Journal of Education and Educational Research
ISSN: 2957-9465 | Vol. 3, No. 1, 2023

Teaching Case Design of Python Data Analysis Course for Non-computer Majors

Ling Wu*, Jinsong Chen, Jian Zhou

Anhui University of Finance & Economics, Bengbu, Anhui 233030, China

* **Corresponding author**: Ling Wu (Email: wudapang1977@163.com)

Abstract: With the development of big data and artificial intelligence technology, the demand for data analysis in various industries is increasing. How to enable non-computer majors to master data analysis skills is a problem worth thinking about in the teaching process. Firstly, this paper analyzes the difficulties of non-computer majors in learning the course of data analysis, and designs a case based on film data. The teaching practice shows that the interest of the course has been greatly improved, and it can more effectively guide the non-computer major students to quickly master the core skills of data analysis.

Keywords: Data Analysis; Python; Nnon-computer Major; Movie Data.

1. Introduction

With the development of various application technologies of bigdata and artificial intelligence, there is an increasingly urgent need for practitioners in various industries to use computer programming to solve practical problems in professional fields. As a programming language, Python has a simple grammar structure, powerful function and a wide range of applications, which is suitable for lower grades and non-computer majors. Learning Python language can cultivate computational thinking and improve innovation ability, so it is widely used in many universities [1-3]. Python has a wide range of applications, but in terms of universality, no matter which major has certain requirements for data analysis and processing. For non-computer students majoring in humanities, social sciences and economic management, they also have many opportunities to encounter large amounts of text and data in their future work [4]. How to enable the students of these majors to obtain the basic data analysis and processing ability is a question worth studying and exploring [5-6].

2. The Teaching Difficulties of the Data Analysis Course

Anhui University of Finance and Economics has put forward the "new economic and management" strategy in 2018, and has opened a data analysis course group for economic and management majors [5]. As the last course of the course group, "Big Data Analysis and Application" is to master the basic methods of data analysis and visualization using Python + Numpy + Pandas + Matplotlib, and initially master practical development skills such as crawler and Chinese text analysis. The content of this course is related to many aspects of data analysis, and various operational changes of data. One function can often be realized in many ways and is more flexible. For most non-computer science students, this course is still relatively difficult. Although students can feel the practicality of the course in the learning process, if they only stay at the level of grammar analysis in the teaching process, they will be very tired to learn, and they need some simple projects to integrate the knowledge points they have learned. In recent years, we have been engaged in teaching Python programming in non-computer science majors, data analysis and visualization. In the teaching process, we have been thinking about and trying to introduce some cases suitable for non-computer major students to learn and feel. After several rounds of teaching exploration, we chose the film data analysis that students are familiar with to carry out teaching activities. The knowledge points of the course are coherent in the form of case practice. Many students are interested in the study of the case, and used their skills to other application scenarios, and received good teaching results.

3. Overall Design of Movie Data Analysis Case

In order to connect the knowledge points and help students grasp the whole process of data analysis in a macro level, we designed a movie data analysis case, and the framework of the whole teaching case is shown in Figure 1.

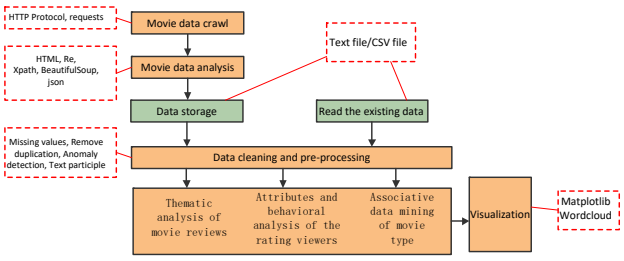


Figure 1. Overall framework of movie data analysis case

3.1. Movie Data Crawl and Analysis

For most non-computer majors, they have never touched the technology of crawling data, which is full of freshness and mystery. First of all, students need to understand that the main purpose of the crawler is to obtain the content of the web page and analyze it. In order, there are many ways to achieve this purpose. For acquiring web pages, two third-party modules requests and crawler framework Scrapy of Python, and regular expressions, Xpath and BeautifulSoup can be used for parsing web page content. The way of obtaining web pages and the way of web page analysis can be freely combined. With the enhancement of the awareness of anti-climbing, we found that the previous way only introduced to obtain HTML

source code + BeautifulSoup parsing is no longer enough. More and more websites hide the data, most of which take the form of dynamic loading and only allow to crawl the limited data. Most of the data returned are in json format, so we also focus on the acquisition and analysis of the data in json format in our teaching.

In the specific case teaching, first we introduce HTTP protocol, HTML tag and json, so that students can know the difference between GET and PUT, status code 200 and 404, the meaning of common HTML tag elements and tag attributes, and the data structure of json object and json array. On this basis, we first climbed the movie reviews of My Motherland and I, Wolf Warrior 2, Titanic, and the way is to obtain HTML source code + BeautifulSoup analysis + CSV save data. Later, we guided the students to use the developer tool to observe how to dynamically load json data on the page. We climbed the 2008-2021 box office list in the Enment movie data list, the user comments and user information of Maoyan movie, and adopted the method of parsing json format + CSV for data saving. Through these examples, students can deeply feel the richness of data formats and the diversity of acquisition means in the network.

3.2. Cleaning and Preprocessing of the Movie Data

The crawled data often has some missing values, dirty data and format non-compliance, which need to be cleaned and pre-processed before analysis. For the crawled movie data, we show two levels of data cleaning to students. The first level is the simple data cleaning such as missing values and repeated values, and the second level is the cleaning of the data format. For example, according to the number of films and box office, 24 films in the list have two genres. Such as Avatar is science fiction/action, Hancock is action/comedy. The genre of the film is separated by /. Such films should be considered twice when conducting the number and box office statistics. This is a small difficulty in data cleaning, and there can be a variety of processing solutions. We guide students to use Pandas to handle the number and box office respectively. After the students made many attempts, the teacher summarized the different processing methods, and the students had a deeper understanding of Panda data processing. The cleaned data can be used for various statistical analyses. Figure 2 is a bar chart and pie chart of the total box office by film type in 2008-2021. It can be seen that in the 14 years of film box office, people still generally like comedy films.

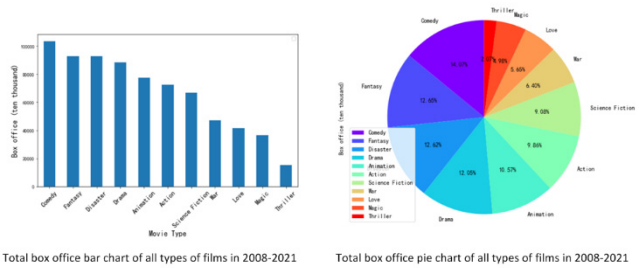


Figure 2. Bar chart and pie chart of the total box office of various genre films in 2008-2021

In addition to using the climbing data, students can also be introduced to some classic movie datasets. For example, some movies datasets are provided on Kaggle, MovieLens, IMDB and other websites. For example, Kaggle provides a set of data of 1000 most popular films from 2006 to 2016, which can guide students to conduct multi-angle analysis, such as obtaining the average score of scores, number of directors, distribution of ratings and film duration, classification of films, etc. MovieLens provides the rating data of multiple users for multiple movies, including movie metadata information and user attribute information. These data can be analyzed by the popular movies with the most reviews, the movies with the high average score, the most controversial movies and so on. These data are well used as material for data analysis.

3.3. Thematic Analysis of Film Reviews

The thematic analysis of film reviews is mainly the production of wordcloud map. The production of the wordcloud map is a very interesting highlight in the data analysis course, which has aroused the great interest of the students. Each movie has thousands of reviews, and users need to divide each Chinese sentence into processing. We chose jieba, which is better at performance, accuracy, and scalability. First, we preprocessed the movie reviews we crawled down, removing expressions and special characters and leaving only words. After using jieba, the number of keywords was counted, and then the word cloud tool WordCloud was used to generate wordcloud maps.

3.4. Attributes and Behavioral Analysis of the Rating Audience

In addition to the analysis of the movie data itself, the attribute and behavior analysis of the movie-viewing users is also an integral part of our case. After climbing the movie Nezha's relevant data of gender, city, score, likes and user level of the review users, the gender distribution, location distribution and scoring distribution of the rating audience can be analyzed, as shown in Figure 3.



Figure 3. Property and behavioral analysis of the rating audience

As can be seen from the Figure 4, this kind of domestic animated blockbuster is more in line with the public taste. According to the gender distribution data, although nearly half of them do not want to disclose their gender, judging from the known data, the ratio of men to female is a little more than 1:1, and the proportion of female audience is slightly higher. In terms of the rating distribution, most of the scores were given full marks, accounting for about 80% of the total.

3.5. Association Data Mining of Movie Types

As a data analysis course, after the students learn the basic content, they can properly learn some data mining and machine learning content. Due to the limitation of class hours, this part of the content should not be too difficult, focusing on the guidance. Machine learning requires a lot of basic knowledge. Therefore, after careful discussion by the teachers, it is found that the Apriori algorithm is well representative and does not involve the basic concept of machine learning temporarily. After using Pandas to organize the data, students can directly use the algorithm, which is suitable for leading beginners into the world of data mining. After introducing the classic beer and diaper stories, the students learned that the

association rules mining helps to find the relationship between different commodity items in the transaction database and find out the customers' purchasing behavior patterns. On this basis, we guide the students to think: similar to the analysis of multiple items in a shopping basket, there are often many types of a movie, such as science fiction/action/plot, etc. Can we analyze the type fields of each movie and find out the movie genre with the highest correlation? Taking the smallest data set provided by MovieLens as an example, the metadata file contains a total of 9742 movies in 22 types, each with multiple types. Before using the Apriori algorithm, we need to use Pandas to organize the data into an input format suitable for the algorithm, and the knowledge points such as group summary, setting index, filtering and sorting, transformation of DataFrame structure, and exclusive hot coding will be used. Organically combines the functions of Pandas data analysis. After understanding the basic concepts of support, confidence and promotion, students used the mlxtend package to analyze the types of 9742 movies and ranked them in descending order, and got the conclusion that Children and Animation were the most related.

4. Conclusion

After completing the data analysis process through case teaching, we assigned a data analysis task in the final big assignment and asked students to write the analysis document. Students also look for their topics of interest and apply the learned skills in multiple scenarios. We found that most students could use the skills they had mastered to complete a relatively complete analytical example, achieving better results. In view of the characteristics of non-computer science students, data analysis courses can help students grasp the core skills of Python data analysis in a short period of time through a complete data analysis process. In order to drive the whole teaching process with the cases, teachers must strengthen their learning and constantly sort out the cases. These cases are characterized by practical value, high fit with the major, exquisite and coherent, which can meet the needs of teaching. In this way, the data analysis class can become colorful, and students can master Python data analysis skills more effectively.

Acknowledgments

This work is supported by General teaching and research project of Anhui University of Finance and Economics, Research and practice of case-driven big data application system development capability training (Grant No: acjyyb2022021).

References

[1] Liu Wenfei, Zhao Mingwei. The Python teaching case practice for humanities and social sciences majors[J]. Computer Education, 2017, (12): 15-19.

[2] Wang Runhong, Liu Yan. Research on university computer public courses based on Python [J]. Computer Knowledge and Technology, 2019,15 (12): 155-156.

[3] Song Tian, Huang Tianyu, Li Xin. Python Language: the ideal choice for programming curriculum teaching reform [J]. University Teaching in China, 2016 (2): 42-47.

[4] Xu Weike. Feasibility analysis of the Python program design course for liberal arts majors [J]. Science, Education and Culture, 2018, (33): 59-61,68.

[5] Wu Ling, Wang Hao, Zhou Jian, et al. The Construction of Data Analysis Curriculum Group of the University of Finance and Economics -- A Case Study of New Economic Management Strategy of Anhui University of Finance and Economics [J]. Journal of Langfang Normal University (Natural Science Edition), 2020,20 (1): 113-116.

[6] Zheng Xiujie. Task-driven non-computer professional Python teaching practice [J]. Computer Knowledge and Technology, 2021,17 (02): 100-101 + 108.