IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Review of Computer Engineering Studies
Vol. 6, No. 4, December, 2019, pp. 81-86
Journal homepage: http://iieta.org/journals/rces

Data Visualization Analysis of Big Data Recruitment Positions in Hangzhou Based on Python

Fangqin Ying[1*], Zhongyue Zhang[2]

[1] School of Dongfang, Zhejiang University of Finance and Economics, Haining 314408, China
[2] School of Mechanical Engineering, Zhejiang University, Hangzhou 310027, China

Corresponding Author Email: zjyfq@zufe.edu.cn

https://doi.org/10.18280/rces.060403

Received: 19 November 2019
Accepted: 26 December 2019

Keywords:
web crawler, recruitment, Python, bigdata, data visualization

ABSTRACT

In the DT era, data will become the main source of energy. The job of "big data" is so popular that it is in short supply for a long time. The purpose of this study was to obtain the big data recruitment information in the ocean of Web, Use Python web crawler technology as a tool to crawl at the recruitment website of https://www.lagou.com ,working place as "Hangzhou", the search criteria for the "big data", using the Pandas + Matplotlib to implement data cleansing, data analysis and data visualization in terms of the job seekers' concerns. The results obtained in this study include company size, education, working years, salary, skill requirements, development prospects, position and job benefits, etc. The results of data visualization provide reference for interested practitioners, but also provide colleges and universities with talent training direction.

1. INTRODUCTION

Mankind is moving from the IT (Information Technology) era to the DT (Data Technology) era [1], big data has been hailed as a new driving force for development in the 21st century. China has proposed to accelerate the implementation of the national big data strategy. In the case of continuously raising of the requirements of big data, however, talent for big data is in short supply, there are many online recruitment website, for instance, 51job.com [2], ChinaHR, Zhaopin.com [3], www.lagou.com, these websites have become the carrier of a large amount of recruitment information. In the ocean of Web, finding information is like finding a needle in the haystack [4]. The search engine is used to find information on the Web. Search engines can be of two types—crawler based and traditional search engines. In a traditional search engine, results are affected by human intervention, for example, users from different fields and backgrounds often have different search purposes and needs, while the results returned by traditional search engines often contain a large number of pages that users do not care about. On the other hand, crawler-based search engines do not have this problem. In order to solve such problems, crawler technology emerges as a result [5-8]. A Web Crawler, also known as a Web Spider [9], is a program that makes a request to a website, obtains resources, analyzes them and extracts useful data. Technically speaking, it is to simulate the behavior of the browser to request the site through the program, crawl the HTML code /JSON data/binary data (pictures, videos) returned by the site to the local area, and then extract the required data and store it for use [10]. www.lagou.com is a vertical recruitment platform of the internet industry in China, Hangzhou is known as a capital of the internet, the demand of "big data" is representative, so this study selects crawl www.lagou.com with the workplace for "Hangzhou", position name for "big data", use Python third-party libraries to develop, use web crawler technology to collect recruitment information rapidly and accurately. Obtain the information as follows: the name of the company, the company size, education, working years, salary, skill requirements, development prospects, position and job benefits, etc., and finally save the data into Excel spreadsheet for later data analysis and visualization display [11]. These visualizations are visual, the results can provide reference for the interested practitioners, and also can provide reference for the direction of talent training in colleges and universities, so as to produce a batch of talents more suitable for the society and workplace.

The remainder of this paper is organized as follows: Section 2 introduces about web crawler technology, Section 3 describes the architecture of the system, and the design and implementation of recruitment data crawler using Python are presented in Section 4. Section 5 describes data analysis and data visualization. The conclusions and future work are given in Section 6.

2. WEB CRAWLER TECHNOLOGY

A mere basic Web crawler is a function with a set of seed URLs as input and a set of crawled webpages as output. This simple function takes URL one by one, gets the webpage, and adds URLs found on this webpage to the list of URLs to be visited further. The architecture of simple crawler technology is shown in Figure 1, which is mainly composed of four aspects: Crawler scheduler, URL Management, Web downloader, Web parser [12].

(1) Crawler scheduler: the entrance of the program, mainly responsible for the control of the crawler.

(2) URL Management: Mainly responsible for extracting the URL address of the web crawler.

(3) Web page loader: fetching the contents of a web page based on the URL by the implementation of urllib2 and request.

(4) Web parser: extract valuable data from web pages.

By using web crawler technology based on python programs, the crawler mobilizes the crawler program to obtain the target data on the web page according to the information provided by the Crawler scheduler. Through the Web crawler technology to realize the extracting of the content on the web page, the architecture of the crawler technology is shown in Figure 1.

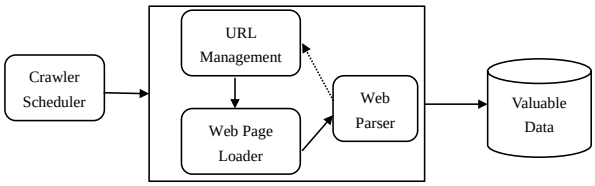

Figure 1. Architecture of crawler technology

3. SYSTEM OVERVIEW

The data in this article is obtained from www.lagou.com by web crawler, using Pandas+ Matplotlib for data scrubbing, data analysis and visualization. Pandas are the Python's core library for data analysis. Matplotlib is a graphics library written in Python language [13] that can easily draw various statistical graphs, such as scatter graphs, bar charts, line graphs, etc. The result of processing analysis needs to be made into visual graphics with obvious effects. This paper is using the version of Python3.6 and the overall architecture of the system is illustrated in Figure 2.

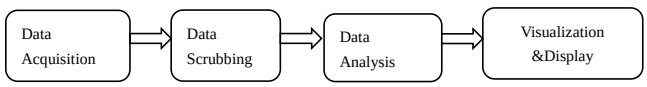

Figure 2. Architecture of the system

4. DESIGN AND IMPLEMENTATION OF RECRUITMENT DATA CRAWLER

The first thing the crawler needs to do is to get the link address to be crawled, then the position information in the link address gets parsed effectively and then is saved in the file of Excel form.

4.1 The crawling process of job information of www.lagou.com

(1) Define the destination page address to be crawled

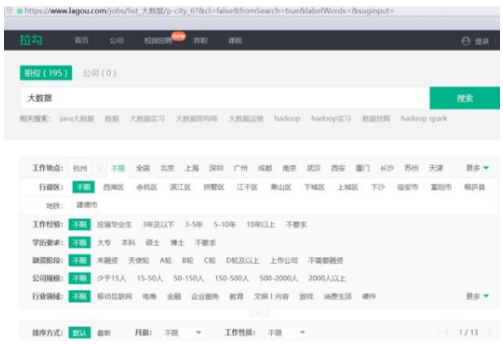

Figure 3. Home page of www.lagou.com

Here's how: The Chrome browser are used to open the browser and go to www.lagou.com home page, search for the position- "big data", workplace- "Hangzhou", as the Figure 3 shows, press F12 on the keyboard to enter developer mode, so we can see the HTML file contents of the page.

(2) View the positionajax.json entry using the XHR in the network Tabs, and define the header information in the web page request, including "Referer" and "user-agent". Figure 4 shows all request header information parameters.

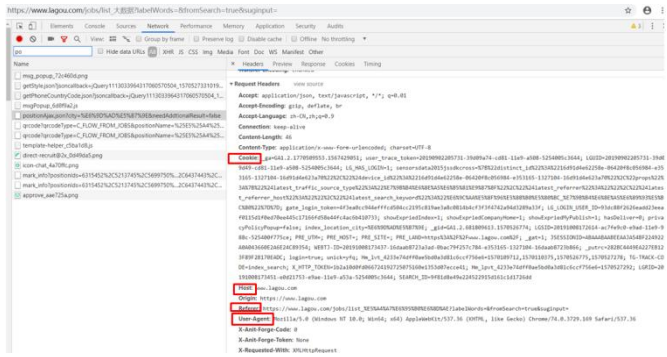

Figure 4. Information of headers

(3) Define web address parameter information

The above information as Referer, User-Agent should be added to the crawler Function to the Headers. The program is shown in Figure 5.

Figure 5. Web address parameter information

(4) Define the Form Data information used to query including the job title to be crawled. As shows in Figure 6 below.

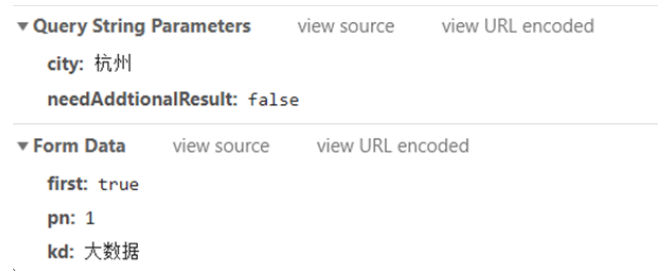

Figure 6. The information of Form Data

(5) Take the random time as the interval, get the cookie information [14], define the current crawl page in the Form Data, and whether the current is the first page, etc., call requests.post() to make a request to the www.lagou.com server.

(6) Each time a page result is fetched, the json() method of the result object is called to convert it to a dictionary for the object's type. Observe the JSON data, where the selected key names are 'companyShortName', 'companySize', 'education', 'workYear', 'positionName', 'positionAdvantage', 'skillLables', 'salary', 'financeStage', 'district'. Program code for crawling position information is shown in Figure 7.

The result after the crawler code runs is shown in Figure 8,

there are total of 13 pages and 183 positions, and the partial data saved by the crawler is shown in Figure 9.

```
# 将请求结果转换为字典
result = req_result.json()
# 提取职位相关的信息并输出
positions_info = result["content"]["positionResult"]["result"]
# for循环遍历职位列表并输出
# 定义一个字典保存每一页的职位信息
positions = {
        '职位名称': [],
        '公司名称': [],
        '融资阶段': [],
        '公司规模': [],
        '行政区': [],
        '工作年限': [],
        '学历': [],
        '薪资': [],
        '职位福利': [],
        '职位需求': []

        }
for i in range(len(positions_info)):
    positions["职位名称"].append(positions_info[i]["positionName"])
    positions["公司名称"].append(positions_info[i]["companyShortName"] )
    positions["融资阶段"].append(positions_info[i]["financeStage"])
    positions["公司规模"].append(positions_info[i]["companySize"])
    positions["行政区"].append(positions_info[i]["district"])
    positions["工作年限"].append(positions_info[i]["workYear"])
    positions["学历"].append(positions_info[i]["education"])
    positions["薪资"].append(positions_info[i]["salary"] )
    positions["职位福利"].append(positions_info[i]["positionAdvantage"] )
    positions["职位需求"].append(positions_info[i]["skillLables"] )
```

Figure 7. Program code for crawling position information

```
爬取第1页的数据
爬取第2页的数据
爬取第3页的数据
爬取第4页的数据
爬取第5页的数据
爬取第6页的数据
爬取第7页的数据
爬取第8页的数据
爬取第9页的数据
爬取第10页的数据
爬取第11页的数据
爬取第12页的数据
爬取第13页的数据
```

Figure 8. The results of Crawls program

	公司名称	公司规模	学历	工作年限	职位名称	职位福利	职位需求	薪资	融资阶段	行政区
0	字节跳动	2000人以	本科	不限	大数据(高	六险一金	[]	30k-60k	C轮	余杭区
1	微拍堂	150-500人	大专	3-5年	大数据开发	六险一金,	['Hadoop'	20k-40k	A轮	西湖区
2	飞猪	500-2000	本科	3-5年	数据开发	业务中台,	['数据挖	25k-45k	不需要融	余杭区
3	字节跳动	2000人以	本科	3-5年	大数据开发	六险一金,	[]	25k-50k	C轮	余杭区
4	Long Bric	50-150人	本科	3-5年	数据开发	五险一金,	['数据库	15k-25k	天使轮	西湖区
5	天宇科技	50-150人	大专	1-3年	大数据开发	待遇高	['数据挖	30k-60k	未融资	上城区
6	云真信	50-150人	本科	1-3年	大数据研发	海量数据	['Scala',	12k-24k	A轮	西湖区
7	杭州志卓	150-500人	本科	3-5年	java(平台	双休	['客户端'	10k-20k	不需要融	西湖区
8	同盾科技	500-2000	本科	3-5年	大数据研发	阿里技术(	[]	20k-30k	C轮	余杭区
9	个推	500-2000	本科	1-3年	大数据研发	16-18薪	[]	15k-30k	上市公司	西湖区
10	杭州志卓	150-500人	本科	1-3年	产品经理	双休	['用户研究	8k-15k	不需要融	西湖区
11	滴滴	2000人以	本科	3-5年	数据开发	技术大牛,	[]	25k-50k	不需要融	余杭区
12	溪鸟物流	150-500人	大专	不限	数据开发	福利待遇	['数据挖	15k-30k	不需要融	西湖区
13	有赞	2000人以	不限	1-3年	大数据开发	发展空间	['Hadoop'	20k-40k	上市公司	西湖区
14	有赞	2000人以	不限	3-5年	大数据开发	上市公司	['Scala',	20k-40k	上市公司	西湖区
15	微博	2000人以	本科	应届毕业	大数据研发	大厂平台	[]	13k-23k	上市公司	拱墅区
16	有赞	2000人以	本科	3-5年	大数据开发	上市公司	['Java',	20k-40k	上市公司	西湖区
17	单创	500-2000	本科	3-5年	大数据开发	前景可期	['算法',	15k-30k	未融资	江干区
18	涂鸦智能	500-2000	不限	3-5年	大数据开发	氛围好、	['Spark',	15k-30k	C轮	西湖区
19	邦盛科技	150-500人	本科	1-3年	大数据平台	六险一金,	['人工智	10k-20k	C轮	西湖区
20	涂鸦智能	500-2000	不限	3-5年	大数据实时	行业独角	['Flink',	15k-30k	C轮	西湖区
21	字节跳动	2000人以	本科	不限	大数据Lea	六险一金,	[]	30k-60k	C轮	余杭区
22	腾讯	2000人以	本科	5-10年	大数据AI(	大数据，A	['C++', '	20k-40k	上市公司	西湖区
23	微拍堂	150-500人	本科	3-5年	资深大数	双休；出	['数据处	20k-40k	A轮	西湖区
24	ZOOM	500-2000	硕士	3-5年	大数据开发	空间大	['Hadoop'	15k-25k	上市公司	西湖区
25	数澜科技	150-500人	本科	不限	大数据平台	大牛云集,	['Java',	15k-30k	A轮	余杭区
26	数美	150-500人	不限	应届毕业	大数据开发	快速发展,	['Spark',	15k-30k	C轮	余杭区
27	广通软件	500-2000	本科	3-5年	大数据开发	技术驱动,	['数据分	15k-30k	上市公司	西湖区
28	心理壹点	500-2000	本科	3-5年	大数据开发	亿级融资	['Flink',	15k-25k	A轮	滨江区
29	网易	2000人以	本科	5-10年	大数据开发	培训发展,	['ETL', '	35k-70k	上市公司	滨江区

Figure 9. Part of the results of the crawler data saved

4.2 Data cleaning

"Salary" processing: since the salary of the same position on the website is not a fixed value, but a range value [15], to facilitate mathematical analysis, a program is written to add two columns, that is maximum salary and minimum salary, the code is as follows:

```
import pandas as pd
data = pd.read_excel('data.xlsx')
data_xinzi=data["薪资"]
value_list_low=[]
value_list_high=[]
for value in data_xinzi:
    value_list=value.split('-')
    value_list_low.append(value_list[0])
    value_list_high.append(value_list[1])
data["薪资下限"]=value_list_low
data["薪资上限"]=value_list_high
```

5. DATA ANALYSIS AND DATA VISUALIZATION

In this part, the key data exported from the lagou.com are made into visual graphs by using the third party libraries of Python, such as Matplotlib [16]. Then import the pyplot plotting module from the Matplotlib library to draw various statistical graphs [17].

5.1 Salary description

Since the salary of position on the website is not a fixed value, but a range value. The mean value of the maximum and minimum salary is taken as the average salary of the position for mathematical analysis.

As can be seen from Figure 10, the mean of minimum monthly salary for big data is about 18,800 Yuan, the mean of maximum monthly salary is about 32,600 Yuan, and the average salary is about 25,700 Yuan. According to the published average urban salary in 2019, the average monthly salary in Hangzhou is 9,978 Yuan, and the average salary in big data is higher than the average salary in Hangzhou. It can also be seen from the histogram that the salary of big data positions distributes mainly in 15,000-20,000 Yuan, followed by 20,000-30,000 Yuan. Positions above 30,000 are dominated by backbone talents who can lead the team, such as big data development experts and directors, who are mainly engaged in big data architecture.

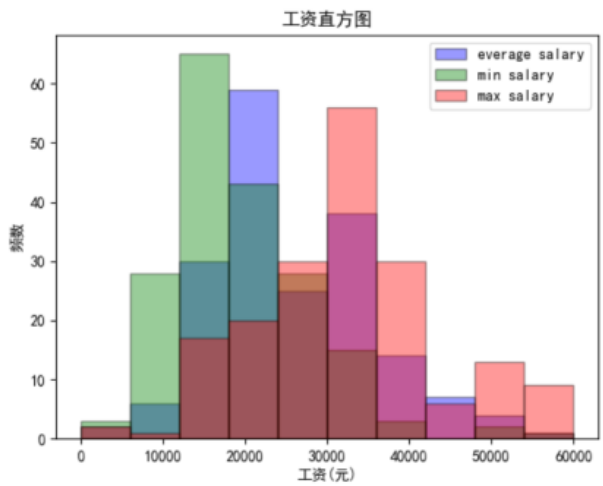

Figure 10. Salary histogram

5.2 Development prospects

Job seekers can judge a company's development prospects by its financing phase. As can be seen from Figure 11, angel wheel investments stage companies account for 1.3%, with high risk but large development and growth space. Series A investment companies account for 6.7% of the total, there have product prototypes and can be launched to the market in face of users, but the revenue is relatively low. Series B investment companies account for 18.7%, there are relatively mature with a relatively clear profit model. Series C investment companies account for 10.7%, their profits will grow rapidly and be basically mature after financing, listed companies and companies that do not need financing account for 49.3%, or nearly half of the total, which can maintain stable growth in earnings. It can be seen that some of the companies in demand for the post of big data are in the stage of vigorous development, including some unicorns and some rise steadily, such as NetEase, Ali and SERVYOU GROUP. Generally speaking, these companies boast great potential and promising prospects. Job seekers can choose start-ups or stable enterprises according to their own preferences.

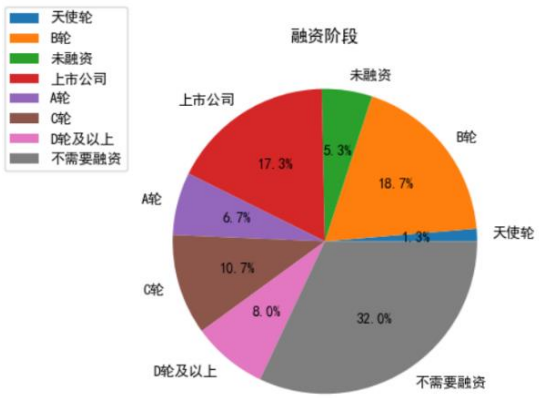

Figure 11. Company financing phase distribution

5.3 District distribution

As can be seen from Figure 12, big data positions in Hangzhou are mainly concentrated in Xihu district, Binjiang district and Yuhang district, which are the largest concentration places of Internet companies in Hangzhou.

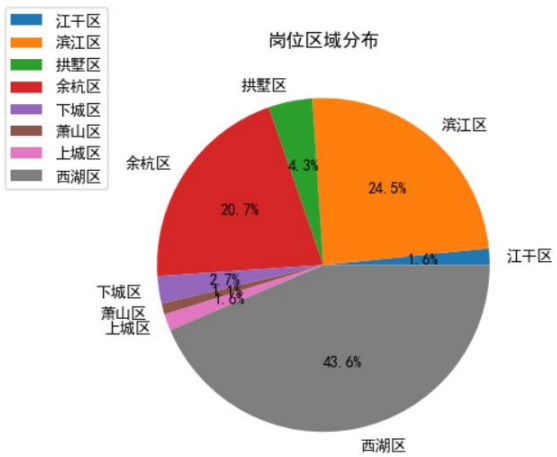

Figure 12. Position district distribution

5.4 Educational level distribution

Education required: bachelor's degree, master's degree, junior college degree and without limitation. As can be seen from Figure 13, the proportion of applicants with bachelor's degree is the highest, accounts for 77.7%. It may be that undergraduates have a good foundation in statistics, advanced mathematics and linear algebra; they have good algorithm and data analysis thinking. The second is junior college, accounting for nearly 9.0%. It can be seen that bachelor's degree and junior college degree can meet the vast majority of job requirements in the market, and the requirement for applicants for master's degree or above accounted for a relatively low number of enterprises, only 3.7%. Further analysis of the original data shows that the positions requiring a master's degree or above are mainly algorithm engineer and data mining that requires higher mathematics. Perhaps due to the Internet industry nature of the recruitment website, the education requirement of Ph.D are not shown here, such talents are more likely to be poached by means of headhunters, internal promotion, etc., rather than published on the recruitment website.

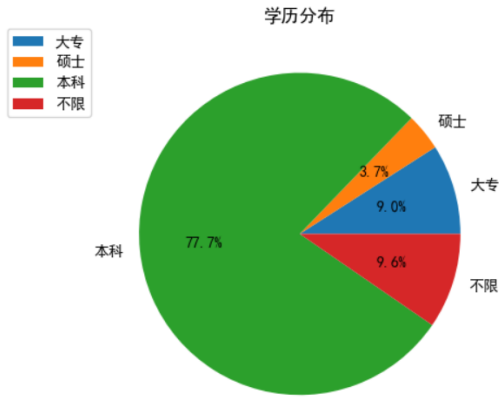

Figure 13. Educational level distribution

5.5 Working years

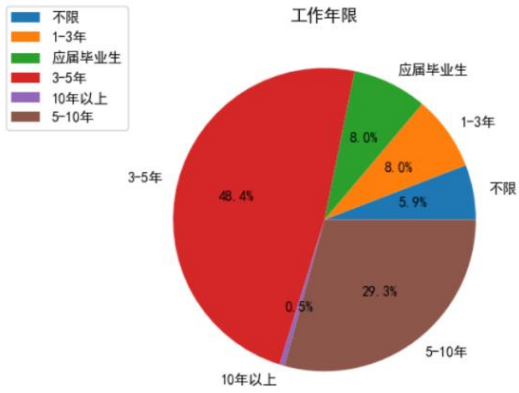

Figure 14. Working years

As can be seen from Figure 14, the positions requiring 3-5 year of work years are the most, accounting for 48.4%, while those requiring 5-10-year work years account for 29.3%. Working years of 1-3-year is required, accounting for 8.0%; the proportion of new graduates and those with unlimited

working years account for 8.0% and 5.9% respectively. It can be seen that the job market is in great demand for talents with certain working experience and ability to work independently or to leading a team. The demand for less than one year's working experience also indicates that this industry has a certain talent deficit.

5.6 Company's size

As can be seen from Figure 15, large companies with more than 2,000 employees account for 13.8%, companies with 500-2,000 employees account for 42.0% and companies with 15-500 employees account for 44.2%. It can be seen that recruitment companies are mainly small and medium-sized companies.

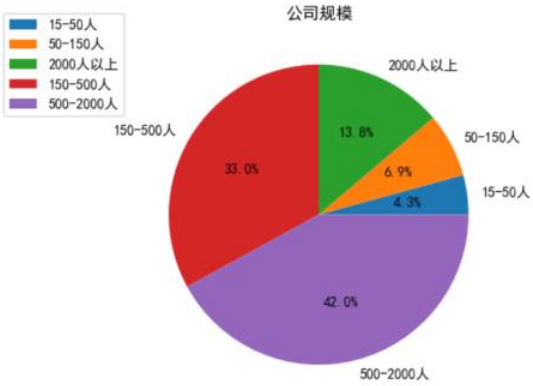

Figure 15. Company's size

5.7 Job benefits

Aside from interest and salary, job benefits are the greatest concern for job seekers. The main process of word cloud is as follows: Firstly, the "job benefits" column is aggregated to produce a string. Secondly, applying Python jieba to word segmentation [18] processing for job benefits; Finally, use the result of word segmentation to generate word clouds by using Python WordCloud [19]. The job benefits word cloud is shown as Figure 16:

Figure 16. Word cloud of job benefits

It can be seen that flexible work, working atmosphere, platform, paid vacation, team, double vacation, equity incentive and training appear relatively frequently in the job benefits. Job seekers can judge whether they can accept the job according to the benefits of the position, and at the same time, they can grasp what the position can bring to them.

5.8 Skill requirements

Only when you know the skill requirements of a position can you prepare for the job search in a specific way. The analysis ideas and process in chart 5.7 are also adopted to obtain the word cloud of "skill requirement", the word cloud is shown as Figure 17.

Figure 17. Word cloud of skill requirements

This shows that the main skills of big data are required to be proficient in Java technology and familiar with Spark, kafka, Hive, HBase, etc.; be familiar with Hadoop distributed system architecture, ETL, SQL database language, data warehouse, machine learning, etc. The company will give priority attention to these who have experience in data visualization, data analysis and mathematical modeling. Job seekers can prepare these relevant knowledge and skills so as to improve the fitness with occupational requirements and enhance their competitiveness. Colleges and universities can change talent training direction according to this.

5.9 Job title

As can be seen from Figure 18, the job titles of big data include data analyst, big data development engineer, operation and maintenance engineer, sales, big data test engineer, big data architect, big data research and development engineer, big data visualization designer, etc. Colleges and universities can determine the orientation of training talents according to this.

Figure 18. Word cloud of job title

6. CONCLUSIONS

This paper introduces the crawl process of web crawler to obtain recruitment information taking www.lagou.com for example, provide more comprehensive analysis on many aspects which job seekers care much for “big data” position, shows the operation process including data scrubbing, data analysis, data visualization, offer some reference for the majority of interested applicants, reduce the asymmetry of information, make search for jobs more efficiently and more closely match, this paper laid a certain foundation for future research work, the future work should be improved according to the following deviations.

Possible deviations:

1. The limitation of sample size may cause errors to the analysis results because of only 183 samples;
2. This paper crawl the recruitment information of www.langou.com only for the Internet industry, actually, other industries also have demands for big data talents, which may lead to incomplete information.

ACKNOWLEDGMENT

This paper is sponsored by Project No. 2018dfy006 of the key projects of School of Dongfang, Zhejiang University of Finance and Economics.

REFERENCES

[1] Ali Research Institute. (2015). Internet+ from IT to DT. Mechanical Industry Press, Beijing.

[2] Wang, T. (2018). Analysis and implementation of recruitment information for software technical personnel based on Python. Fujian Computer, 11: 118-119. https://doi.org/10.16707/j.cnki.fjpc.2018.11.058

[3] Liu, G.P., Liu, N., Duan, H.Y. (2018). Talent recruitment data collection based on focused web crawler technology. Computer Programming Skills & Maintenance, 5: 69-70.

[4] Kumar, M., Bhatia, R.., Rattan, D. (2017). A survey of Web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6): 1-8. https://doi.org/10.1002/widm.1218

[5] Wang, B.Y. (2017). Research on Python-based web crawler technology. Digital Technology and Application, 5: 76.

[6] Madhusudan, P.A., Lambhate, P.D. (2017). Deep web crawling efficiently using dynamic focused web crawler. International Research Journal of Engineering and Technology, 4(6): 3303-3306.

[7] OH, H.J., Won, D.H., Kim, C., Park, S.H., Kim, Y. (2018). Design and implementation of crawling algorithm to collect deep web information for web archiving. Data Technologies and Applications, 52(2): 266-277. https://doi.org/10.1108/DTA-07-2017-0053

[8] Kim, T.J., Kim, H.J. (2017). Machine learning-based topical web crawler: An ensemble approach incorporating meta-features. Journal of Engineering and Applied Sciences, 12(18): 4651-4656. https://doi.org/10.36478/jeasci.2017.4651.4656

[9] Guo, L.R. (2017). Programming of web crawler based on Python. Electronic Technology and Software Engineering, 23: 248-249.

[10] Xing, L., Wu, M.N. (2018). Design and application of recruitment theme crawler. Computer Knowledge and Technology, 14(25): 73-75.

[11] Wu, Y.C. (2019). Discussion on web data grabbing and analyzing with Python crawler technology. Computer Era, 8: 94-96. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2019.08.027

[12] Chang, F.J., Li, Z.H., Wen, J., Chang, F.J. (2019). The design and implementation of recruitment data crawler using Python. Software Guide, 7: 16-17. https://doi.org/10.11907/rjdk.191156.

[13] Huang, Q. (2019). Data visualization method and system implementation based on Python. China Computer & Communication, 14: 137-140.

[14] Li, P. (2019). Research on Python-based WebCrawler and anti-reptile technology. Computer & Digital Engineering, 47(6): 1415-1420. https://doi.org/10.3969/j.issn.1672-9722.2019.06.028.

[15] Jia, N.Y. (2019). Job data analysis based on python crawler-take www.lagou.com as an example. Information Technology and Informatization, 4: 64-66. https://doi.org/10.3969/j.issn.1672-9528.2019.04.018

[16] Li, J.H. (2018). Data analysis based on Python. Electronic Technology & Software Engineering, 17: 167.

[17] Wes MCKinney. Xu Jingyi. (2019). Python for Data Analytics: Data Wrangling with Pandas, NUMPy, and IPython, China Machine Press, Beijing.

[18] Zhu, Y.Z., Jing, J. (2019). Chinese word segmentation technology based on Python language. Communications Technology, 52(7): 1612-1619. https://doi.org/10.3969/j.issn.1002-0802.2019.07.012.

[19] Yan, M., Zheng, C.X. (2018). Word segmentation and word cloud production in Python environment. Modern Computer, 34: 86-89. https://doi.org/10.3969/j.issn.1007-1423.2018.34.021