Project Proposal for
Research on data related occupations in the employment market of China

1. Introduction

In this project, I am planning to research the job related with data such as data engineers, data scientist and data analysts, in the employment market of the main cities of China. Since the class we are taking is Data Manipulation and data science and data engineering is developing rapidly, so I want to find the employment situation related with data. The reason why I choose Chinese employment market is that unlike the field of data science in America, it is still in a developing stage in China, and the need is surging in China, announced by the media at least. So I want to find the detail and the actual conditions of the employment market's feedback to the popularity of data in China. I want to figure out the demand of the employment market such as the number of positions and the percentage data-related jobs occur in IT employment market, and the requirements from the companies such as educational background and working experience, the salary level in different cities, different requirements, different kinds of jobs.

2. Data Sources

The data sources I choose are from two big websites about recruiting. They are lagou.com and zhaopin.com, and there are a large number of positions posted by companies especially from the IT and internet on both of the websites. 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 of the websites have complete information, including position name, salary, job nature, education background requirement, work experience requirement, city, company name, company size, industry field and etc. The information is crawled from the website, and JSON API is applied on lagou.com, while MySQL is utilized to store the web information crawled from zhaopin.com. The specific website is http://www.lagou.com/jobs/positionAjax.json and http://xiaoyuan.zhaopin.com/. For every website, I crawl 5000 positions information as described before.

3. Data Manipulation

The information from zhaopin.com will be crawled by *Beautifulsoup* first, then be stored directly in the database in MySQL. While the information from lagou.com will be crawled by *Scrapy* through JSON API, then it will be stored in a *.csv* file. Then the data could be read directly from the MySQL database and the *.csv* file. Next we could use *pandas* to process and analyze the data and the datasets could be combined by *dataframe*. Given the positions from the two websites, we can generate the position number, the rate, the requirement, the skills, the salary, the company size in different cities from the pre-intermediate to the advanced. By combining them we can find the situation where date related occupations when compares to the other occupations in the employment market and figure out the preference of the common and the advanced employment market.

4. Visualization

The visualization process mainly utilizes *matplotlib*. The percentage, position number, the salary of data related occupations in the common and the advanced employment market would be visualized only by combining the two datasets. Therefore, we could figure out the factors that affect the salary generally. So the visualization of the output will be histograms, line graphs showing the number and pie charts showing the percentage.