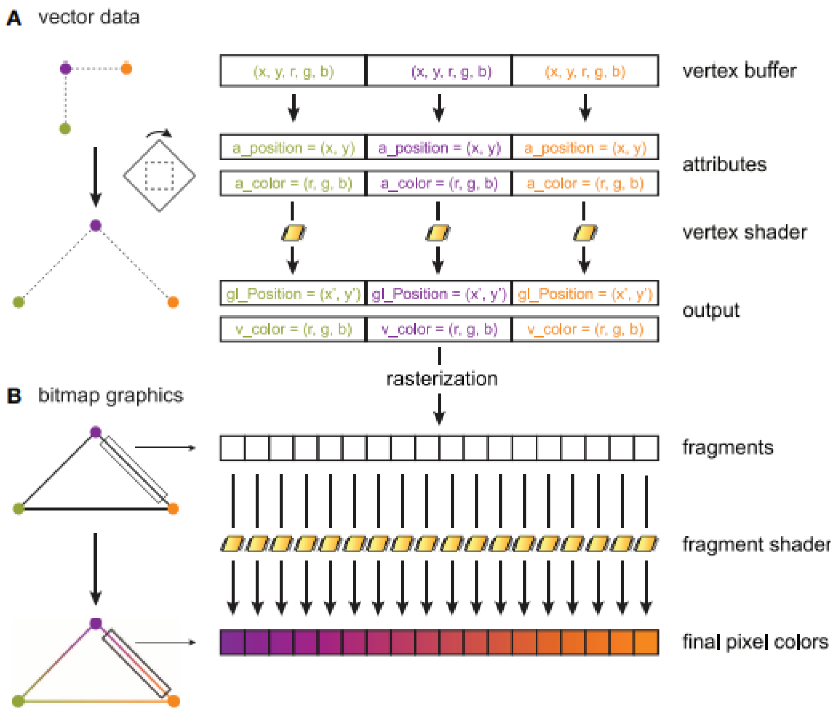


Figure 3 Vector data and bitmap graphics

3. Method

3.1. Matplotlib

Matplotlib is a Python drawing library, which allows users to easily graph data and provide a variety of output formats. Matplotlib can be used to draw various static, dynamic and interactive charts. It is a very powerful Python drawing tool. We can use it to present a lot of data more intuitively in the form of charts. Matplotlib can easily complete line chart, scatter chart, contour chart, bar chart, histogram, 3D graphics, and even graphic animation.At the same time, Matplotlib uses a hierarchical structure to form a code base. At the top is its state machine environment, which is provided by the matplotlib. pyplot module. At this level, simple program functions can be used to add line, image, text and other related drawing elements to the visualization graphics to be generated by the current program.

```
import matplotlib.pyplot as plt
import math
import numpy as np
t = np.arange(0,2.5,0.1)
y1 = np.sin(math.pi*t)
y2 = np.sin(math.pi*t+math.pi/2)
y3 = np.sin(math.pi*t-math.pi/2)
plt.plot(t,y1,'b*',t,y2,'g^',t,y3,'ys')
plt.show()
plt.axis([0,5,0,20])
plt.title('My plot', fontsize=20, fontname='Times New Roman')
plt.xlabel('Counting', color='gray')
plt.ylabel('Square values',color='gray')
plt.text(1,1.5,'First')
plt.text(2,4.5,'Second')
plt.text(3,9.5,'Third')
plt.text(4,16.5,'Fourth')
plt.text(1.1,12,r'$y = x^2$', fontsize=20, bbox={'facecolor':'yellow','alpha':0.2})
plt.grid(True)
plt.plot([1,2,3,4],[1,4,9,16],'ro')
plt.plot([1,2,3,4],[0.8,3.5,8,15],'g^')
plt.plot([1,2,3,4],[0.5,2.5,4,12],'b*')
plt.legend(['First series','Second series','Third series'], loc=2)
plt.show()
```

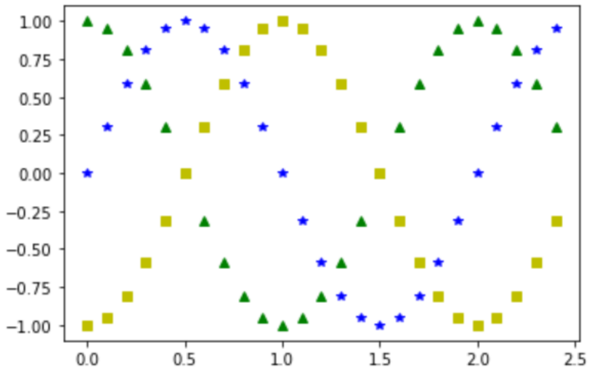


Figure 4 Draw simple point graph with Matplotlib

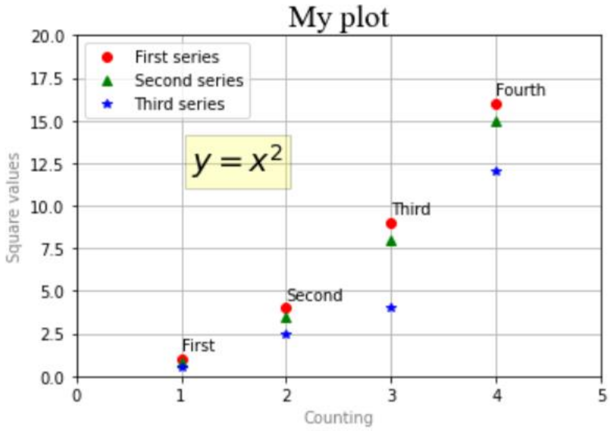


Figure 5 Draw a simple grid diagram with Matplotlib

3.2. NumPy

NumPy is a Python package, more specifically it is the basic package of high-performance scientific computing and data analysis. It stands for "Numeric Python". It is a library composed of multidimensional array objects and a collection of routines used to process arrays. Numeric, the predecessor of NumPy, was developed by Jim Huguin. Another package Numarray has also been developed, which has some additional functions. In 2005, Travis Olivant created the NumPy package by integrating the functions of Numarray into the Numeric package. This open source project has many contributors.

In general, the NumPy library has the following characteristics:

a. The core part of the NumPy library is the array object. It encapsulates n-dimensional arrays of homogeneous data types, and its functions will be presented in the form of demonstration code.

b. All elements in the array must be of the same type and have the same size in memory.

c. Array elements can be described by index, and the index sequence number starts from 0.

d. The dimension of NumPy array is called rank, and each linear array is called an axis.

e. Provide broadcast function for calculation between arrays.

f. A tool for integrating C/C++/Fortran code.

g. Linear algebra, Fourier transform, random number generation and other functions.

3.3. Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides an advanced interface for drawing attractive and informative statistical graphs. For a brief introduction to the ideas behind the library, you can read introductory notes or papers. Visit the installation page to learn how to download the package and start using it. You can browse the sample library to learn something you can do with seaborn, and then view the tutorial or API reference to learn how to do it. To view the code or report errors, visit the GitHub repository. The general support problems are mostly on the stackoverflow website, which has a special seaborn channel.

4. Experiments

4.1. Visualization of film reviews

With the booming of the film industry and the increasing competition in the film market, more and more studios are trying to understand user preferences through the changes in user ratings for different film genres. Ratings are the feedback of the audience from the point of view of acceptance, depending on the artistic quality of the film itself and the audience's needs, i.e. the extent to which the film meets the audience's expectations. Then it is an important question to understand the influence of film genres and ratings on audience's choice. In the following, we take movie data as an example and use python programming to visualize the relationship between the number of reviews and ratings to further analyze users' sentiment towards movies[12-14].

The graph below shows the joint distribution of the number of comments and ratings. Compared with other common analytic visualizations, Python's joint distribution can directly calculate the correlation and show it in the graph while visualizing, no separate calculation is needed.

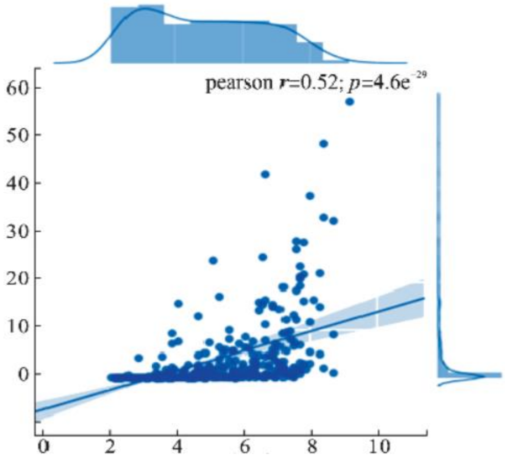


Figure 6 Correlation Analysis

The correlation coefficient is 0.52, which is a moderate correlation (the correlation coefficient is between -1 and 1, and the closer the correlation coefficient is to 1, the greater the positive correlation between the number of reviewers and the box office). This is statistically significant.

In general, movies with less than 100,000 reviewers and a large number of reviewers are also rated relatively high.

4.2. Visualization of tourist attraction data

Data acquisition refers to the use of a device that automatically collects data from various data sources into a device, based on where to go website, through the Python crawler technology to obtain the corresponding data, the process of data acquisition: initiating a request, obtaining the address, analyzing the web page, extracting data, and storing data.

Data crawling is mainly realized by Python crawler technology[15], by launching a request to the specified website, obtaining the data from the server's response, parsing it, and then storing it in Ex-cel. This crawl uses the Requests library, which is a third-party database for Python that automatically decodes the content from the server and makes informed guesses about the corresponding encoding based on HTTP headers, and makes a request to the specified URL to get the response information of the data page that we want to crawl. Crawling web data is most often used is the general crawl framework, its biggest role is to allow users to effectively, stable and reliable crawl the web content.

To get the data of Xi'an tourist attractions on the Where to go website, we need to get the URL link of the corresponding data to initiate the request. When we open the Where to go website, we find that the attraction data is divided into 49 pages, each with a different URL, so we need to find the relationship between the URLs of each page. By flipping through the pages, we found that except for the first page, the URL of each page has a page= parameter that changes at the end, that is, starting from 1, the page number corresponds to the number of each page, so we can build a for loop for requesting URLs.

There are many parsing libraries in Python, and it is most convenient to look at the web page data and choose the BeautifulSoup library to parse the response. object s', and then extracts the data corresponding to that tag through its .text property. To find a single tag you can use the find() method, and to find multiple tags you can use the find_all() method, where the key is to pinpoint the tag with a statement.When the data of the attraction is acquired and stored, it is necessary to process the data and eliminate the data that do not meet the visualization requirements. The first step is to de-duplicate the data. Excel has its own de-duplication function, so the data can be de-duplicated by Excel. After the de-weighting, the data is checked again and found to meet the analysis requirements, and no other data processing operations are needed.After the data is processed, the data should be visualized and analyzed to show the hidden information in the form of graphs. Tableau is a visual analysis tool that can help us to make visual images quickly, and it is simple to operate, easy to learn and has a good user experience.

The word cloud is a cloud-like color pattern composed of words, and the size of the area occupied by a word in the word cloud represents its frequency, which is visually more intuitive. By analyzing the word cloud of Xi'an attractions, we can easily find the characteristics of Xi'an[16]. The mention of words such as culture, history, and museum indicate that Xi'an is a typical cultural ancient city, while the mention of words such as leisure, entertainment, and paradise reflect that Xi'an has a relaxed and diverse tourism environment.

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
filename = "covid19.txt"
with open(filename) as f:
mytext = f.read()
wcloud = WordCloud(width=2800, height=1600).generate(mytext)
plt.imshow(wcloud)
plt.axis('off')
plt.show()
```

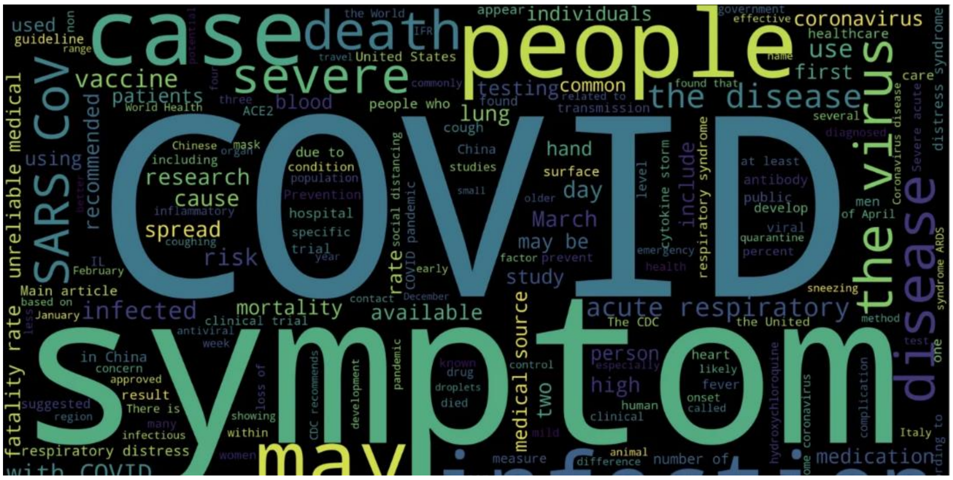


Figure 7 Analysis of Scenic Spots' Word Cloud

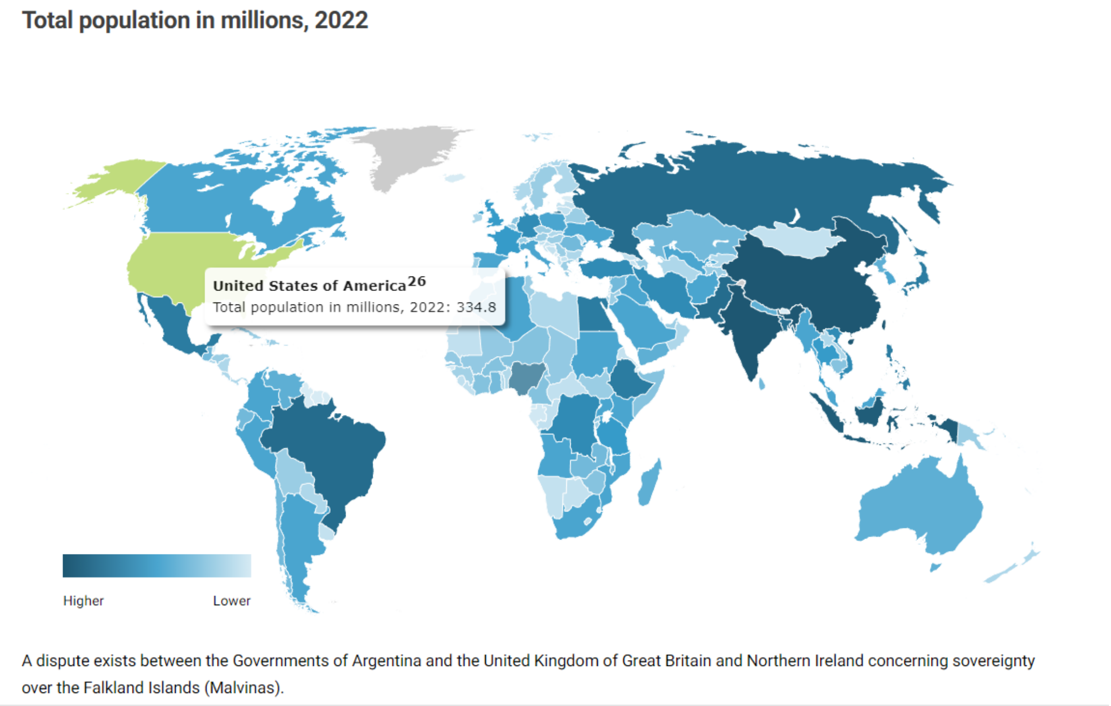
4.3. The Application of Visualization Technology in Multimedia Technology and the Expansion of Crawler Technology

Multimedia technology integrates multiple media such as text, graphics, images, sound, animation and video as well as the combination of various media[17].So as to intuitively, vividly and vividly express space environment information in the form of visual, auditory, tactile and other media. The visualization method transforms a large amount of unordered spatial environment information into intuitive information that is easy to operate and perceive by human beings. The combination of multimedia technology and visualization, GIS and cartography has promoted the emergence and development of spatial information visualization. This paper discusses and studies spatial information visualization based on multimedia technology from the following aspects.

Python has the unique advantages of the language, making it suitable for data analysis and data visualization in various fields, including the application of multimedia visualization technology. The definition and expression of visual icons are described with Python language and mathematical methods. The following is a case of python based multimedia visualization.

The introduction of data related to multimedia technology requires Python crawler applications. Teng Yifang et al. [18] successfully realized the data extraction of multimedia related materials. The combination of data crawler and data visualization can complete a powerful multimedia platform with visual shock, real-time data update. In addition, with the rapid update and development of Python, audio and video visualization processing has also won a new breakthrough.

On the official website of the United Nations, the World Population Dashboard multimedia displays global population data, including fertility, gender parity in school enrolment, sexual and reproductive health information, etc. In summary, these data reveal the health and rights of people around the world, especially women and young people. The figures here are from UNFPA and other United Nations agencies and are updated annually. Python language has its unique advantages in crawler technology and visualization technology, and has unlimited prospects in the future multimedia and big data era.

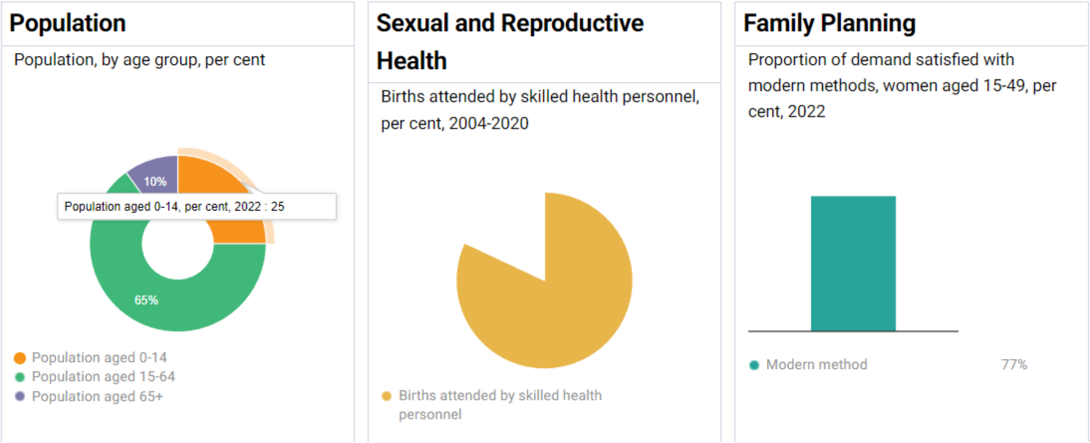


Total population in millions, 2022

United States of America[26]
Total population in millions, 2022: 334.8

A dispute exists between the Governments of Argentina and the United Kingdom of Great Britain and Northern Ireland concerning sovereignty over the Falkland Islands (Malvinas).

Figure 8 Data visualization multimedia platform based on global population for official statistics of the United Nations[19]



Population

Population, by age group, per cent

Population aged 0-14, per cent, 2022 : 25

Sexual and Reproductive Health

Births attended by skilled health personnel, per cent, 2004-2020

Family Planning

Proportion of demand satisfied with modern methods, women aged 15-49, per cent, 2022

Figure 9 Multimedia Visualization of the UN Official Website Supported by Big Data[20]

5. Conclusion and Future Work

This paper describes the development of Python-based data visualization aspects and analyzes the application of the language in data crawling and visual analysis. By analyzing the movies released on Douban.com, we can give some production tips to movie companies and provide important reference indicators for users to watch movies. It is clear from the example that using Python language for visualization analysis can better apply a large amount of complicated data, make the visualization results more visual and more intuitive to be understood by the public, and greatly improve the work efficiency. M atplotlib and Seaborn, the famous Python plotting libraries, can easily plot various statistical graphs, and with the powerful data processing and scientific computing capabilities of the Python language, they have excellent performance in the analysis and processing of film data. Compared with the existing ThemeRiver technology and common network diagram and line graph visualization, Python data crawling and visualization analysis has simple and efficient characteristics, and can draw statistical graphics that other drawing tools cannot draw, which has a good development prospect.

With the development of technology, database technology based on computer network technology has been used in all aspects of our lives, providing new ways to process and integrate information and data, and VB programming development can support the construction of database programs, not only to promote the use of database space, but also to provide binding services for the database. However, when using VB programming development, it is not possible to operate directly on the database, and it is necessary to copy the information in the database by means of object variables and then realize VB programming development. This requires staff to make good use of various key technologies, understand the problems that may be encountered in the application of database access technology in VB programming development, and do a good job of false alarm prevention and security protection.

References

[1] Zhang Jinlei, Zhang Baohui, Liu Yonggui. Research on the Application of Data Visualization Technology in Teaching [J]. Modern Distance Education Research, 2013 (06): 98-104+111.

[2] Shen Enya. Big Data Visualization Technology and Application [J]. Science and Technology Bulletin, 2020, 38 (03): 68-83.

[3] Zhang Hao, Guo Can. Research on Application Trend and Classification of Data Visualization Technology [J]. Software Guide, 2012,11 (05): 169-172.

[4] Chen Jianjun, Yu Zhiqiang, Zhu Yun. Data visualization technology and its application [J]. Infrared and Laser Engineering, 2001 (05): 339-342.

[5] Ren Yonggong, Yu Ge. Research and Progress of Data Visualization Technology [J]. Computer Science, 2004 (12): 92-96.

[6] Wei Yiyang, Wu Yifan, Li Yong.Research on the Application of Python Technology in Data Visualization [J]. Fujian Computer, 2022,38 (01): 27-31.

[7] Cao Shengjia; Zeng Yunhan; Yang Shangru; Cao Songlin. Journal of Physics: Conference SeriesVolume 1757, Issue 1. 2021.

[8] Tian Xueli, Guo Zhibin, Liu Mengxian. Python based web data crawling and visual analysis [J]. Computer Knowledge and Technology, 2022,18 (06): 24-26. DOI: 10.14004/j.cnki.ckt.2022.0312.

[9] Zhu Yongzhi, Jing Jing. Research on Chinese Word Segmentation Technology Based on Python Language [J]. Communication Technology, 2019, 52 (07): 1612-1619.

[10] Zhao Hanyuan. Visual Analysis of Book Data Based on Python Crawler [J]. Electronic Technology and Software Engineering, 2021 (14): 178-179.

[11] Jia Yanping, Zhai Jingang. Visualization analysis of tourist comment data based on Python reptile technology [J]. Journal of Anyang Normal University, 2021 (05): 51-54. DOI: 10.16140/j.cnki.1671-5330.2021.05.013.

[12] Cai Wenle, Zhou Qingqing, Liu Yuting, Qin Lijing. Visualization analysis of Douban film review data based on Python reptiles [J]. Modern Information Technology, 2021,5 (18): 86-89+93. DOI: 10.19,850/j.cnki.2096-4706.2021.18.022.

[13] Gao Wei, Sun Panpan, Li Dazhou. Visual analysis of movie data based on Python reptiles [J]. Journal of Shenyang University of Chemical Technology, 2020,34 (01): 73-78.

[14] Cheng Wenying, Li Xiumin. Research on Python based movie data crawling and data visualization analysis [J]. Computer Knowledge and Technology, 2019, 15 (31): 8-10+12. DOI: 10.14004/j.cnki.ckt.2019.3647.

[15] Jia Yanping, Zhai Jingang. Visualization analysis of tourist comment data based on Python reptile technology [J]. Journal of Anyang Normal University, 2021 (05): 51-54. DOI: 10.16140/j.cnki.1671-5330.2021.05.013.

[16] Zhu Yongzhi, Jing Jing. Research on Chinese Word Segmentation Technology Based on Python Language [J]. Communication Technology, 2019, 52 (07): 1612-1619.

[17] Wang Jianhua. Research on Spatial Information Visualization Based on Multimedia Technology [J]. Journal of Surveying and Mapping, 1999 (01): 94+87.

[18] Teng Yifang Design and Implementation of Ontology based Multimedia Material Web Crawler [D]. Jilin University, 2015.
[19] https://www.un.org/en/global-issues/population.
[20] https://www.unfpa.org/data/world-population-dashboard.