**基于互联网大数据的计算机相关专业招聘分析与应用（分析报告）**

团队成员：简雁，武盛弘，郗化鹏，李静，徐静如

（所在单位：曲阜师范大学软件学院，所在地区：山东省济宁市曲阜市，邮政编码：273100）

摘要：近几年来，计算机和信息技术迅猛发展并且普及应用，行业应用系统的规模迅速扩大，所产生的数据呈爆炸性增长，因此提高对数据的提取、处理和分析能力，通过进一步加工实现数据的“增值”已经成为现在世界的迫切需求。本项目通过编写python网络爬虫程序，使用requests库、bs4库等爬取互联网中近期与计算机相关专业的招聘信息，收集各专业学科学习课程内容。通过数据清洗、转化、可用信息提取等技术手段获得可用的原始数据。对原始数据进行初步的分析后，利用jieba、gensim等python第三方机器学习库，把项目要求的最终结果通过可视化的手段展现出来。

关键词:机器学习;网络爬虫;jieba分词;gensim;知识图谱

正文：

# 1、引言

## 1.1系统的背景

随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长，动辄达到数百TB甚至数十至数百PB规模的行业大数据已远远超出了现有传统的计算技术和信息系统的处理能力。信息技术发展进入了大数据时代，因此，寻求有效的[大数据处理](http://www.raincent.com/list-10-1.html)技术、方法和手段处理数据，利用这些数据，挖掘数据的深层价值，是未来信息科技发展的趋势，也是大数据技术产生的背景。

在竞争日益激烈的现如今社会，大学生毕业之后的就业问题已然成为了社会关注的热点问题，面对严峻的就业形势，我们更加意识到大学生不仅需要扎实的专业知识技能，对本专业及相关专业的就业城市、薪水、排名、优势等前景的了解也是至关重要的。做好充分的了解有利于大学生正确定位自己，正确的选择职业，从而更好的把握机遇，更好的实现就业，拥有一份体现自身价值与个人能力相匹配的职业，为自己的未来找到正确的发展方向。

## 1.2系统的意义

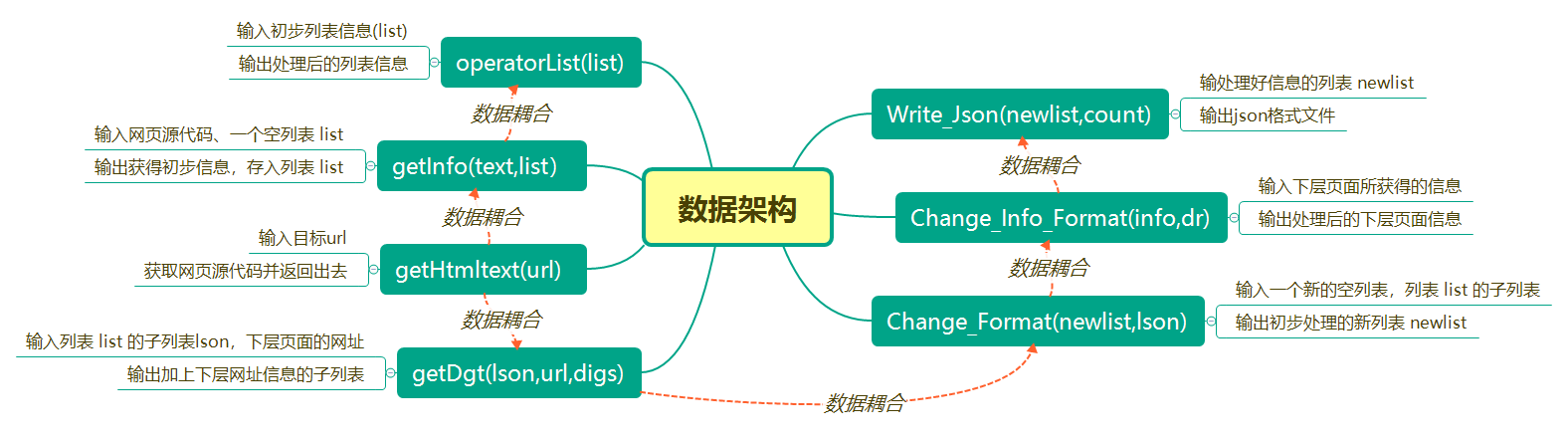
由于新时代对人才需求的增长和高校扩招的影响，我国高等教育已由精英化教育向大众化教育转型。在转型过程中大学毕业生人数出现了急剧增加的趋势，高等教育难以适应经济、社会发展和就业市场变化所带来的需要，使得大学生的就业问题愈发严峻。针对当前这种现状，利用大数据分析对大学生职业提供导向与建议具有重要的现实意义。

该系统可以帮助用户正确认识到自己的就业方向和优势劣势，帮助用户发掘自己的潜能，从而确定自己的职业发展道路。可以增强寻找职业时的目的性和计划性，提升成功的机会。职业寻找要有计划、有目的，不可盲目地尝试，很多时候我们的职业生涯受挫就是由于生职业划没有做好，好的计划是成功的开始。另外可以为用户正确定位，现今社会竞争十分激烈，要想在这场激烈的竞争中脱颖而出并保持立于不败之地，必须对自己的职业方向、优势有较好的理解，以便更好的把握机遇。

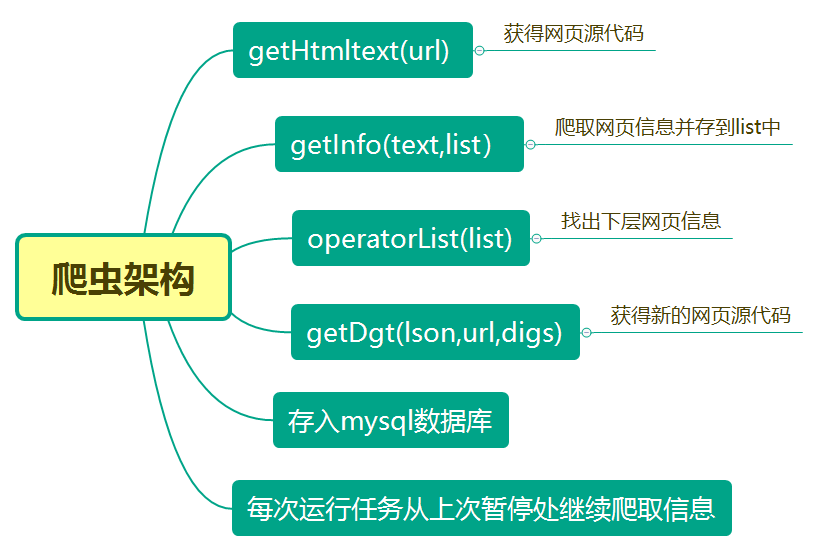
# 2、数据处理过程

整个数据处理流程可以概括为三步，分别是采集、清洗和转化，具体步骤如下：

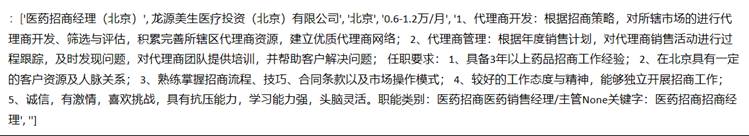
## 2.1数据采集

本项目数据采集利用Mysql数据库来接收本小组从前程无忧、齐鲁人才网、拉勾网等多个网页上爬取的大约43万的数据。

具体采集过程中，首先,getHtmltext()函数中采用requests库中的requests等方法进行网页总体源代码信息的爬取，requests是用python语言基于urllib编写的，采用的是Apache2 Licensed开源协议的HTTP库，Requests比urllib更加方便，节约大量的工作。Beautiful Soup借助网页的结构和属性等特性来简易的解析网页， 然后再调用getInfo()函数，该函数使用bs4库中的BeautifulSoup类对已爬取的网页源代码进行初步的处理，通过对BeautifulSoup类对象使用soup.find\_all()函数对网页所需信息进行一些初步的查找和提取并将其写入列表中，对列表操作进行初步数据处理，再根据该列表中标明工作具体要求网址的项进入下层网页，最后采用类似的操作对具体工作信息再进行数据的初步爬取与处理。



爬取的网页内容最终都存在于一个多维列表之中，大致格式如下图：



## 2.2数据清洗

### 2.2.1初步处理过程：

初步处理主要是通过调用字符串处理工具正则表达式（re）库，对信息进行正则匹配及处理，通过观察可以发现具体处理的地方是：

1、工资：由一个整体分割为最高工资和最低工资

2、关键词：使用jieba分词根据具体信息（info）字段内容将分词结果取词频较高的词集作为关键词，其中关键词分词的规则进行了一定的人工干涉

### 2.2.2数据清洗过程

在获取网页数据时，我们发现，在诸如职位名、具体信息等数据项中因不明原因出现了参差不齐的数据，其中不仅有格式混乱，字符混用的情况，更有内容掺杂None或乱码或缺少部分数据项的数据。通过将获取的数据转化为字典并检测相关键是否存在完成了对缺少数据项数据的清洗，通过将文件重新编码解决乱码问题，通过正则表达式替换解决了None和中英文字符混用的问题，最后将清洗后的数据放入字典设置数据库表项主键的方式完成了数据的去重。

## 2.3数据转化

我们使用了MySQL数据库,在Navicate For MySQL中创建数据库和对应表后设置表项主键为(职位名,公司名)，设置表项为职位名，公司名，工作地点(省)，工作地点(县市区)，最低工资，最高工资，具体信息，关键词，各表项均为varchar(num)类型并设置非空。通过python的MySQLdb库连接数据库后将清洗后的数据产生的字典按键写入数据库完成了对数据的转化。

# 数据分析与挖掘方法

对数据进行分析和挖掘主要目的在于发现数据隐藏的信息，对上一步产生的数据库文件进行分析知，隐藏的信息可能位于职位名，工作城市，具体位置，最低工资，最高工资，岗位职责，任职要求上。通过分析相同或相似职位名可得到职业需求量，通过分析工作地点进行相关统计可以得到工作的城市分布图，通过分析最低工资最高工资可以得到部分省县市区的最低最高薪资分布，通过分析具体信息和提取的关键词可以得到企业对招聘人员的具体要求。

由上可知我们可以从已导入的数据获取的隐藏信息有:企业对大数据要求最迫切的前十名招聘职位（通过计算职位名或具体信息出现大数据关键词的职位出现的次数获得）、大数据职位需求量最高的前十名城市（通过计算职位名或具体信息出现大数据关键词的城市出现的次数获得）,计算机相关专业需求量前十名（通过计算职位名或具体信息或关键词中计算机相关专业出现的次数获得），计算机专业薪水最高的前十名招聘职位（通过计算职位名或具体信息或关键词中涉及计算机的职位按薪水排序获得），提供简历由已有数据中获得最匹配职业前十名（通过分词产生词向量后由机器学习进行相似度匹配获得）。

由上可知，为了精确提取职位名和具体信息，我们使用了jieba分词进行关键词的提取，并将分词后的数据进行词频计算选择前几的词当做该招聘信息的关键词写入数据库。剩余隐藏信息一部分可由sql语句查询对应表项获得，另一部分可由sql语句查询生成的关键词获得。为了保证生成关键词的准确性，我们使用了用生成的关键词产生词云人工筛选的方式，通过数十次的人工筛选确保了关键词的粒度和精确性。

在获取了较精确的关键词后，我们可以通过gensim库提供的机器学习的内容进行训练，之后通过计算相似度来获得提供简历与数据库中较为匹配的数据。。

# 结论



**参考文献：**

[1] 昂温 G，昂温 P S. 陈生铮，译. 外国出版史. 北京：中国书籍出版社，1988： 12-15

[2] Peebles P Z, Jr. Probability, random variables, and random signal principles. 4th ed. New York: McGraw Hill, 2001：25-26

[3] 程根伟. 1998年长江洪水的起因. 见：许厚泽，赵其国. 长江流域洪涝灾害与科技对策. 北京：科技出版社，1999：32-36

ECharts: A Declarative Framework for Rapid Construction of Web-based Visualization

Deqing Li, Honghui Mei, Yi Shen, Shuang Su, Wenli Zhang, Junting Wang, Ming Zu, Wei Chen.

Visual Informatics, 2018 [[PDF]](http://www.cad.zju.edu.cn/home/vagblog/VAG_Work/echarts.pdf)

**命题老师评语：**

**联络方式：**

（１）通讯地址：

（２）电话：

（３）Ｅ－ｍａｉｌ：