**Data-related Occupations in the Employment Market of China**

**Jun Wang** [wanjun@umich.edu](mailto:wanjun@umich.edu)

**1. Motivation and Summary**

Unlike data science in America, it is still in a developing stage in China, and the need for the field is surging and the potential power of the industries related with data is huge in China, announced by the media at least. I made a research on the data-related jobs, such as data engineers, data scientists and data analysts, in the employment market of the main cities of China like Shanghai, Beijing and Guangzhou, where the related industries are concentrated. Hence, the research would figure out the detail and the real situation of the employment market’s feedback to the popularity of data in China, including positions amount, cities, salary levels and education requirements, varying from the advanced employment market and the common employment market, so that we can evaluate the situation of data-related occupations.

**2. Data Sources**

Given we are trying to research both the advanced employment market and the common employment market, we should use two appropriate data sources. The data sources, lagou.com and zhaopin.com, are both big websites for recruiting especially for jobs related with IT. However, lagou.com focus on the more advanced applicants, employees and employers, while zhaopin.com is designed for more widely and common and pre-intermediate employment market including campus recruiting. Positions posted on both the websites have complete information, including position name, salary, education background requirement, city, company name, company size, industry field, etc. The information is crawled from the zhaopin.com, and JSON API is applied on lagou.com. The specific JSON API is <http://www.lagou.com/jobs/positionAjax.json> , and the website fetched for zhaopin.com is <http://sou.zhaopin.com> . In the project proposal, I planned to get fetch the webs from <http://xiaoyuan.zhaopin.com/> for zhaopin.com, however, the information for salary is incomplete in this website, so I switched to <http://sou.zhaopin.com>. Apparently, the former will give us the JSON-format data, and the later only gives us the html-format data. For both the data sources, I got 5000 posts for IT, which contain almost all the IT occupations including the data-related, and 1500 posts for the data-related jobs. The *.sql* files I submit are fetched on Oct. 13, 2016 and all the posts were posted within the last three months, which could reflect the change of the employment market recently.

**3. Data Manipulation Methods**

In the project proposal, I planned to store the information fetched from zhaopin.com in the *.csv* file, however, I decided to store all the information in *MySQL* database for processing data directly and conveniently. The JSON file from lagou.com is relatively simply to manipulate, almost all the values could be obtained from the loaded JSON. Then I got the values for *company\_name, position\_name, company\_size, city, industry, salary, jobnature, education, workyear*, and stored them in *MySQL* database. However, the process to parse the webs from zhaopin.com is relatively complicated. The values of different variables appeared in the different positions of the html, so to get them I had to design different patterns. Finally, I got the values for *company\_name, position\_name, company\_size, city, salary, education*, and stored them in *MySQL* database.

The information posted on <http://sou.zhaopin.com> was incomplete sometimes, and the formal of its contents is irregular and not fixed, so I keep fetch until I get 5000 posts. And I had to eliminate the specific characters when I fetched webs from zhaopin.com to keep consistency. In fact, I have tried three different keywords and fetched the information to find keywords which could effectively filter useless information and find data-related posts precisely. Initially, I used *data analysis* as the keyword to select data-related posts in zhaopin.com, however, I found that many occupations which were not related with data were also selected because almost all the occupations need “data analysis”. Then I used *statistics* as the keyword, given many students from statistics department are going to be engaged in data-related jobs, however, the posts amount was too small. Finally, I chose *data mining* as keyword, which could contain most of data-related occupations in China.

The other challenge is to deal with the incompatibility from Chinese words. To get and store and manipulate the information represented as Chinese, I had to convert the keywords to utf-8 format firstly. In the fetching process, I should convert specific characters to utf-8 format to transfer them into consistent format. In the analysis process, to manipulate the information in the data frame, I also should convert it. And I also set the default encoding of system as utf-8.

The salary is important to our analysis, but the formal of the information in these two datasets is inconsistent, and I had to deal with the situation where digits and Chinese characters are mixed. First, I eliminated all the characters which are not digits, then I transfer the remaining digits in the consistent format, so that I could use the transferred data in the same way.

The results like positions amount, salary should be joined together to make comparison. Hence, I calculated the results respectively, then I join the results including positions amount and salary from these two datasets then we can visualize them.

For the workflow of the source code, *zhaopin\_it.py, zhaopin\_data\_mining.py, lagou\_it.py, lagou\_data\_analysis.py* are for fetching information from webs and JSON data, *visualization.py* is to read data from the database and get the desired result and join the result together. For the codes, *zhaopin\_it.py* fetched 5000 IT posts, *zhaopin\_data\_mining.py* fetched 1500 data-related posts. In both files, *get\_save\_url* fetched the web, *get\_post\_fact* parsed the web, then all the values were stored in the *MySQL* database. *lagou\_it.py* get 5000 IT posts and *lagou\_data\_analysis.py* get 1500 data-related posts from the JSON.

**4. Analysis and Visualization**

Before the analysis and visualization, we should read the data from the database first, then we should perform the join process. After executed *pd.read\_sql*, we could get the information for data-related posts in different cities and different education for lagou.com and stored them in *pd.Dataframe*. While for zhaopin.com, we directly selected the information for data-related posts in different cities and different education from *pd.Dataframe.* The salary we get for each post has minimum salary and maximum salary, and its unit is thousand RMB, so to analysis the salary level, we should use functions *lagou\_split, zhaopin\_split\_min, zhaopin\_split\_max, lagou\_average\_salary\_min, lagou\_average\_salary\_max, average\_salary* to the average minimum salary and average maximum salary for every desired dataframes. Given the dataframes and the results from *average\_salary*, we could combine them then to analysis.

After combined the results of the two datasets, we could get many interesting relationships and insights. I made the visualization using *Excel*.

Figure 1 shows the number of data-related posts in different cities from the two tables, *test\_lagou\_data\_analysis.sql* and *test\_zhaopin\_data\_mining.sql*. Apparently, Beijing has most of the positions no matter in the advanced employment market or in the common employment market. And the advanced one provides more positions than the common one in these three cities, which shows that the advanced employment market in the major cities of China provides more opportunities than the common.Figure 2 provides more evidence, the salary levels in Beijing is higher than Shanghai’s than Guangzhou’s no matter in the advanced employment market or in the common employment market, and the advanced employment market provides higher salary that the common one in the three cities. Figure 3 shows that the advanced employment market provides higher average minimum salary and higher average maximum salary than the common one, which shows where the advanced employment market advances. However, Figure 4 shows that the percentage of data-related posts in IT posts is still low, while the advanced employment still has the higher percentage than the common one. But the good news is that Figure 5 shows that in both the advanced employment market and the common employment market, data-related posts has higher average salary level than IT posts. And the difference between the average level of data-related posts and IT posts in zhaopin.com is larger than it in lagou.com. Then we are desired to analyze the educational element. Figure 6 shows that in most of time, undergraduate is required, while PhD is rare. However, Figure 7 shows that the salary level is linearly correlated with the education requirement. The higher the education is, the higher the salary level is. The exception is it seems that in the advanced employment the undergraduate education and the master education doesn’t make too much difference.

After the above analysis, we could conclude that, Beijing provides the most opportunities and salary for data-related occupations, and the opportunities provided by the advanced employment are more than the common one, so if you want to be engaged with data-related occupations in China, go to Beijing, go to the advanced employment market. And if you are an undergraduate, it’s OK for you to seek for the data-related job, and the salary is higher than the average salary of IT. If you want to get higher salary, to be advanced, to get higher education. Although the data-related occupations in the employment market is still in the developing stage, the relative high salary and the potential power will attract more and more employees and employers.

Figure 1. Data-related posts amount in different cities.

Figure 2. Average salary of data-related posts in different cities.

Figure 3. Average salary of data-related posts in lagou.com and zhaopin.com.

Figure 4. The percentage of data-related posts in IT posts.

Figure 5. The average salary comparison: IT VS Data-related posts.

Figure 6. The education requirement for data-related posts.

Figure 7. Average salary for different education requirements.