Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 6, No. 2, 2022

Analysis of the Employability of Fresh Graduates in Double-carbon Industries Based on Web Crawler and Visualization Technology

Zehan Zhou[1], Haoyang Wu[2], Zhen Chen[3]

[1]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China
[2]School of Management Science and Engineering, Anhui University of Finance and Economics, Bengbu 233030, China
[3]School of Economics, Anhui University of Finance and Economics, Bengbu 233030, China

Abstract: Web crawlers are usually used to obtain a large amount of information, and combined with visualization technology, the obtained information can be expressed more intuitively. The dual-carbon industry has developed rapidly since the release of the dual-carbon policy, but the fresh graduates do not know enough about their employment information. Based on this, this article uses python to crawl the dual-carbon job descriptions on the job-hunting network for fresh graduates, and visualize them. Finally, make reasonable suggestions for the school and the government.

Keywords: Web crawler, Double carbon, College student employment, Data visualization analysis.

1. Introduction

Following President Xi Jinping's announcement at the 75th United Nations General Assembly that China strives to peak carbon dioxide emissions by 2030 and strive to achieve the goal of carbon neutrality by 2060, the two-carbon-related fields have sprung up rapidly, greatly promoting the With the expansion of talent demand in the dual-carbon field, a large number of related jobs were born. Carbon peaking and carbon neutrality are closely related to each of us. College students and people from all walks of life all over the country are paying more and more attention to the dual-carbon field. At the same time, the number of graduating college students across the country is increasing year by year, the pressure of college graduates and employment is high, the employment situation is grim, and the situation of difficult employment and slow employment has always existed. At this time, there is a large talent gap in the emerging dual-carbon fields, and a large number of relevant talents are urgently needed.

In this regard, in order to help college graduates to clearly obtain relevant information about dual-carbon positions, do a good job in their precise positioning, clarify their own ability requirements, and better choose relevant positions in the dual-carbon field, this article will take the job-hunting network for fresh graduates as an example, through python Reptile conducts in-depth research on information about jobs in the dual-carbon field to help college graduates easily find relevant jobs in the dual-carbon field.

2. Reptile Design Process

2.1. Demand analysis

The purpose of crawlers is to extract structured and large amounts of data from websites and store these data permanently on computer hard drives. This paper selects MySQL database to save structured data, uses this database to clean the data, then exports the processed data to Excel, and finally uses the wordcloud visualization library and Tableau drawing software to display and analyze the data. [1]

2.2. Crawler Modular Design

Modular crawler programs have high scalability, so they can adapt to different application scenarios. By combining different modules, a complete web crawler system can be constructed, and modular programs are more convenient for testing and post-maintenance.

Web scraping module analysis:

The essence of a crawler program is to simulate a client requesting web page information, and crawl the required data information from the web page source code returned to the web page. This article is divided into web scraping module, web page source code parsing module, data storage module, these three modules

The blocks are combined into a complete web crawler system.

2.2.1. Web Scraping Module

First build the first level url list. For multiple crawling of the same IP in a short period of time, the IP address will be blocked by the website, so the technology of proxy IP pool is used here to access. In order to avoid being discovered by the other party, it is also necessary to join User-Agent to disguise itself as a proxy server. By constructing a proxy IP pool and a proxy pool composed of many user agents, the combination of access IP and user generation is randomly selected each time, so as to disguise itself as a user access from different IPs, which greatly reduces the probability of being anti-crawlers. In the "Job Network for Fresh Graduates", by analyzing the url addresses of 100 pages with "carbon" as the key word, it can be found that the end of the url address of each page differs by 10, that is, the initial url address code is designed.[2]

```
# 初始化url列表
url_list = ['https://s.yingjiesheng.com/search.php?word=%E7%A2%B3&sort=score&start={}'.format(i) for i in
            range(0, 10, 10)]
```

2.2.2. Web Page Source Code Parsing Module

Considering that there will be asymmetric crawling information and 50 types of errors, the if nested structure is used here to prevent asymmetric crawling information and set timed sleep to prevent the server from crashing due to high-intensity crawling.[3]

```python
        if len(titles) == len(url_deeps) == len(info_datas) == len(info_sources):
            num = 1
            for title, url_deep, info_data, info_source in zip(titles, url_deeps, info_datas, info_sources):
                info_list = []
                response_deep = requests.get(url_deep)
                if response_deep.status_code == 200:
                    response_deep.encoding = 'GBK'
                    try:...
                    except BaseException as e:
                        print(f"二级页面解析错误, 报错为: {e}, 页面地址为: {url_deep}")
                else:
                    print(f"二级页面请求错误, 状态码为: {response.status_code}")
                # sleep with random time
                time.sleep(random.randint(1, 5) + random.random())

        else:
            print("一级页面的解析过程中发生字段信息不对等情况")


    except BaseException as e:
        print(f"一级页面解析错误, 报错为: {e}")
else:
    print(f"一级页面请求错误, 状态码为: {response.status_code}")
# sleep with random time
time.sleep(random.randint(1, 5) + random.random())
page += 1
```

Then a double-layer traversal is performed. After extracting the source code of the first-layer URL, analyze the current text to find the key information that the user needs. According to the user's needs, it is also necessary to know the name of each type of work and the corresponding web page link. The information required by the user exists under a-href. All job names are stored in the title list, and all job links are constructed into complete URLs and stored in the url_deep list corresponding to the title list. In the same way, information such as date can be obtained.

```python
# create html_object and soup_object for parse html
soup = BeautifulSoup(response.text, 'lxml')
html_obj = etree.HTML(response.text)

# parse the first html[title,data,info_source,url_deep]

# get title info
titles = soup.find_all('h3', class_='title')
# get url_deep
url_deeps = html_obj.xpath('//*[@id="container"]/div[1]/ul/li/div/h3/a/@href')
# get the source of info with title
info_sources = html_obj.xpath('//*[@id="container"]/div[1]/ul/li/div/p/text()')
# get info data
info_datas = html_obj.xpath('//*[@id="container"]/div[1]/ul/li/div/p/span/text()')
```

Next, we use the requests library to parse the current web page, and we can also use the proxy IP pool and the user agent pool to randomly select and match the information to crawl information more smoothly. According to the nature of the dual carbon post. Use the Beautifulsoup library to extract the information from the parsed web page. At this time, the following design code will be used[4]

```python
soup_deep = BeautifulSoup(response_deep.text, 'lxml')
work_place = soup_deep.find_all('span', class_='i')[1].get_text()
job_needs = soup_deep.find_all('div', id="wordDiv")
```

2.2.3. Data Storage Module

The information obtained from all the double-carbon positions on the graduate job search website is relatively large data, so next, save the data as a csv file, use the jieba library to segment the text, and then use the Counter method to perform word segmentation. Word frequency statistics, and finally remove stop words, leaving effective words to improve analysis efficiency and analysis results. It lays the foundation for the following job data analysis.

```python
import csv


def write_csv(Filename, info_list):
    times = 1
    # csv_title
    csv_title = ["标题", "发布时间", "发布来源", "工作地点", "招聘要求", "招聘要求页面链接"]
    with open(Filename, 'a+', encoding='utf-8') as f:
        csv_obj = csv.writer(f)
        if times == 1:
            csv_obj.writerow(csv_title)
            csv_obj.writerow(info_list)
        else:
            csv_obj.writerow(info_list)

    times += 1
```

3. Data Visualization Analysis

In the above, the required data is imported into the csv file and excel file through the designed crawler program, and then the jieba library is used for data cleaning. In the current time period, the Graduate Job Search Network has released a total of 1,000 Java job information. The following will use the Tableau drawing software and the Wordcloud library to perform a visual analysis of the data to obtain the Graduate Job Search website's data on dual carbon jobs.

The data analysis results are as follows:

With the continuous growth of the number of fresh graduates, the employment pressure of the whole society is also increasing. At the same time, the rise of jobs in the dual-carbon field has led to a large shortage of talents in dual-carbon positions and the pressure of the employment situation of college students. , more and more college students and social people are paying more and more attention to jobs and employment opportunities in the dual carbon field.

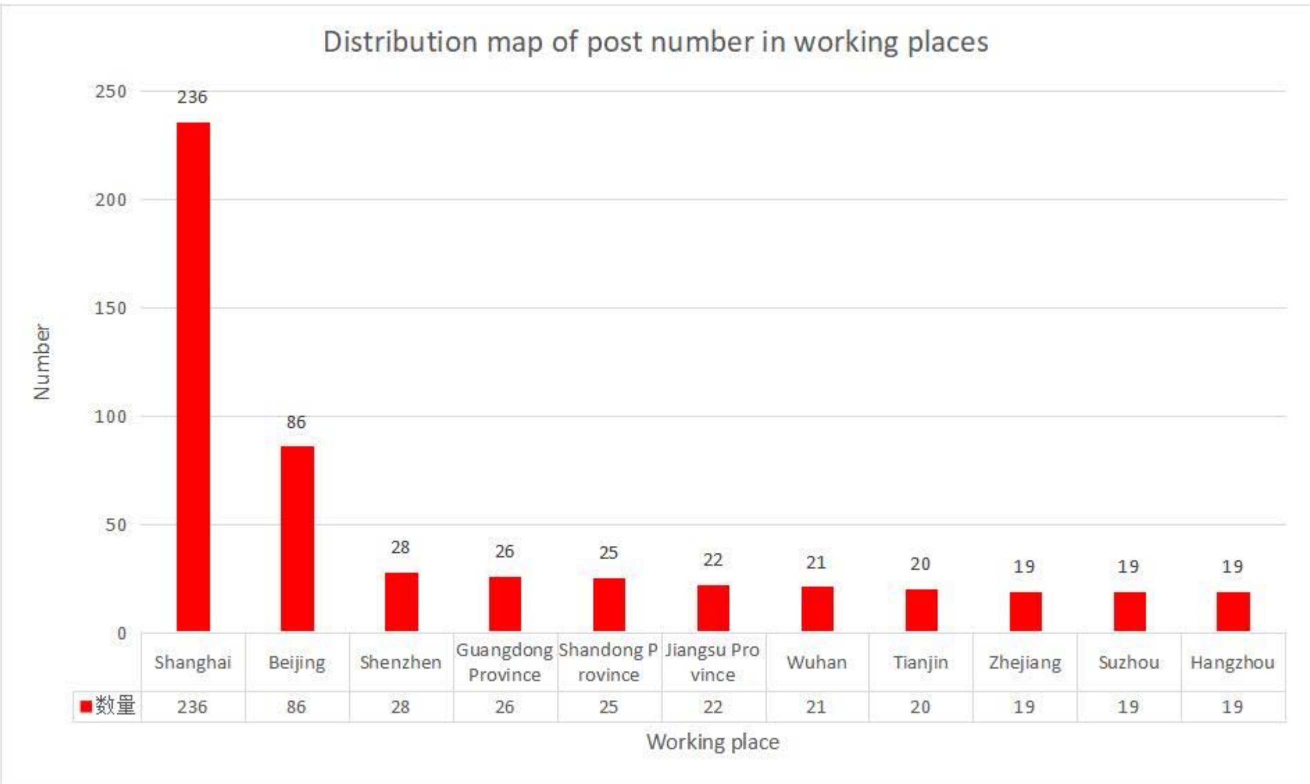

Figure 1. Region - Distribution of Number of Positions (Part)

After analyzing the data extracted by the research institute, the top cities and provinces in terms of the number of carbon-related jobs provided across the country are shown in Figure 1.

As can be seen from Figure 1, the distribution of the two-carbon jobs is relatively broad. The distribution locations are more diverse and random. From a specific analysis, the distribution of the number of positions has the following characteristics:

(1) The quantity gap is large. As can be seen from the figure, the number of jobs provided by the two cities of Shanghai and Beijing is 236 and 86, respectively, far ahead of other cities.

Especially in Shanghai, the number of jobs related to dual carbon provided is much higher than that provided by other cities in China. In addition to the large gap in the number of jobs in the top rankings, the number of jobs in the latter cities is relatively close.

(2) Radiation characteristics. The cities with the highest number of dual-carbon jobs are evident in the Yangtze River Delta region with Shanghai as the core. From this, it can be seen that the dual carbon industry also has the characteristics of developed cities radiating around.

(3) Uneven distribution. The dual-carbon jobs are mainly distributed in the central and eastern regions, and the jobs provided in the western region are lower than those in other regions of the country.

(4) Widely distributed. Although the spatial distribution of dual-carbon jobs is uneven, they are involved all over the country, which is reflected in a wide range, which further reflects the importance that my country attaches to the realization of the "dual-carbon" goal.

(5) There is a strong correlation with the degree of economic development. Jobs are mainly distributed in economically developed places such as first-tier cities and new first-tier cities, such as traditional Beijing, Shanghai, Guangzhou and Shenzhen, and new first-tier cities, Hangzhou, Wuhan, etc. These cities provide a large number of jobs related to dual carbon. It can be seen that the distribution characteristics of the number of dual-carbon jobs are related to the degree of economic development.

Our team further digs the data through certain methods, and counts the job names provided by various companies related to the double carbon field. It is found that the names of the double carbon related positions are not uniform. Figure 2 below is a statistical display of some job titles.

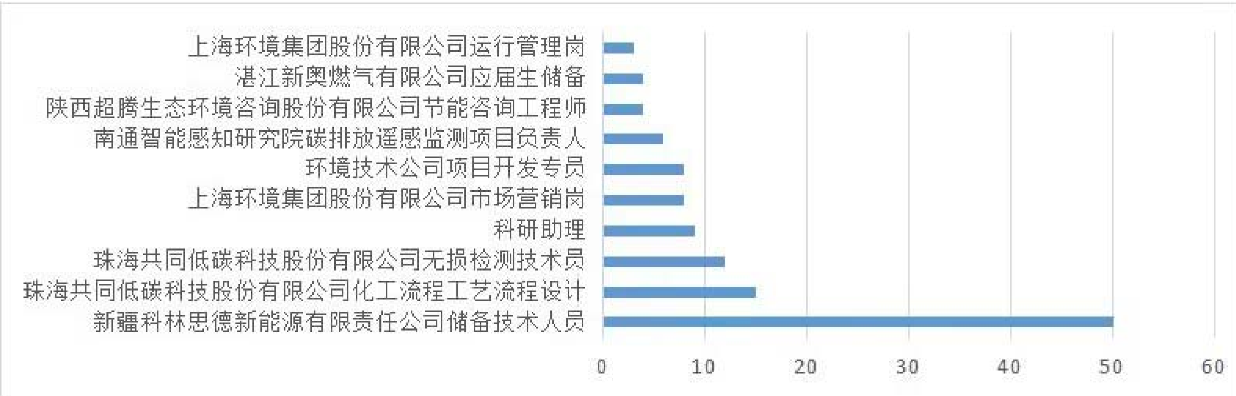

Figure 2. Statistical chart of word frequency of job title

After specific analysis, we can see the following characteristics:

(1) The number of positions is in a similar state. It can be seen intuitively from the figure that, except for some enterprises, the number of positions provided by most enterprises does not exceed 20;

(2) Diversification of job titles in the dual-carbon field. Various regions and enterprises do not have unified job titles in the dual-carbon field, but various ones, each with its own naming method, reflecting various and complex characteristics;

(3) Although there are various job titles in the dual-carbon field, there is still a certain tacit understanding for the naming of certain jobs. It can be seen that although the company does not have a unified job title, the company prefers words such as "engineer" and "technical personnel". Words such as "Company Reserve Technician" and "R&D Engineer" in the picture can verify this well. It can be seen that each company still has a certain tacit understanding of the naming of job titles;

(4) The industries involved in the jobs in the dual carbon field are very extensive. For example, some are engaged in scientific research and development, some are engaged in product marketing, some are engaged in product research and development, and some are engaged in carbon management and so on. Therefore, positions in the dual carbon field are not only engaged in environmental protection related work, such as finance, product sales, research and development, scientific research and other aspects.

In addition to regular data analysis, we use the third-party WordCloud library to draw word cloud graphs using the frequency of words in the text as a weight. By configuring the corresponding parameters, loading the word cloud text information, and outputting the word cloud file, we obtained the cloud map of the job analysis words and job demand words, so that the data can be analyzed intuitively. The analysis results are as follows.

Figure 3. Job analysis word cloud

Through the analysis of Figure 3, it is not difficult to find that low-carbon-related jobs have a certain diversity and randomness in regional and industrial distribution. However, from the overall analysis of the cloud map, the distribution of low-carbon jobs has the following characteristics:

(1) It has the characteristics of a wide range, and low-carbon jobs have a certain distribution in various regions of the country. For example, Shanghai and Hangzhou in the Yangtze River Delta, Sichuan and Tibet in the southwest, Xinjiang and Qinghai in the northwest, Heilongjiang in the northeast, Guangdong and Guangxi in the southeast, etc., show characteristics that are all over the country in terms of scope. To a certain extent, this reflects the importance attached by governments and enterprises across my country to energy conservation and emission reduction, as well as the widespread popularity of low-carbon jobs.

(2) Concentrated in large and medium cities, especially those dominated by high-tech enterprises. The high frequency of cities such as Shanghai, Hangzhou, and Zhuhai in the word cloud map indicates that large and medium-sized cities have a more significant demand for low-carbon emission reduction.

(3) Relying on the existence of high-tech enterprises and institutions of higher learning. The high frequency of local high-tech enterprises and local colleges and universities in the word cloud map indicates that low-carbon jobs are generally opened by high-tech enterprises and are closely related to colleges and universities.

Figure 4. Word cloud map of job requirements

Through the analysis of the job demand word cloud map, it is not difficult to find that the job demand for low-carbon jobs has the following characteristics:

(1) Higher education requirements for job seekers, with a bachelor’s degree or above. In the cloud map, the keywords of undergraduate, undergraduate and above, and doctorate appear most frequently, and there are relatively few jobs for junior college students.

Through the analysis of the job demand word cloud map, it is not difficult to find that the job demand for low-carbon jobs has the following characteristics:

(2) Higher education requirements for job seekers, with a bachelor’s degree or above. In the cloud map, the keywords of undergraduate, undergraduate and above, and doctorate appear most frequently, and there are relatively few jobs for junior college students.

(3) The majors are dominated by traditional engineering, but there are also liberal arts majors such as mathematics and accounting. Covering electronics, machinery, chemicals, materials, computers, automation, economics, management, law and other fields.

(4) There are higher requirements for personal physical

fitness and willpower. Key words such as hard work, strong sense of responsibility, and good health appear in the cloud map with high frequency, reflecting from the side that some low-carbon jobs have problems such as harsh working environment and high work intensity due to the pollution of factories and chemicals.

(5) The work content focuses on technical applications. Key words such as R&D, operation, testing, and scientific research appear prominently in the cloud map, which fully shows that low-carbon positions place higher requirements on the scientific research level and technical level of the incumbents.

4. Conclusion

In this paper, through the configuration of Python and MySQL Server, a web crawler data collection and analysis system based on the graduate job search website is created. The system can log in to the graduate job website and obtain page information, analyze the URL in the page, and at the same time The URL after the screening is filtered again, and the data obtained by the user is stored in the database. On this basis, the data will be deeply excavated, that is, a series of data analysis methods will be used to obtain information about job requirements. , the city where the post is located and a series of important information, providing useful reference and reference for the majority of employees.

Acknowledgment

This work is supported by the Anhui University of Finance and Economics Undergraduate Innovation and Entrepreneurship training program, Project number: 202210378334

References

[1] Wu Xuekai, Liu Tianbo, Hu Wenxin. Employment analysis of Java industry based on web crawler [J]. Science and Technology Information, 2021, 19(02): 13-16. DOI: 10.16661/j.cnki.1672-3791.2008 -5042-2682.

[2] Xiang Boliang, Tang Chunchun, Qian Qian, etc. Analysis of employment data based on web crawler [J]. Intelligent Computer and Application, 2020, 10(01): 223-226+230.

[3] Chen Fen, Zhang Xiaolan. The application of crawler technology based on Python language—taking the statistical analysis of information on campus dynamic sections of the official website of colleges and universities as an example [J]. Journal of Xiamen City Vocational College, 2022, 24(03): 86-91

[4] Huang Minhao, Ding Lang, Zhang Xuelian. Python-based web crawler and text visualization [J]. ComputerProgrammingSkillsandMaintenance, 2020(07):2425.DOI:10.16184/j.cnki.comprg.2020.07.009.

[5] Li Pei. Research on Python-based web crawler and anti-crawler technology [J]. Computer and Digital Engineering, 2019, 47(6): 1415-1420.

[6] Wang Bin. Design and implementation of public opinion management system based on focused crawler [D]. Shanghai Jiaotong University, 2016.

[7] Ou Gengxin.Development of GUI Applications for Groundwater Modeling Using Python[J]. Groundwater, 2020, 58(4):91-94

[8] Lin Jie. Research and implementation of theme web crawler [D]. Wuhan University of Technology, 2011.

[9] Xue Wei, Yuan Yuan, Dong Siqin, et al. Urban catering data analysis based on visualization technology [J]. Science and Technology Information, 2020, 18(18): 17-18.

[10] Lu Shufen. Design and Implementation of Web Crawler System Based on Python [J]. Computer Programming Skills and Maintenance, 2019(2):67-68.