职达

第九届中国大学生服务外包创新创业大赛

队伍成员：

曾永家右　黄泰北　赵辉

龙兰心　罗雪蕾

基于大数据的人岗匹配系统“职达”

解析规则

职达信息技术有限公司

达芬奇队

指导教师：徐进 尹帮旭

目录

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

```
_id: ObjectId("5ac59b2caa8f30142ceb0926")
url: "https://jobs.51job.com/beijing-cyq/100622418.html?s=01&t=0"
salary: "2.5-3万/月"
work_city: "北京-朝阳区"
release_date: "04-05发布"
work_nature: "全职"
work_experience: "3-4年经验"
min_degree: "本科"
hire_num: "招1人"
job_class: "大数据开发/分析 数据开发"
job_name: "大数据开发工程师"
post_requirements: "工作职责:
1、参与大数据和混合现实（Mixed Reality， MR）相关系统和产品的设计和开发；
2、参与优化数据处理和分析流程，应对..."
company_name: "北京全时联盟便利店有限公司"
company_nature: "民营公司"
company_size: "1000-5000人"
company_class: "快速消费品(食品、饮料、化妆品),互联网/电子商务"
welfare: "五险一金 专业培训 绩效奖金 弹性工作"
```


图 3-1

此为清洗后数据


```
_id: ObjectId("5ac6fc67aa8f302814514ec4")
url: "https://jobs.51job.com/beijing-cyq/100622418.html?s=01&t=0"
job_name: "大数据开发工程师"
job_class: "大数据开发/分析 数据开发"
c_name: "北京全时联盟便利店有限公司"
c_nature: "民营"
c_size: "1000-5000"
c_class: "快速消费品(食品、饮料、化妆品),互联网/电子商务"
salary: "25000.0-30000.0"
city: "北京"
w_nature: "全职"
date: "04-05"
num: "1"
degree: "本科"
experience: "3"
welfare: "五险一金,专业培训,绩效奖金,弹性工作"
skill: "kafka,nosql,redis,sql,分布式,数据分析,数据挖掘,数据清洗,机器学习,聚类"
major: "计算机,机器学习"
```


图 3-2