

**目录**

[一、数据库展示 1](#_Toc511939111)

[二、解析规则 1](#_Toc511939112)

[三、例子展示 2](#_Toc511939113)

# 一、数据库展示

数据库名称为big\_data，里面有两个集合(collection)，分别为raw\_data和regulated\_data，raw\_data为爬虫结果数据库，regulated\_data为进行数据清洗后所得数据库。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | raw\_data | |  | regulated\_data | |
| 编号 | Name | Type |  | Name | Type |
| 1 | url | String |  | url | String |
| 2 | salary | String |  | salary | String |
| 3 | work\_city | String |  | work\_city | String |
| 4 | release\_date | String |  | release\_date | Date |
| 5 | work\_natrue | String |  | work\_natrue | String |
| 6 | work\_experience | String |  | work\_experience | String |
| 7 | min\_degree | String |  | min\_degree | String |
| 8 | hire\_num | String |  | hire\_num | Integer |
| 9 | job\_class | String |  | job\_class | String |
| 10 | job\_name | String |  | job\_name | String |
| 11 | company\_name | String |  | company\_name | String |
| 12 | company\_nature | String |  | company\_nature | String |
| 13 | company\_size | String |  | company\_size | String |
| 14 | company\_class | String |  | company\_class | String |
| 15 | welfare | String |  | welfare | String |
| 16 | post\_requirements | String |  | major | String |
|  | skill | String |

表1-1

# 二、解析规则

下面逐条解释数据解析的过程的过程

1、url -> url

保持不变。

2、salary -> salary

在raw\_data中，salary的类型有：‘a-b万/月’，‘a-b万/年’，‘a-b千/月’，‘a-b元/月’，‘a-b元/天’，先进行切片，获取最后三个字符前面的字符，然后根据‘-‘进行分割，将a, b转换成浮点数，再根据最后三个字符，将a，b分别转换为每个月对应的工资，单位为元/天，仍记作a，b，最后进行字符串的拼接，将‘a-b’存入regulated\_data。

3、work\_city -> work\_city

在raw\_data中，work\_city为城市名，或城市名-xx区，城市名一般为两个字，也有少量三个字或四个字的城市名，对此，我们自行创建了一个列表，如果work\_city前三个字或前四个字再此列表，则取其名字，否则 work\_city进行切片，取前两个字符，存入regulated\_data。

4、release\_date –> release\_date

在raw\_data中，release\_data为‘mm-dd发布’，只需对其进行切片，去除后两个字符，再将其转换为Date类型，最后将其存入regulated\_data中即可。

5、work\_natrue -> work\_natrue

保持不变

6、work\_experience –> work\_experience

在raw\_data中，work\_experience的类型有：‘经验不限’、‘a年经验’、‘a-b年经验’，若为‘经验不限’，则将‘0-0’存入，若为‘a年经验‘，则将’a-a’存入，若为‘a-b年经验’，则将‘a-b’存入。

7、min\_degree -> min\_degree

保持不变

8、hire\_num -> hire\_num

在raw\_database.big\_data中，hire\_num的类型有‘招a人‘、’招若干人‘，若为‘招a人‘、则将a存入，若为‘招若干人‘，则将‘0’存入，此处‘0’只做标记用。

9、job\_class -> job\_class

保持不变

10、job\_name –> job\_name

保持不变

11、company\_name -> company\_name

保持不变

12、company\_nature -> company\_nature

保持不变

13、company\_size -> company\_size

保持不变

14、company\_class –> company\_class

保持不变

15、welfare > welfare

保持不变

16、post\_requirements –> major

首先对专业建立一个列表，然后将post\_requirements中出现的所有专业存到lean\_database.regulated\_data。

17、post\_requirements –> skill

首先进行统计分析，获取大数据岗位所学的技能，然后建立一个技能列表，post\_requirements中出现的所有技能存到regulated\_data。

# 三、例子展示

下面我们展示数据清洗的一个例子：

此为清洗前数据

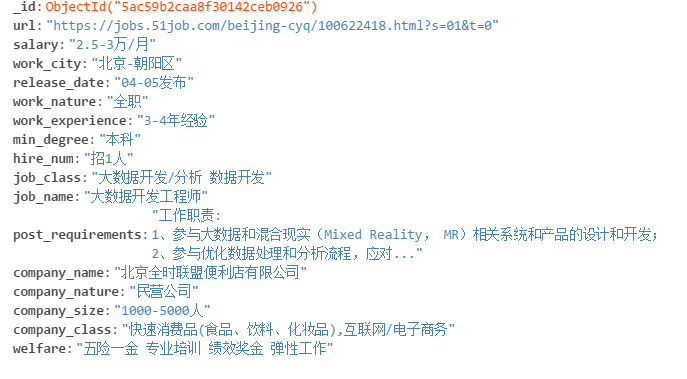


图3-1

此为清洗后数据

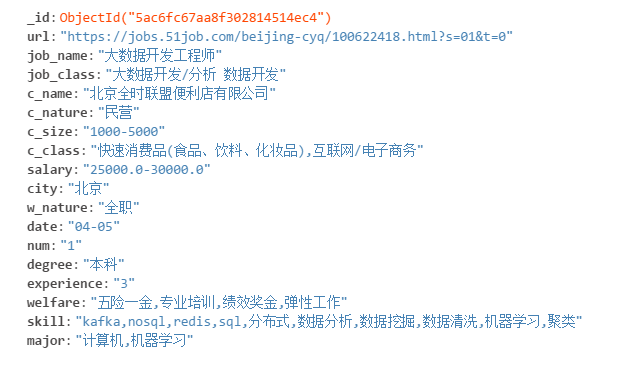


图3-2