# 招聘数据采集项目

## 仿百度爬虫

|  |  |
| --- | --- |
| 软件环境 | Eclipse 控制台 mysql |
| 时间 | 2017-03-21 |
| 创建人 | 马义玺 |
| 版本 | V0.1 |

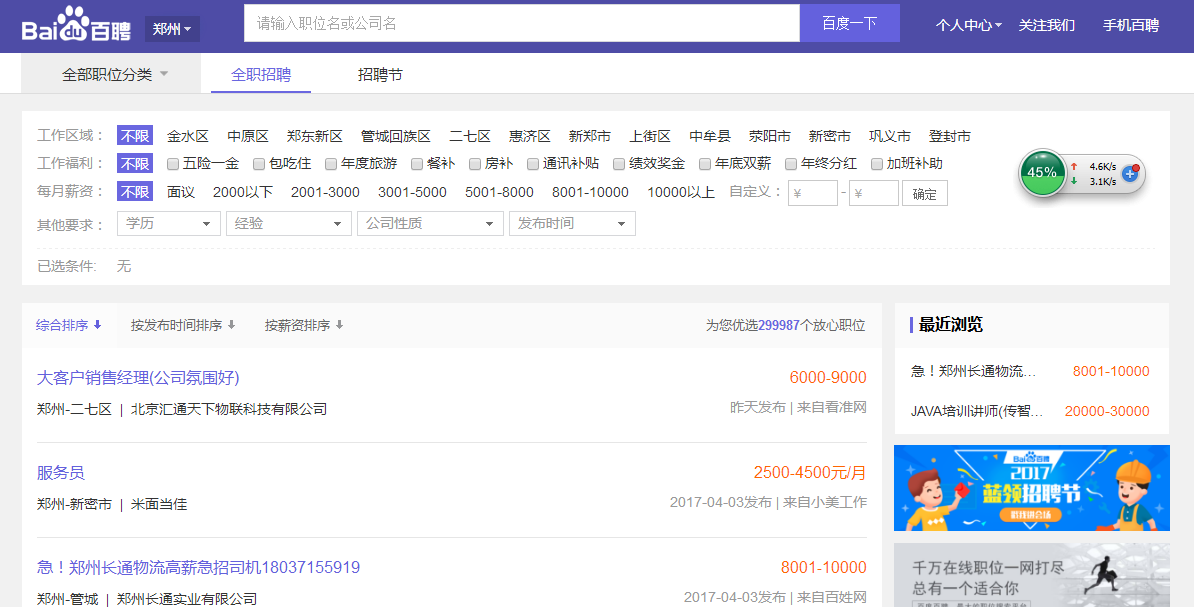
## 前置条件：

学完课程

|  |  |  |
| --- | --- | --- |
| JavaOOP | Java面向对象编程 | 软件技术基础：软件开发相关专业术语； |
| Java开发环境的搭建:JDK安装和环境变量配置，开发工具Eclipse介绍； Java语言的特点、Java程序开发流程、JRE及JVM； HelloWrold：创建工程、Java工程和代码结构解析，Java基本的语法结构； |
| Java语法基础：计算机中的进制转换，Java中的数据类型； 常量和变量，运算符与表达式，运算符优先级；类型转换； |
| Java语法基础：分支语句； |
| Java语法基础：循环语句； |
| 函数和函数的重载； |
| 数组的应用和针对数组的常用算法； |
| Java中类和对象的基本概念，类实例化和对象的调用； Java中类的封装；成员变量、属性、方法、构造函数； Java中的包； 访问修饰符； this、static、final关键字及其用法； 内部类，匿名内部类； |
| Java中的继承；super； Java中多态及多态的实现：抽象类和接口； |
| Java中的常用类：Object,数据类型封装类；字符串类String和StringBuffer；日期处理相关类；Math类,BigInteger等； |
| 异常；log4j; |
| Java集合框架；泛型集合；Collections/Arrays; |
| Java I/O；Java多线程； |
| 项目案例： |
| Mysql | MySQL数据库基础应用 | 数据库基础概念； Mysql数据库的创建和基本管理； |
| 表的创建和管理；用户的创建于管理；DDL，DML，DCL语句分类 |
| 常用函数（字符，日期，数学）； |
| 数据查询和操作：CRUD、基本查询 |
| 表的高级查询：运算符、模糊查询、排序、分页、分组查询（max/min/sum/count/avg) |
| 笛卡尔积问题、 多表联合查询（内连接，左连接，右连接）、子查询； |
| union/union all;数据库三范式 |
| mysql数据库的引擎介绍；mysql数据库的其他对象：view，procedure，trigger，event |
| JDBC | JDBC数据访问 | Junit，注解；JDBC核心API:Connection、Statement、ResultSet； |
| sql注入； PreparedStatement和CallableStatement接口；工具类MyDBUtils类；JDBC事务处理； |
| jdbc 大数据的操作，jdbc批处理，jdbc事务处理 |
| 多表操作；javabean，DAO模式；工作最佳实践：java日期和sql日期的转换；金钱的数据类型 |

效果：

百度招聘



## What

根据智联招聘的数据来源， 采集公司招聘信息的爬虫系统（分三步，第一采集网址，第二采集内容），同时把数据录入到mysql数据库中

**Where**

找工作的时候

锻炼java基础项目的时候

**When**

锻炼项目的时候用

需要自动采集数据的时候

需要数据库存储的时候

## 业务需求：

我们生活会有很多重复，繁重的工作来人工处理，

我们可以使用 java 自动处理， 代替手工，提高工作效率

举例：招聘信息，找工作

无效信息过滤

## 思路：

1. 输入城市和职位
2. 智联招聘网站招聘信息的网址
3. 得到单个公司发布的职位招聘信息
4. 把招聘信息存放到数据库mysql中
5. 查询出来数据，并实现到eclipse控制台

**How**

1. 掌握mysql 数据库
2. 掌握java基础，类对象
3. 掌握如何连接数据库

4. 研究java版的采集网页源代码的框架 httpclient

5. 查看html页面 框架 jsoup ，相当于xpath方式， 也可以用正则表达式，或者是前后截取的方式（abc 取b的话就是前面是a后面是c的中间的字符）。

6. 采集到的数据保存到临时实体对象里面

7. 把临时实体对象存到mysql数据库中。

8. 清空临时实体对象。

9. 采集的数据元素有： 职位名 职位月薪 工作地点 发布日期 工作性质 工作经验 招聘人数 职位类别 职位描述 公司介绍 工作地点 公司名称 公司规模 公司性质 公司主页 联系电话 来源网址。

参考图片：

http://jobs.zhaopin.com/333373986250073.htm



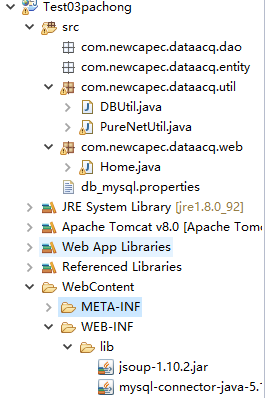
自己看的：

1. 采集招聘信息职位网址列表
   1. <http://sou.zhaopin.com/jobs/searchresult.ashx?jl=%E9%83%91%E5%B7%9E&kw=java&sm=0&p=1>
2. 采集单个职位信息
   1. 已经完成， 采集后生成sql语句
3. 入库到mysql
   1. Jdbc DBUtil db\_mysql.properties
4. 在控制台输出

## 遇到的问题

1. Split 如果是 | 前面要加两个\\
2. 网址中竟然还有 CZ 开头的
3. 导包的地方容易弄错

# 系统演示效果



>>>>>>>>>>>>>>欢迎使用招聘数据采集项目<<<<<<<<<<<<<<<<<

请输入你期望工作的城市(如： 郑州)：

郑州

请输入你期望的工作职位（如： java）：

java

insert into job(jobName,companyName,xinzi,addDate,jobMemo,companyMemo,disable) values('Java实习生','郑州戌轮教育科技有限公司','6001-8000元/月?','2017-04-01','岗位要求： 1. 在校生、应届毕业生及工作经验一年以下人员。 2. 大专及以上学历，理工科专业，计算机专业优先。 3. 有良好的逻辑思维能力，基本的英语阅读和书写能力。 岗位职责： 1. 从事公司项目组项目研发工作；? 2. 实习期4.5个月，转正后可选择到上海或者郑州任职，薪资上海6K起，上海4K起。',0);

影响了 1 行

insert into job(jobName,companyName,xinzi,addDate,jobMemo,companyMemo,disable) values('高级Java工程师','中机商城有限公司','10001-15000元/月?','2017-04-01','职位描述： 1、负责接口平台主要模块代码编写； 2、参与接口平台相关项目的开发工作。 岗位要求: 1、精通Java，有5年以上Java较深层次开发经验，具备很强的Java功底； 供平台，?',0);

影响了 1 行

。。。。。。（此处省略好多行）

》》》》》》》》采集完毕《《《《《《《《《《《

请输入你期望工作的城市(如： 郑州)：