2021 2nd International Conference on Artificial Intelligence and Education (ICAIE) | 978-1-6654-2492-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIE53562.2021.00048

Python-based crawler recruitment website program defects and visual analysis-taking 51job recruitment website as an example

Weihua Yang
Dalian Polytechnic University
Dalian, Liaoning, China
e-mail: jasonyangwh@sina.com
* Corresponding author: 970464975@qq.com

Xin Wang*
Dalian Polytechnic University
Dalian, Liaoning, China
e-mail: 970464975@qq.com

***Abstract*—With the development of science and technology such as the Internet, and the explosive growth of the amount of information, we generate data all the time every day. Obviously we have come to the era of big data. Because we have a sufficiently powerful amount of data, our research can not only be carried out in the main part of the normal distribution, but also makes it possible for the long tail part to become a part of our research. Therefore, data mining is particularly important among them, and it has become one of the main directions of research today. Combined with the current hottest employment issues, this article mainly studies how to use Python, a programming language with a rich language library and powerful computing functions, to crawl the recruitment information process of the largest recruitment website 51job, based on the large amount of information crawled You can find effective information and conduct specific research on employment positions, regions, positions, salary and other directions.**

Keywords-Python; web crawler; recruitment network

I. Introduction

With the popularization of the Internet, more than 90% of recent graduates now choose to find their job intentions on recruitment websites. In order to adapt to this trend, more and more companies publish their own recruitment information in various departments. Recruitment websites lead to cumbersome information. Choosing a company that meets your requirements is often time-consuming and laborious, and it is difficult to find a suitable job position. Therefore, based on the above problems, we use the Python request library to crawl out the recruitment information and save it locally. At this stage, there are many codes for crawling recruitment websites, but they are not complete enough. Based on them, this article has carried out problems such as the coding of the csv form, the unification of the salary unit, and the prevention of excessive http connections from being closed.

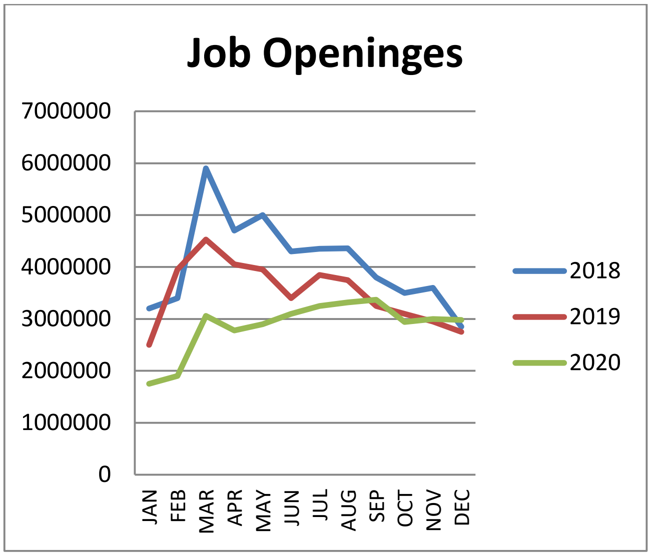


Figure1. 2018-2020 number of recruitment positions

II. Related Technical Analysis

A. Python

Python is the programming language most used by people nowadays. Its language is simple, with good interpretability and interactivity, and its codes are mostly based on English words, which are easy to read. The rigorous overall code structure makes it more efficient than similar languages and is easier to maintain and find errors.

Among the many programming languages, Python is the most friendly for beginners, without complicated nesting and fuzzy replacement, spending the same time, learning Python has a very high cost performance. According to a report, at the end of 2017, in the statistics of all development languages, it can be known that the total number of users of python has surpassed c and java as the number one popular language. As the number of users increases, the Python language environment becomes more and more perfect, and the database becomes more and more powerful. It is a powerful tool for data mining and analysis.

978-1-6654-2492-9/21/$31.00 ©2021 IEEE
DOI 10.1109/ICAIE53562.2021.00048

III. CRAWLER PROGRAM DESIGN

The topic web crawler is essentially a search program based on the Web. It traverses websites that are related to the topic according to a preset topic, automatically collects valuable network information, and analyzes the correlation relationship. The content of low-level information is pruned to form an initial set of web pages in this way .[1]

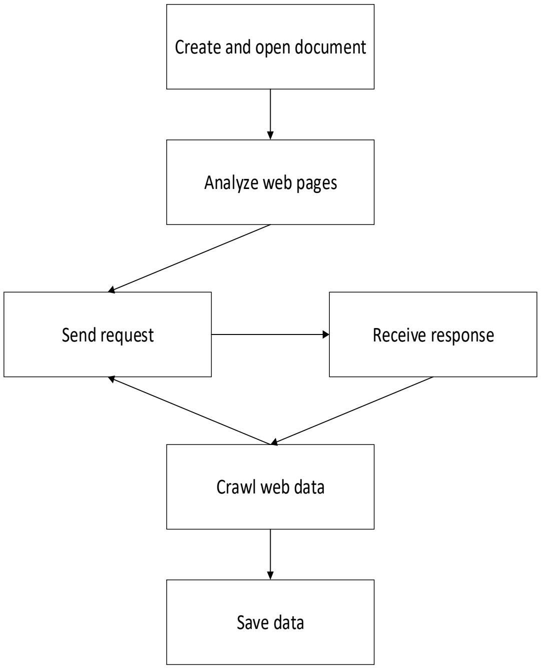


Figure2. Flow chart of web crawler design

A. Create a csv that saves the written data

Use the open statement to create a job1.csv file. When writing data, you should pay attention to the encoding form of the crawling webpage. You can see that the encoding form of the webpage on the 51job website is in GBK form, so we also write The same encoding form should be adopted to avoid special characters in the writing process, leading to errors that cannot be written. We force the errors to be ignored. This is also an innovation of this article. In the era of big data, data is booming, with small errors. It can be ignored in the face of huge data. Then use the for loop statement to write the data required for each page of the web page into the csv file. The specific code is as follows:

```
f = open('job1.csv','w',encoding='gbk',errors='ignore')
list = []
for a in list:
    f.write(a)
    f.write(',')
    f.write('\n')
```

B. Analyze web pages

rules of each page URL. In order to facilitate the construction of the rules, the URL is divided into two parts, url0 and url_end, and constructed into a loopable URL, get its list, and view it at the same time Write the code on the web page to find the information code we need, such as: job title, salary level, company name, company nature, industry, functional category, region, experience requirements, academic requirements, number of recruits, release date, job information, etc.

```
url0=
"https://search.51job.com/list/120000,000000,0000,00,2,99,+,2,"

url_end =
".html?lang=c&postchannel=0000&workyear=99&cotype=99&degreefrom=99&jobterm=99&companysize=99&ord_field=0&dibiaoid=0&line=&welfare="

url = url0 + str(i) + url_end
```

At the salary level, we can see that their units are not in a unified form, so when we crawled the data, we added innovative points to unify the salary units and write them into the document. Delete all the rows with keywords of "10,000/year" and "yuan/day" in the salary column, and transform the data of keywords with "thousand/month and "10,000/month" into K as the unit.

C. Make request and accept response

We define a function named res = requests.get. After sending the get request, we get the HTML page, and use XPath technology to get the string of the last page, that is, get the total number of pages on the website for the keyword of the job. And print it out. XPath is the XML path language, which is a language used to determine the location of a certain part of an XML (a subset of standard general-purpose markup language) documents [2]. With the URL list obtained in the previous step, the get method of the request library is used again to achieve web page crawling, and then the XPath technology is used to obtain the specific information of the corresponding post [3]. Use etree.HTML() to parse the HTML document object in string format. Only after parsing can the content in the html source code be obtained through xpath. The specific code is as follows:

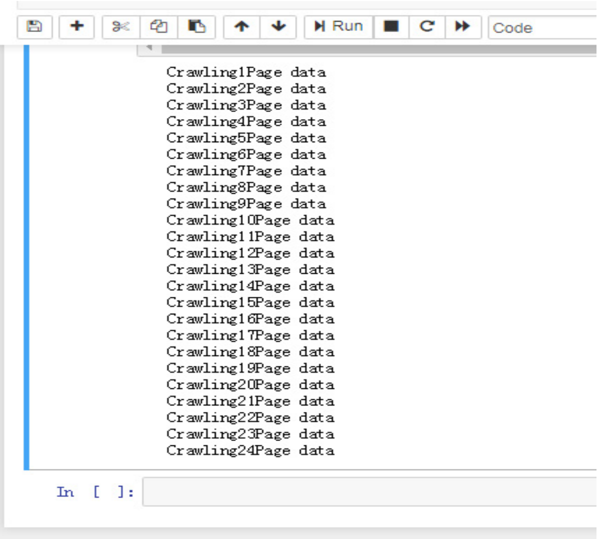
```
res = requests.get(url=url, headers=headers,verify=False)
 res.encoding ='gbk'


  p = res.text
  ex = r'job_href\":(.*?),'
  p1 = re.findall(ex, p) # re.findal
   # print(res.text)
   for a in range(len(p1)):
   url = p1[a][1:-1].replace("\\", "")
    page = requests.get(url=url, headers=headers)
     page.encoding ='gbk'
   tree = etree.HTML(page.text)
```

```
# etree.HTML()

title = tree.xpath('/html/body/div[3]/div[2]/div[2]/div/div[1]/h1/text()')[0:1]
```

During the crawling cycle, the program will open to access multiple http, causing too many openings, causing the program to freeze. The IP will think that there is a security problem and cannot continue to access the website, so that the crawling process will not reach the amount of crawled data. Forced to stop. In order to solve this problem, we introduced InsecureRequestWarning to disable security request warnings so that the crawling process can continue smoothly. The specific code is as follows:

```
from requests. packages. urllib3.exceptions import InsecureRequestWarning

requests.packages.urllib3.disable_warnings(InsecureRequestWarning)
```

After the above acquisition, the corresponding data can be crawled out.



```
Crawling1Page data
Crawling2Page data
Crawling3Page data
Crawling4Page data
Crawling5Page data
Crawling6Page data
Crawling7Page data
Crawling8Page data
Crawling9Page data
Crawling10Page data
Crawling11Page data
Crawling12Page data
Crawling13Page data
Crawling14Page data
Crawling15Page data
Crawling16Page data
Crawling17Page data
Crawling18Page data
Crawling19Page data
Crawling20Page data
Crawling21Page data
Crawling22Page data
Crawling23Page data
Crawling24Page data

In [ ]:
```

Figure3. Crawling data graph

IV. Data preprocessing

Utilize pandas functions,Pandas has functions that can filter and delete recurring tables, and retain no duplicate information. Removal of worthless data information: When obtaining valuable and useless information data, such as the content of a post, other related job information data except for the Java development engineer, the useless data information should be removed in time, so that it can be obtained Accurate Java and development post data information.[4]

V. Recruitment information visualization

Extract big data-related positions from all recruitment information, and analyze which cities across the country have greater demand for big data-related positions. Based on this, we drew a visual map of the number of big data jobs across the country. This kind of map more vividly and directly reflects the number of people in need of big data jobs in various regions of the country. At the same time, we can clearly see that our city, and The distance between urban areas with high demand for jobs in order to better analyze the selection of your own company and career.

According to the map, we can see that the demand for big data-related jobs in Hubei, Zhejiang, Guangdong, Beijing, Shanghai and other cities is very huge, far exceeding other regions. Among them, in recent data, Hubei The demand for occupies the first place, which shows that Hubei has huge room for development in this field in the near future.

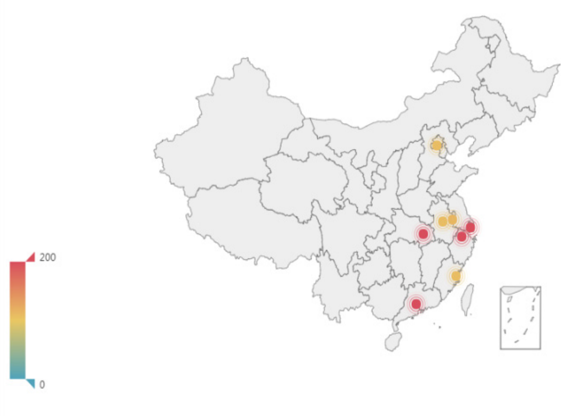


Figure4. National Big Data Job Demand Map

According to the filtered information, we extracted the experience requirements of the personnel required by each position. In order to analyze more intuitively, we made a pie chart. It can be seen that at this stage, there is a shortage of big data-related positions. The demand for fresh graduates without work experience is still very large. Similarly, because the college entrance examination chooses a major for better employment prospects, so this analysis can also be used as a basis for choosing a volunteer for the college entrance examination.

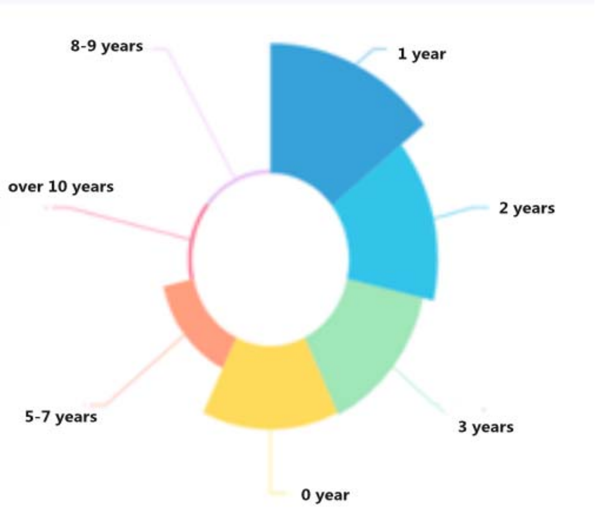


Figure5. Experience requirement pie chart

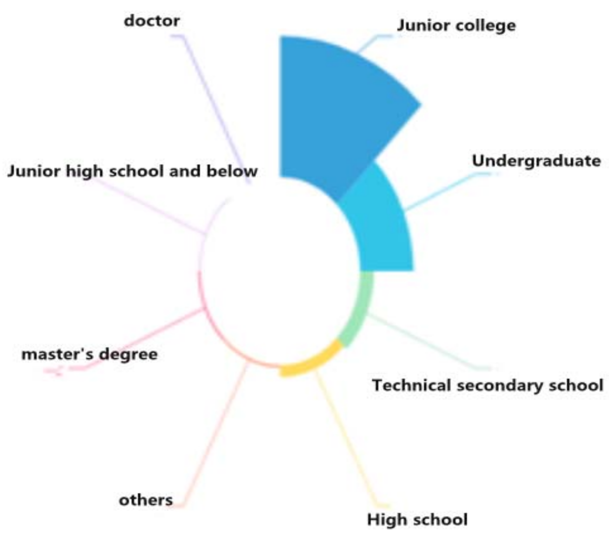


Figure6. Education requirements pie chart

VI. Conclusion

This article uses Python to crawl information on the 51job website, and adds to the predecessor’s basic code the innovation and progress of preventing writing errors, preventing http open too much and stopping the program, and salary unit conversion, so that those who are looking for a job can be more direct.It is convenient to find a job that suits you, and it also provides a basic data mining method for forecasting the development of future career positions. It also has certain research and practical significance for the national macro policy and which professional talents to be cultivated.

According to the visual analysis of the data, we can also clearly see the current educational and work experience requirements for big data-related positions, as well as the demand for their positions in various regions across the country, not only for our professional courses such as big data, data analysis. It provides reference suggestions for students from the background to choose related occupations and companies, and the data obtained by crawlers can make predictions on future job demand analysis, predict the development prospects and requirements of each job, so as to find a better career for themselves In addition, you can perform cluster analysis based on the crawled data, and then perform statistical analysis, so that for graduates after the college entrance examination, the choice of university's major and region will also have a certain reference value.

References

[1] Wang Kuo, Fei Chenjie, Liu Bosong. Research on topic crawler of convolutional neural network fused with LDA [J]. Computer Engineering and Application, 2019, 55(11): 123-128.

[2] Qi Peng, Li Yinfeng, Song Yuwei. Web data collection technology based on Python[J]. Electronic Science and Technology, 2012, 25(11): 118-120.

[3] Pan Chengjia, Liu Dongdong, Recruitment information crawling and analysis based on python [J]. Computer Knowledge and Technology, 2020, 16(27): 102-103.

[4] [2] Li Pengjie. Application of PLC technology in electrical engineering automation control Analysis [J]. Architecture Engineering Technology and Design, 2019, (16): 3588.