

**课 程 报 告**

**课程名称： 大数据导论**

**专业班级： CS1707**

**学 号： U201714785**

**姓 名： 聂鸿勇**

**U201714786**

**王占成**

**U201714792**

**李澳阳**

**指导教师： 崔金华**

**报告日期： 2019年12月12日**

**计算机科学与技术学院**

**目 录**

[1 引言 1](#_Toc27055563)

[**1.1课题背景与意义 1**](#_Toc27055564)

[**1.2国内外研究现状 1**](#_Toc27055565)

[**1.3项目的主要研究工作以及各项工作需求分析 2**](#_Toc27055566)

[**1.4组内分工 2**](#_Toc27055567)

[2 爬虫 3](#_Toc27055568)

[**2.1数据集说明 3**](#_Toc27055569)

[**2.2应用技术说明 3**](#_Toc27055570)

[**2.2.1 技术选取 3**](#_Toc27055571)

[**2.2.2 技术实现 3**](#_Toc27055572)

[**2.3项目过程 4**](#_Toc27055573)

[**2.4项目结果分析 4**](#_Toc27055574)

[3 总结与展望 5](#_Toc27055575)

[**3.1遇到的问题 5**](#_Toc27055576)

[**3.2心得体会 5**](#_Toc27055577)

[**3.3未来展望 5**](#_Toc27055578)

# 1 引言

## 1.1课题背景与意义

（1）背景：

近年来，中国高校毕业生数量逐年增多，大学生面临严峻的就业形势，毕业生的就业受到了极大的挑战，大学生在毕业后能否顺利就业，能否找到一个让自己满意的工作已成为全社会普遍关注的热点问题。大学毕业生就业问题日趋严重，而国家对青年的关注永远是最强烈的。当代大学生是社会建设的接班人，是人才的主要来源，要建设好社会，当然不能缺了人才，种种国情考虑下，青年就业问题，尤其是大学毕业生就业问题已成为社会关注的难点问题之一。自2006年起，中国每年需要就业的大学生超过400万人，而大学生就业率一直徘徊在70%左右，现在更是到了60%的边界，这些数字表明，大学毕业生客观上已经成为中国城镇就业的主要目标群体，“大学生就业难”这一问题已经成为一个客观事实。

（2）意义：

鉴于当前大学生就业问题，我们选择了招聘信息分析。通过搜集各大公司的招聘信息，分析招聘工作的工作类型、工作地点、薪资等，来让大学生们对当前的就业风向有一个全面、良好的认识。

## 1.2国内外研究现状

（1）国内研究现状：

国内不同的研究分析表明，当前大学生深造比例持续上升，毕业生就业压力没有明显的增加，但80%的应届大学生选择了直接就业或创业，只有20%的大学生选择正在读研或准备考研。与此同时，“北上广深”就业比例仍然占有很大的比重，但是“新一线”城市的吸引力不断增加。新一线城市纷纷上演“抢人大战”，频频出招争夺人才，取得了显著的效果。在“新一线”城市就业的应届本科毕业生中，外省籍毕业生占比从2014届的27.9%上升到了2018届的37.3%。

对于不同的专业就业满意度而言，信息安全、软件工程、网络工程、通信工程、数字媒体艺术就业满意度比较高，语文教育、法律事务、初等教育就业满意度比较低。

（2）国外研究现状：

对美国的就业现状分析而言，拥有大学学位的成年人不到20%，从1990到2018年期间，青年大学生毕业生的整体就业率有所下降，闲置人数增加。而且黑人与白人相比，其失业率和闲置率较高。虽然白人毕业生的失业率基本上恢复到2000年水平，但其他种族的失业率仍高于白人，并且明显高于2000年的水平。可见美国大学生的就业形势也不容乐观。

## 1.3项目的主要研究工作以及各项工作需求分析

根据项目的选题，选择一个具有较大参考意义的数据来源。经过对许多网站的浏览以及小组间的讨论，决定选取拉勾网作为目标网站进行爬取。并且在爬取信息之前，确定所要爬取的信息区域。最终将爬取到的信息存储起来，交付给其它小组处理、分析。

## 1.4组内分工

本次作业我们小组负责数据级工作，小组成员具体分工如下：

李澳阳：课题选取、针对反爬取处理、数据爬取

聂鸿勇：基础爬取、针对反爬取处理、数据爬取

王占成：课题选取、基础爬取、数据爬取

# 爬虫

## 2.1数据集说明

数据来源：拉勾网（[www.lagou.com](http://www.lagou.com/)）

获取方式：通过编写Python爬虫脚本爬取，以调用站内关键词搜索接口的

方式进行爬取，因此会出现部分数据重复的情况。

数据集大小：总共爬到约1GB的原始数据，190多万条记录

数据格式：为便于处理分析，采用CSV格式存储，字段之间由逗号分隔

数据关键属性：职位ID，职位名称，公司ID，公司名称，公司规模，

产业领域，技能要求，城市，薪资，工作经验，学历要求等

数据完整性：数据集内容基本完整，部分记录有少量字段缺失

## 2.2应用技术说明

### 2.2.1 技术选取

选用python语言编写爬虫脚本，原因是python具有如下几个优点：

* 语法接近自然语言，简单易上手，开发速度快
* 抓取网页的接口相比一些静态语言更简洁，便于使用且学习成本低
* 对网页访问提供了完整的API，能模拟浏览器的行为应对反爬虫
* 有强大的第三方库如beautifulsoap对爬取到的网页进行处理

### 2.2.2 技术实现

最初计划直接爬取HTML文档然后使用beautifulsoap解析出需要的数据，爬取过程中发现网页限制访问前30页的数据，因此能爬到的数据量非常有限。后来通过对网页抓包，找到了网页调用的Ajax API接口的URL。尝试直接访问该URL，发现服务器不返回数据。

仔细分析后推测该网站使用了反爬虫技术。再次对浏览器的访问进行抓包，发现服务器返回HTML网页时还在响应头返回了一个cookie字段，推测是客户端下一次请求需要将该cookie字段放入请求头中。使用Postman验证该推测，发现服务器能正常返回数据，说明成功伪装为浏览器绕过了服务端的反爬虫验证。接下来着手编写爬虫代码。

使用Python的Requests库来模拟浏览器发送HTTP请求，获取服务器返回的接口数据。然后使用csv库将获取到的数据以csv格式写入文件中。

由于服务器有反爬虫手段，每次请求会对cookie进行验证，故还需使用request\_html库，在每次请求接口数据前额外发送一次请求，并将服务器返回的cookie添加到下一次请求的请求头中，即可绕过反爬虫验证。

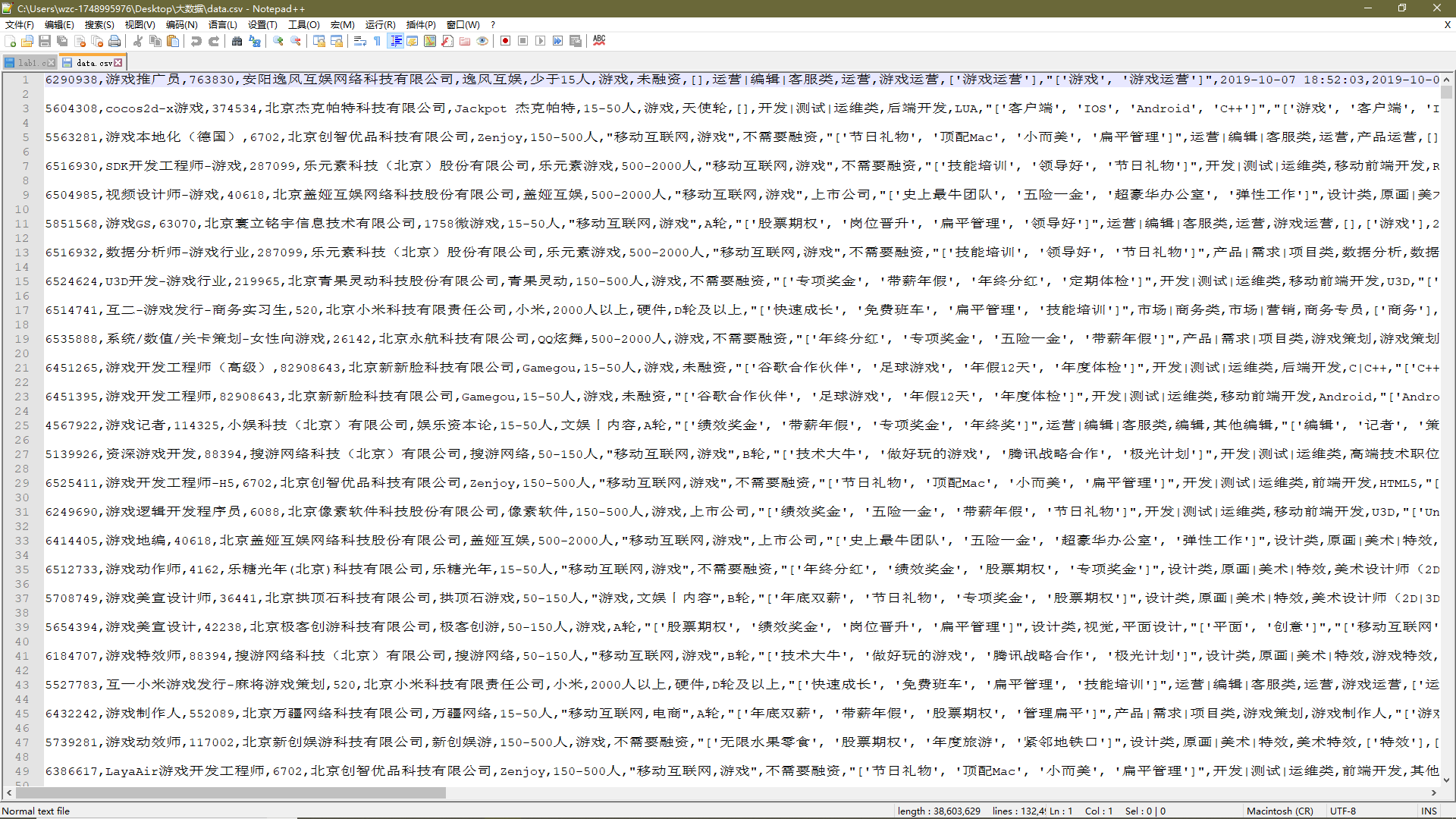
## 2.3项目过程

1. 确定项目方向及需要爬取的数据，对数据进行初步分析，定义数据结构，选择数据的存储方式。
2. 使用python编写爬虫代码，初步实现爬虫能够对拉钩网进行单页少量数据的爬取的功能。
3. 进一步完善爬虫代码，实现自动分地区、页面进行大量数据的爬取。
4. 对网站数据进行爬取。
5. 在进行大量数据爬取的过程中，发现网站存在反爬虫，在代码运行过程中会被对方服务器中断连接。
6. 根据反爬虫的种类，修改爬虫代码，使其能够避开反爬虫，实现连续不断的爬取功能。
7. 对爬虫代码进行爬取地区、页数的调控，使得小组成员可同时运行爬虫程序对网站进行数据爬取，加快数据集收集的进度。

## 2.4项目结果分析

爬到的结果共计190万余条。

爬到的部分结果如下图所示：



# 3 总结与展望

## 3.1遇到的问题

由于服务器有反爬虫手段，每次请求会对cookie进行验证，故还需使用request\_html库，在每次请求接口数据前额外发送一次请求，并将服务器返回的cookie添加到下一次请求的请求头中，即可绕过反爬虫验证。

## 3.2心得体会

任何数据本身都有其可拓展的价值，要善于发现、利用数据的价值。数据获取是大数据项目建设中的第一个环节，也是后续大数据应用的基础，因此在项目的开始就要明确项目所需要的数据，对数据进行分析并对各属性有一定的初步了解，确定好爬虫代码中的数据结构，这样能够更好地编写爬虫代码，避免不必要的麻烦。

在对数据进行收集的过程中需要解决：如何进行有效的数据存储、如何保证数据的高并发访问等问题。大数据可以广泛应用于各个领域，通过对大量数据的挖掘、对不同数据集进行针对分析后，能够总结经验，可在一定程度上预测发展趋势，为用户辅助决策做出巨大贡献。

## 3.3未来展望

随着中国经济的高速发展，就业的难度肯定是会不断地降低。而且国家对于政策上地支持力度不断地加大。通过扩大投资来保增长和保就业，对企业减轻负担和加大扶持等措施稳定就业岗位，并且通过积极的就业政策扩大就业和促进创业。加大政策扶持力度，鼓励和支持有利于扩大就业的劳动密集型产业、中小企业、非公有制经济和服务业更好发展，更多吸纳就业。在一系列积极的宏观经济政策和就业政策作用下，这种严峻的就业形势将会随着经济复苏而逐步缓解。

但就业形势很大程度上取决于经济的走向，随着经济全球化的程度不断加深，金融危机等情况会对就业形势产生巨大的影响。因此，我们在通过积极的政策改善就业情况的同时也要提前制定好对困难的应对措施，这样才能使我国大学生的就业形势稳定并且不断地改善。