Visual Analysis of Computer Hot Technology Based on CSDN

Roukai Huang[1], Shuyan Yu[2]

[1]648870158@qq.com, [2]ysy@usx.edu.cn

Yuanpei College, Shaoxing University, Shaoxing, China

Abstract—With the progress of information technology, there are more and more data on the Internet. How to make good use of these massive data and dig out useful values has become a topic that many researchers are studying. In this paper, massive data of CSDN Q&A are collected, and Python's powerful third-party libraries are used to visually analyze hot technologies, their development trends and technical categories. Scikit-Learn library is used for machine learning, and random forest regression algorithm is used to train hot technology prediction modules, optimize parameters and make predictions. Based on the research above, this paper finds that the computer technology will be further common, and the skills that programmers need will also increase.

Keywords—Hot Technology; Visual Analysis; Machine Learning; CSDN

1. INTRODUCTION

1.1 Foreword

With the continuous expansion and deepening of the computer application field, more and more people choose to enter the big field of computer, their level is also different, and the problems they encounter are also very different. Technical Q&A websites such as CSDN have been born gradually. People tend to look up Q&As through technical forums or Q&A sites or post questions themselves, etc. The CSDN platform in China provides convenience for relevant personnel in the computer field, and there is the StackOverflow platform in foreign countries. When a large number of users pour into CSDN, massive data will be generated. By studying these data through data visualization, data mining and other means, you will find the laws hidden behind the data, gain value beyond the data itself, and then understand the current state of the computer field, to meet the needs of its own development in the computer field.

1.2 Background and current situation at home and abroad

In the past few years, information technology has continuously penetrated and brought forth new innovations in various fields such as society, economy, and life. With the support of a series of emerging technologies such as mobile computing, Internet of Things, and cloud computing, more and more data have been generated. What is behind these data and how to make good use of these data has become one of the key directions for people to study. Therefore, there are many visual analysis studies based on big data. For example, the analysis of Hot Topics of College Online Public Opinion Based on Baidu Tieba, take Baidu Tieba of a university as the data source for hot

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334604

topic analysis, and put forward relevant guidance strategies to promote the all-round development of college students and promote the construction of campus harmony[1]. However, its data source is too narrow, only the data collected from a certain university does not cover most universities, and the conclusions drawn are only limited to this university. Another example is the Research on the Prediction of the Scale of Online Hot Topics. The selected data is too far from the time, which has a certain impact on the accuracy of the results[2]. As for the research on visual analysis of public traffic smart card, Xia Ting and others designed and implemented a multi-view fusion visual analysis system to solve the problem that the existing analysis methods are difficult to carry out multi task analysis in different temporal and spatial distribution and regional passenger flow time series change[3]. Regarding the visual analysis and research of MOOC, Tian Shenghu et al. conducted related explorations and put forward some development suggestions through the bibliometric and knowledge map analysis of more than 7,600 related articles included in CNKI[4]. But there is less research on technical spots in the computer field.

In foreign countries, StackOverflow is a well-known computer-related IT technology question and answer website. Ahmad Diyanati et al. proposed a method to determine the level of expertise of StackOverflow programmers based on user comment mining[5]. Programmers with high professional levels will ask higher-level questions or solve higher-quality problems.

1.3 Technology selection

There are many different implementations in the community for data collection. For example, use the common requests library to send network requests, use xpath to parse the web page, and finally use the csv library to save the data. In this article, the scrapy library is used for data collection, because it integrates many functions and has a powerful and general project template, which can help to quickly develop a data collection program[6]. The re library is used to perform string fuzzy matching to extract the required information. It is a built-in library of Python and is implemented based on regular expressions. Compared with other information extraction methods, it is faster and more efficient.

For data visualization analysis, this paper uses numpy, pandas, matplotlib and seaborn libraries for implementation. Numpy is a library composed of multi-dimensional array objects and a collection of routines for processing arrays. It can be used to process high-dimensional data conveniently and quickly. Pandas is an open source Python library based on numpy and is widely used for rapid analysis of data, as well as data cleaning and preparation. Matplotlib is the most basic, excellent and commonly used visualization library in Python. It has a set of drawing interfaces imitating MATLAB's function form[6][7]. Seaborn is an advanced visualization library based on matplotlib. Compared with matplotlib, it is simple and convenient to use, and the drawn graphs are more beautiful[8].

For hotspot technology prediction, this paper uses sklearn library to implement, all of its object interfaces are simple, and the commonly used machine learning methods are encapsulated, including regression, dimensionality reduction, classification, clustering and other methods.

Technically speaking, it is feasible to use Python data collection technology to obtain CSDN data, and then use Python data analysis technology to analyze the content obtained by CSDN question and answer to further study the hot technologies in the CSDN computer field. It is also feasible to study this topic from the perspective of CSDN information, because most of its users are

discussing related technologies in the computer field, and hot technologies can be analyzed from them. Every CSDN registered user can become the issuer and answerer of the question.

To sum up, for the data analysis of CSDN Q&A, after learning from different current mainstream technologies, this paper determines to use the scrapy framework and re library to analyze, collect and save CSDN Q&A data, and use numpy, pandas, matplotlib and seaborn libraries to visually analyze the hot technologies in the computer field from three aspects: hot technologies, the development trend of hot technologies and the technologies that people of different technical categories need to master. Finally, it predicts the hot technologies in the computer field through sklearn library.

2. ANALYTICAL MODEL BUILDING

2.1 Construction of data analysis model of hot technology in computer field based on Python

Firstly, determine the collected data, and obtain all the ID, title, content, reward C currency, reward amount, release time, number of answers, number of visits, questioner ID, whether to close the question, whether to adopt, label information, category information and other information of CSDN Q&A by the end of 2021 on the basis of web crawler based on CSDN. Then use the data cleaning method to preprocess the data, remove duplicate values, outliers, missing values, etc., eliminate irrelevant categories, and then call DataFrame.to_ csv (filepath, index=False) saves the obtained data locally in the form of csv file. Next, the preprocessed data is analyzed by Python data analysis technology, and the visual display of data analysis is carried out by drawing bar chart, line chart and wordcloud chart. According to the data analysis, a series of results of visual analysis of technology hotspots in the computer field based on CSDN are obtained, and the research conclusion is further drawn. Finally, the machine learning related algorithms are used to predict the hot technologies in the computer field, and the relevant prediction conclusions are drawn.

2.2 Key module process construction

Data collection module, data preprocessing module, data visualization analysis module and machine learning prediction module are the key links in the construction of data analysis model of technical hotspots in the computer field. The following will analyze the construction process of each key module.

2.2.1 data collection module

The construction of the data acquisition module uses the Python scrapy library and the re library to obtain and parse the CSDN Q&A webpage. The target of the crawler is to obtain all the ID, title, content, reward C coin, reward amount, release time, number of answers, number of visits, questioner ID, whether to close, whether to adopt, label information, category information and other information of CSDN Q&A by the end of 2021.

Firstly, the browser developer tool is used to capture packets online and analyze which request the question data list is. After finding the corresponding request, it is found that the request has encryption parameters, which is a string obtained by some encryption algorithm on other

parameters. After crawling for a certain number of times, it is found that the data obtained through the API interface is not enough to support this data analysis. Therefore, the next method is adopted.

Click a question to enter the question details page and analyze its web source code. It is found that the fields required for this data analysis are in a script tag, which is a JavaScript object. It can be matched with regular expression through Python re module, and then converted into Python dictionary by JSON Library to analyze it and obtain the required fields. After observing the website, it is found that the last number is a string of numbers. After analysis, it is found that this number is the ID of the question. By observing the number behind the latest question website, the for loop can be used to traverse the ID from 1 to eight million, send a request, and obtain all question data. Lastly save the data obtained each time through the built-in module of the scrapy and write it into the CSV file. In order to improve the collection efficiency, eight million IDs are divided into four processes to collect at the same time, and four corresponding CSV files are generated.

2.2.2 Data preprocessing module

Merge the four files generated by data collection into one file. By calling os.listdir (path) obtains all file names in the data storage directory, and then through the for loop and pandas.read_csv(filepath) reads each CSV file into memory and adds it to a general list, then combines all DataFrame in the list into one through pandas.concat (objs), and finally calls to_ csv(filepath, index=False) saves the merged data as a CSV file.

Perform data preprocessing on the merged data. Call the info() and describe() methods to observe the overall situation of the data.

- The votesCount field is empty, so this field will be deleted.
- Comparing the sum of the status field with the total number of data, it is found that its value is all 1, so this field is deleted, because all 1 has no meaning.
- Assign an empty string to a record without a title.
- Assign '[]' to records without categoryList.
- Call drop_duplicates() on the ID field to remove the duplicate.
- Delete the record with negative createdAt field, because the field is the timestamp of the publishing question, and there should be no negative number

Call the apply() method to pass in the anonymous function and list derivation, extract the information in the tagInfo and category fields, and convert the dictionary list into a string (list). For example, extract and convert [{'tagId': 3, 'tagName': Android ',' categoryId ': 0,' type ': 0}] into ['android']. Call the to_datetime() method to convert the createdAt and updatedAt timestamps into datetime format and display them as formatted dates, such as 2012-09-05 19:40:10. Finally, by observing the category field, remove some irrelevant records and call to_csv(filepath, index=False) method saves the processed data locally in the form of CSV file

2.2.3 Data visualization analysis module

The construction of data visualization analysis module is to visualize through wordcloud library, and draw line chart and bar chart through seaborn library. The design of wordcloud diagram needs to use wordcloud library. Through multiple aggregation operations on the data, some data can be counted to meet the needs of some analysis topic. For example, count the number of times each user asks questions in each category, and study what technologies people need to master in a certain direction.

The design of line chart and bar chart needs to use the drawing tool seaborn. It is encapsulated on the basis of matplotlib. The drawing is more beautiful and it is more convenient to use. The new data is obtained by aggregating the data, and the barplot () and lineplot () methods of seaborn are called to draw the bar chart and line chart.

2.2.4 Machine learning module

The construction of the machine learning module mainly uses the RandomForestRegressor class in the sklearn library, and then divides the training set and test set through the train_test_split() method in the library. The grid search is carried out through the GridSearchCv() method, and the relevant parameters and their values are selected. A set of model parameters with high accuracy can be obtained through the grid search. Based on this model, the hot technologies in the computer field are predicted. Random forest is an integrated learning algorithm, which has been applied in many fields[9][10].

The data analysis model of hot technology in the computer field based on Python data analysis technology. Due to the simplicity of Python data analysis technology and powerful third-party library function, the model construction process is relatively clear and the implementation technology is relatively convenient. Using Python based data analysis technology can also complete the aggregation and statistics of data, which is convenient for visual operation. Moreover, because it has many third-party libraries for machine learning, such as scikit-learn, choosing one of them can make the model algorithm very concise, and it also has certain advantages in model operation efficiency.

3. DATA ANALYSIS OF TECHNICAL HOTSPOTS IN COMPUTER FIELD BASED ON CSDN

3.1 Experimental data

Taking CSDN Q&A as the data source, this paper collects the data from September 2012 to December 2021. The total number of questions is 794232, with 271323 questioners. The main collected fields are ID, title, content, reward C currency, reward amount, release time, number of answers, number of visits, questioner ID, whether to close the question, whether to adopt, label information and category information.

Data collection process: use Python scrapy, re and json libraries to obtain and parse CSDN Q&A web pages. Then, the collected fields are saved in the form of CSV files through the scrapy built-in components' items, pipelines and CSV library. Because the number is too large, it is divided

into four processes for collection. Because of the use of scrapy, each process is also multi-threaded, which can improve the collection efficiency.

3.2 Experimental results and analysis

Based on the experimental data, the questions in CSDN Q&A are visually analyzed.

Aggregate through the groupby(by) method of DataFrame object. The by parameter refers to the specific fields to aggregate. It can be a string, but only one field can be aggregated; It can also be a list, which can aggregate multiple fields. Then perform statistical operations through the agg(func) method of DataFrameGroupBy object, where func parameters can be a method, string, list and dictionary, and the method returns a DataFrame object. If a method is passed, it can implement user-defined statistics and carry out relevant operations according to the passed method; If a string or list is passed in, its value can be statistics related operations such as mean and sum, which can be applied to each column; If you only want to perform statistical operations on the specified columns, you can pass in a dictionary in the form of {'field name': 'statistical operations'}. Then you can use the reset_index() and rename() methods to operate on the results to make them more intuitive.

After obtaining the statistical results, you can draw the bar chart by seaborn's barplot(x, y) method, where x is the field name, and y is its specific value; draw the line chart by the lineplot(x, y) method; wordcloud graphs can be drawn through the WordCloud library; information such as abscissa name, ordinate name and title can be set through the methods such as xlabel(xlabel), ylabel(ylabel), title(title) in the pyplot class in matplotlib. Finally, the drawn figure can be saved locally by the savefig(fname) method.

3.2.1 Hot technology analysis

Seaborn is Python's drawing tool, and pandas is Python's library dedicated to data analysis. By aggregating the question creation year and technology category, count the total number of questions, views and answers of each technology category every year. Then filter out the records in recent 5 years, aggregate the technology categories, and count the total number of questions, views and answers, so as to obtain the total number of questions, views and answers of each technology category in recent 5 years. Then sort the total number of questions, select the top 10 to draw the histogram, and the results are shown in the figure below.

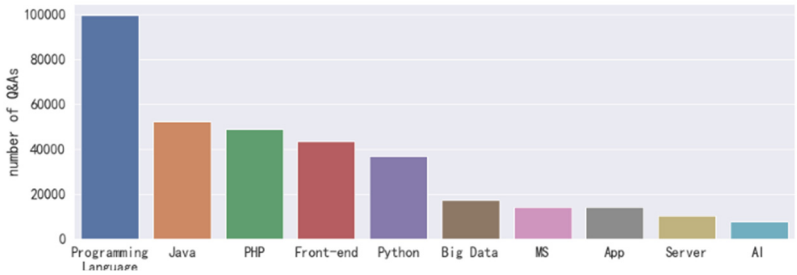

Fig. 1. histogram of total number of Q&As by technology category in 2017-2021

As can be seen from Fig. 1, there are about 100000 questions about programming language, which is more than twice that of Java, which ranks second. The main reason is that the

programming language is a large category. Strictly speaking, Java, PHP and python can be regarded as. Users do not deliberately choose the language they use when choosing the category, which may have an impact on the solution of the problem. The second is Java, PHP and front-end. The first two are mostly used by back-end developers, so we can know that the programmers who make products (front-end, back-end, web pages, mini program, etc.) still account for the majority in society. The last one is an emerging direction - artificial intelligence, mainly because it has only developed in recent years, and the threshold is also high.

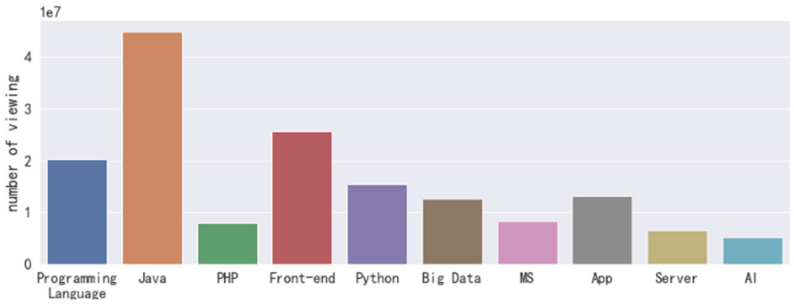

Fig. 2. histogram of total number of views by technology category in 2017-2021

As can be seen from Fig. 2, the number of views on Java is the highest, and there is a significant gap between PHP and it, indicating that most people have given up PHP and turned to Java. Of course, it may also be because of time. Many questions about PHP have been raised before.

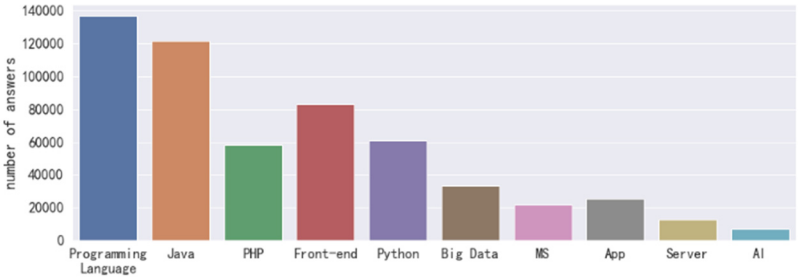

Fig. 3. histogram of total number of answers by technology category in 2017-2021

As can be seen from Fig. 3, the number of answers about programming languages is the highest, of course, in part because the number of questions is much higher than others. This is followed by Java, front-end and Python. The lowest is still artificial intelligence.

Combining the three pictures, it can be seen that the hot technologies in today's society are still front-end and back-end, and of course Python has also accounted for a part. The answer rate for questions about Java is the highest, which shows that people's discussions on Java are very intense, and it shows that the usage rate of Java is higher than others. This is followed by big data and front-end, illustrating how these technology categories are dominant in society. However, the answer rate of artificial intelligence is the lowest, indicating that there are fewer people engaged in the technology and there is more room for development.

3.2.2 Analysis of hot technology development trend

According to the statistical results obtained in the previous section, with the year as the abscissa, the total number of questions, views and answers as the ordinate, call the lineplot() method to draw a line graph. The development trend of hot technologies is shown in Fig. 4.

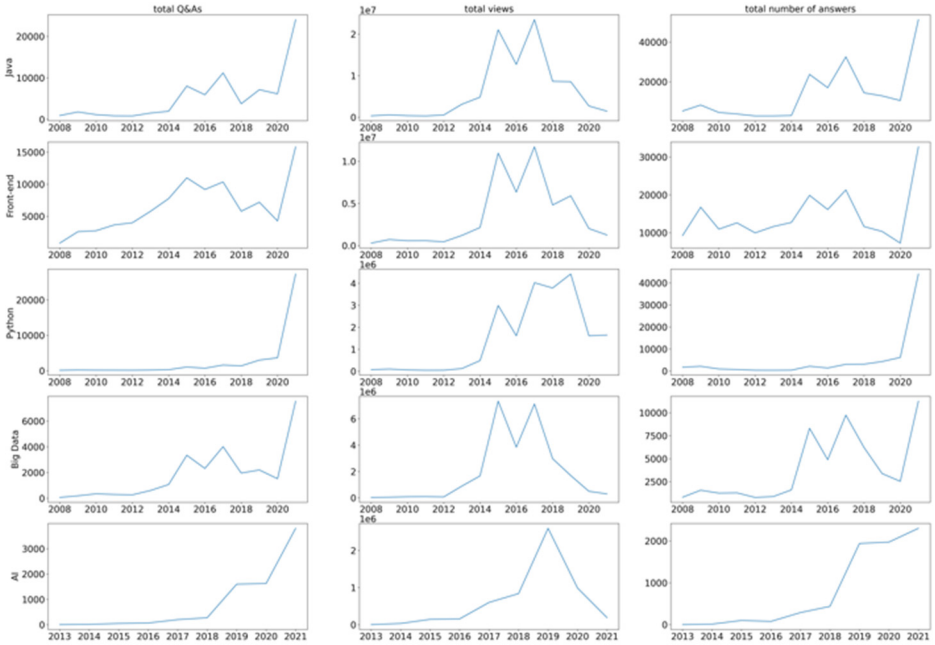

Fig. 4. a line chart of the development trend of some hot technologies

As can be seen from Fig. 4, except for Python and artificial intelligence, the two technologies that have become popular in recent years, all others have experienced the shock stage, but the total number of problems in 2021 has reached a record high. In particular, Python has increased by nearly 6 times, indicating that there are a lot of people learning Python, and the problems they encounter are also very rich. Compared with previous years, the total number of views is also far behind, mainly due to the fact that the questions were raised relatively recently, and there is no accumulation in previous years. The overall upward trend in the number of responses indicates that the discussion on these technology categories is still intense. To get a better understanding of the data, let's look at the average number of responses to each question each year. The result is shown in Fig. 5.

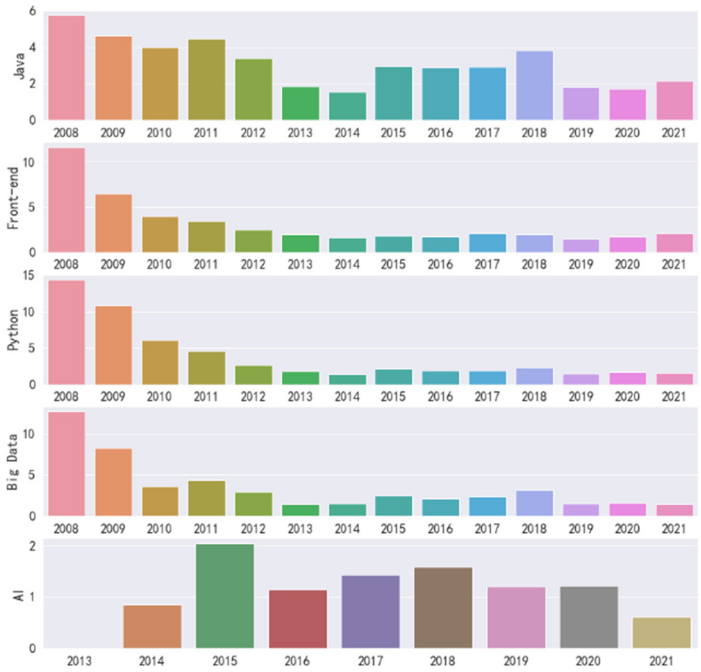

Fig. 5. histogram of the ratio of answers to questions per year

It can be seen from Fig. 5 that the ratio of the number of answers to the number of questions generally shows a downward trend, and the number of answers to the questions raised in 2008 is the highest. This is mainly due to the fact that at the beginning of the site's establishment, people at that time had not been exposed to such sites and were highly motivated. And the problems people encounter may not be well solved. After searching this Q&A, if they solve the problem, they may record it in this Q&A, which will let more people know and solve the problem. Of course, due to its large time span, it does not rule out future generations to answer and supplement the previous questions. In recent years, the average number of answers to each question is about 2. However, in the direction of artificial intelligence, there is a situation of less than 1 in 2021, which shows that there are still many problems that have not been answered and solved, and there is great room for its development.

3.2.3 Analysis of technologies to be mastered by personnel of different technical categories

Firstly, aggregate the users and technology categories, get the label of each user's question, and remove the duplicate. Then, based on the above operations, aggregate the technology categories, count the number of times each label appears, and call the wordcloud library to visualize the wordcloud. You can intuitively understand what technologies people in each technology category need, and the results are shown in the figure below.

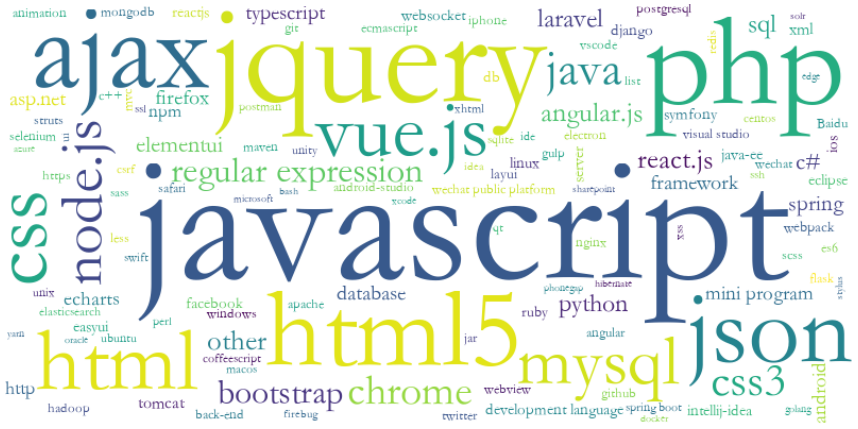

Fig. 6. wordcloud of technology used in front-end

As can be seen from Fig. 6, in addition to the three piece front-end set of HTML, JavaScript and CSS, you also need to master jQuery, which is a JavaScript library, which greatly simplifies JavaScript programming; ajax and json, the former is that the front-end sends requests to the back-end server, while the server usually returns data to the front-end in the form of the latter; Regular expression is also one of the required courses for front-end developers. It is often used in scenarios such as form verification. Of course, there are also frameworks such as vue.js, angular.js and react.js to help front-end developers develop easily and quickly.

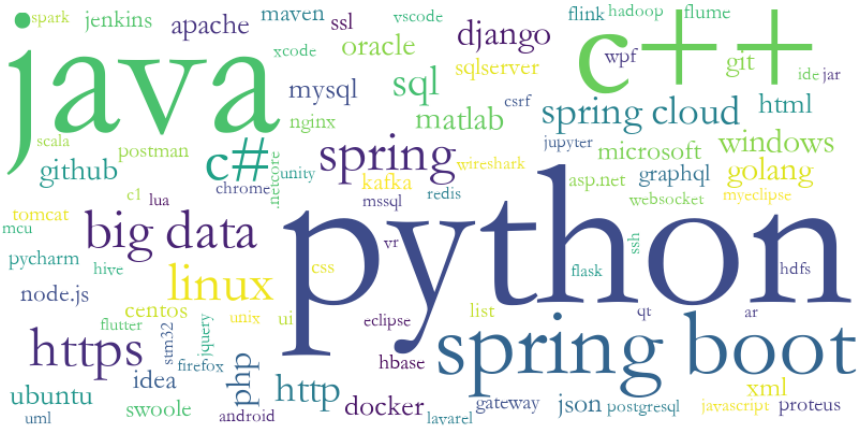

Fig. 7. wordcloud of technology used in back-end

As can be seen from Fig. 7, Python and Java are necessary to engage in back-end development. To use Java, you must be familiar with frameworks such as spring, springboot, and springcloud; Using python, people choose Django more. Compared with flask, another back-end Library of python, it is large and comprehensive, with a large number of common tools and components, and even with admin in the management background, which is suitable for rapid development of enterprise websites with perfect functions. For databases people usually use mysql.

Fig. 8. wordcloud of technology used in Python

As can be seen from Fig. 8, programmers engaged in Python development usually use django and flask libraries for back-end service development, and some are for data collection. Their discussion of the list data type in Python is also very intense, so it is very important to learn list well. Of course, Python is also used for data mining. Python has a powerful third-party library, which can easily perform data analysis, artificial intelligence, machine learning, neural networks, and deep learning related work. According to a specific direction you are interested in, you can study it in depth.

4. PREDICTION OF TECHNICAL HOTSPOTS IN COMPUTER FIELD BASED ON CSDN

4.1 Data preparation

4.1.1 Data preprocessing

Before training the model, the dataset must be preprocessed. Common ones are dealing with duplicate values, outliers, and missing values; formatting timestamps and converting them into the format of years, months, days, hours, minutes and seconds.

The related operations are similar to those mentioned above, deduplicate the id, remove the records less than 0 in the createdAt field, fill in the missing values for categoryList and title, and extract the year, month, day, hour, minute, and second fields from the createdAt timestamp, and set the dictionary list type of the category field is converted to a string.

4.1.2 Feature selection

Aggregate the data by category, year, and month, and count the total number of visits, likes, postings, total reward C coins, total reward amount, total answers, and number of completed questions for each category each year and month.

Select fields other than category and visitsCount as features, and select the fields of visitsCount as predicted labels. Three categories of Python, Java, and front-end are selected to conduct experiments on predicting relevant content.

4.2 Grid Search

A set of parameter values with better accuracy among the candidate parameters can be obtained by grid search. Firstly build a random forest regression model without any parameters, and then prepare a set of parameters:

- n_estimators: the number of tree models in the random forest, set it to range 50-200.
- criterion: the index used to measure the quality of branches, so it can choose "mse" mean square error or "mae" mean absolute error.
- max_depth: the maximum depth of the tree, the tree will continue to grow to no more than this depth, set it to a range of 1-10.

The search is performed by calling the fit() method. After completion, the best_params_ attribute can be used to view the optimal parameter value and the best_score_ attribute can be used to view the optimal accuracy under the optimal parameter value.

4.3 Predictions and conclusions

Use the optimal parameters obtained by grid search to perform random forest regression modeling, and pass the features and labels of the training set through the fit() method to train the model. Generate the data set to be predicted from January to December 2022, in which the total number of likes, the total number of posts, the total number of reward C coins, the total number of reward amounts, the total number of answers, and the number of completed questions are replaced by historical averages. The prediction results are shown in Fig. 9, Fig. 10 and Fig. 11.

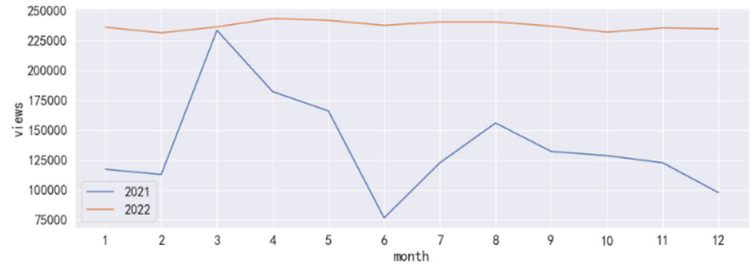

Fig. 9. line chart of Python's 2022 browsing forecast compared with 2021

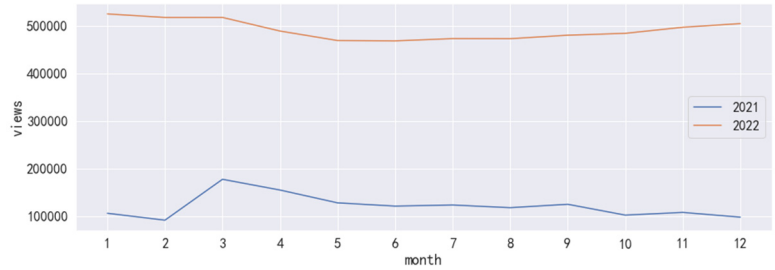

Fig. 10. line chart of Java's 2022 browsing forecast compared with 2021

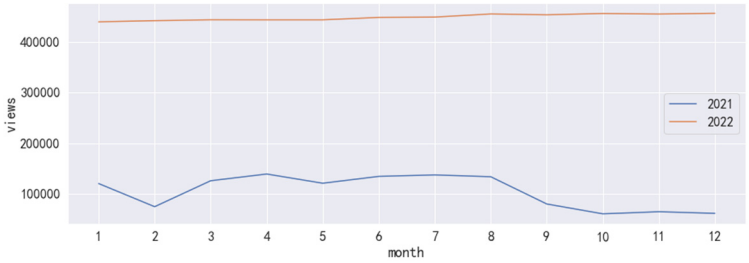

Fig. 11. line chart of Front-end's 2022 browsing forecast compared with 2021

From the forecast comparison results of the three line charts above, it can be seen that the number of views related to Python, Java and front-end issues in 2022 will be higher than in 2021. Among them, Java has the largest increase in views, exceeding last year's peak by nearly 200000. However, compared with the actual data of last year, the predicted results are too flat, there is no large increase or drop, and its accuracy has yet to be verified.

5. CONCLUDING REMARKS

Based on CSDN Q&A data, this paper uses scrapy framework in Python to efficiently send network requests and save data, in which re module is used to match and extract the required string information, and csv module saves the data as a local csv file. Then, numpy and pandas modules are used for data preprocessing and data analysis. Matplotlib and seaborn modules are used for data visualization. Hot technologies in the computer field are explored from three aspects: hot technologies, hot technology development trends and technologies that people of different technical categories need to master. At last, the RandomForestRegressor algorithm and GridSearchCV grid search in sklearn module are used for machine learning, and the page views of Python, Java and front-end quiz in 2022 are predicted.

The paper make visual analysis of technical hotspots in computer field from many aspects, so as to get reliable conclusions. However, there are still some shortcomings in the prediction of technical hotspots, and a satisfactory model has not been trained for prediction.

REFERENCES

[1] J. H. Li and X. Hu, “Analysis of hot topics of college online public opinion based on Baidu Tieba,” Computer and Modernization, no. 9, pp. 12-18, 2020.

[2] T. Y. Luo, “Research on the prediction of the scale of online hot topics,” Journal of Intelligence, vol. 35, no. 10, pp. 181-184+145, 2016.

[3] T. Xia, H. Niu, L. K. He, X. P. Fan and M. Zhu, “Visual analysis based on public traffic smart card data,” vol. 37, no. 6, pp. 1750-1754, 2020.

[4] S. H. Tian, J. F. Yao and T. Y. Cui, “Current situation, hotspots and development suggestions of MOOC research in china—visualization analysis based on knowledge mapping,” vol. 39, no. 1, pp. 25-30, 2019.

[5] A. Diyanati, B. S. Sheykhahmadloo, S. M. Fakhrahmad, M. Hadi Sadredini and M. H. Diyanati, “A proposed approach to determining expertise level of StackOverflow programmers based on mining of user comments,” Journal of Computer Languages, vol. 61, pp. 101000, 2020.
[6] X. Q. Zhang, “Analysis and implementation of web crawler based on scrapy framework”, Programming Skills & Maintenance, no. 2, pp. 18-19+44, 2022
[7] C. K. Li and Z. Q. Xiao, “Web page data analysis and visualization based on Python,” Information Recording Materials, vol. 22, no. 9, pp. 203-205, 2021.
[8] J. Q. Ma and G. W. Teng, “Application research of big data visualization based on matplotlib,” Computer Knowledge and Technology, vol. 15, no. 17, pp. 18-19, 2019.
[9] L. Zhao, G. Q. Sang, W. Wu, C. J. Liu and J. N. Wang, “Critical rainfall prediction model for mountain torrent disaster based on random forest regression algorithm,” Journal of University of Jinan (Science and Technology), no. 4, pp. 1-9, 2022.
[10] Y. R. Deng, X. D. Cheng, F. Tang and Y. Zhou, “The control of moldy risk during rice storage based on multivariate linear regression analysis and random forest algorithm,” Journal of University of Science and Technology of China, vol. 52, no. 1, pp. 47-54+72+55, 2022.