泰迪内推平台招聘与求职双向推荐系统构建

摘要

针对企业招人难和求职者找工作难的问题，泰迪内推平台通过对招聘信息进行分析研究，了解不同职业领域的需求特点，推出了招聘与求职双向推荐系统，经过信息获取和处理、岗位画像分析、模型建立等流程，为广大求职者和企业提供有价值等的就业招聘信息。

关键字

信息爬取，岗位画像，求职者画像，岗位匹配度，求职者满意度，岗位人才双向推荐

目录

第十一届“泰迪杯” ..... **错误！未定义书签。**
泰迪内推平台招聘与求职双向推荐系统构建 1
摘要 1
1. 引言 3
1.1 背景 3
1.2 问题重述 3
1.3 相关工作 3
1.4 解决流程 4
2. 数据信息爬取 4
2.1 招聘公司信息爬取 5
2.2 求职者信息爬取 5
3.数据预处理 6
3.1 文字列表元素化处理 6
3.1.1 公司招聘信息的文字列表元素化处理 6
3.1.2 求职者信息的文字列表元素化处理 7
3.2 信息数字化处理 8
3.2.1 学历信息数字化处理 8
3.2.2 最低薪资和最高薪资的数字化处理 9
3.1.3 经验要求、岗位人数和工作经验数字化处理 10
3.3 公司 ID 和求职者 ID 数据处理 10
4.招聘与求职信息分析 11
4.1 招聘信息画像的建立 11
4.2 求职者信息画像的建立 15
4.3 招聘市场趋势分析 18
4.3.1 工作地区数据分析 18
4.3.2 工作职位数据分析 18
4.4 模型评估及优化 19
5.岗位匹配度和求职者满意度模型的构建 19
5.1 数据预处理 19
5.2 评分模型的建立 20
5.3 匹配度和满意度结果的输出 21
6.招聘求职双向推荐模型的建立 22
6.1 岗位匹配的计算分析及模型 22
6.2 岗位和求职者的选择算法分析及模型 22
6.3 双向推荐模型的优化 23
总结 18
参考文献 25

1. 引言

1.1 背景

在新时代背景下，随着大学生毕业人数不断增加，大学生求职问题已成为广泛关注的社会热点。而且受疫情影响，诸多企业的招聘都改为线上进行，脱离时间和空间的限制，招聘需求不断上涨，有近六成企业招聘需求增加，其中需求量较大的科技研发、数字化、蓝领技能岗位都存在不同程度的人才短缺。但从人才供给来看，应届生数量增加，2022 年高校毕业生达到创纪录的 1076 万人，而且部分企业校招开展暂缓或推迟，因此出现校招需求缩减或冻结，这些因素都加剧了应届生就业的严峻形势。基于种种因素，出现就业竞争压力大、招聘与求职信息不对称等现象。

泰迪内推平台是聚焦于“大数据+”和“人工智能”领域的求职招聘网站，该平台融合了多家企业发布的招聘信息，同时平台也为求职者提供求职信息的展示。为缓解毕业生就业压力，同时满足企业对人才的需求，泰迪内推平台会定期为高校学生提供优质岗位推荐，解决毕业生就业的同时也缓解企业用人难的问题，为校企之间搭建起资源互换的桥梁，力求实现人才的供需对接和教育资源转化，通过深化产教融合，促进教育链、人才链、产业链与创新链有机衔接。

因此，对招聘信息进行分析研究，了解不同职业领域的需求特点，挖掘兴起的数据类行业相应的人才需求现状及发展趋势，为广大求职者提供正确的就业指导有着重要意义。

1.2 问题重述

问题描述：当今时代虽然招聘人才的公司和寻找合适岗位的人才都有很多，但是因为信息闭塞等原因，部分公司无法为岗位找到最合适的人才，而部分求职者也难以找到理想的合适岗位，从而导致好的岗位空缺和人才流失。

问题的解决：对招聘岗位和求职者分别进行信息的爬取收集，在获取到他们各自的数据后，为每一个招聘岗位和求职者建立画像，从岗位类别、工作行业、薪水要求、技术要求等多个特征方面进行分析，通过我们建立的模型得知岗位的匹配度和人才的满意度，最后将匹配率最高的岗位与求职者进行双向推荐，达到公司岗位得到理想人才、求职者找到理想工作岗位的效果。

1.3 相关工作

岗位和求职者信息的爬取：通过 python 代码的编写，实现在泰迪内推的职位招聘页面和人才求职页面获取到大量他们的信息。

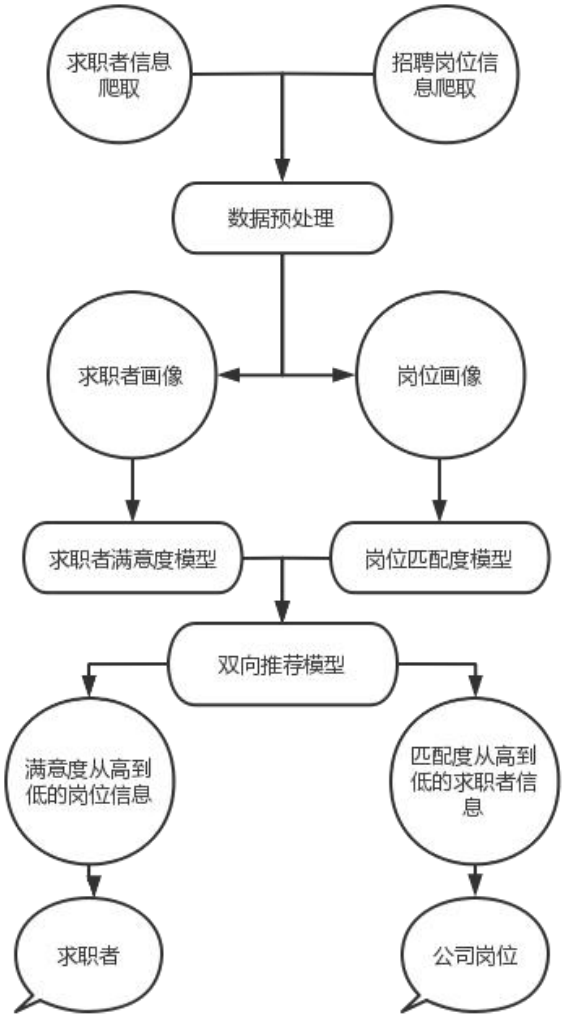
招聘岗位和求职者的信息画像的建立：在获取岗位的企业名称、招聘 ID 等

信息和求职者的工作地点、工作行业、求职者 ID 等信息后，分别用他们的特征信息为他们建立词云画像。

岗位匹配度和求职者满意度模型建立：建立计算出多位求职者与公司所招聘某一岗位的匹配度，以及某一位求职者对多位招聘岗位的各自的满意度的模型。利用该模型，每一个招聘岗位可以获取求职该岗位的人才与之匹配度的信息，每一位求职者也可以获取所求职的岗位的满意度的信息。

招聘求职双向推荐模型的建立：在岗位匹配度和求职者满意度模型的基础上，建立一个为每一个岗位推荐最匹配他们的人才，为每一个求职者推荐能使他们最满意的岗位的模型。

1.4 解决流程



2. 数据信息爬取

在信息爬取的技术方面，我们主要从主页面和详情页面两大板块进行信息的爬取，以此解决爬取到出现在详情页面但不在主页面出现的信息，以及出现在主页面但不在详情面出现的信息，这样可以获取到更多特征信息。

在网页内容的爬取中我们使用了 Selenium 模块。并安装了相应的浏览器驱动对信息进行爬取。

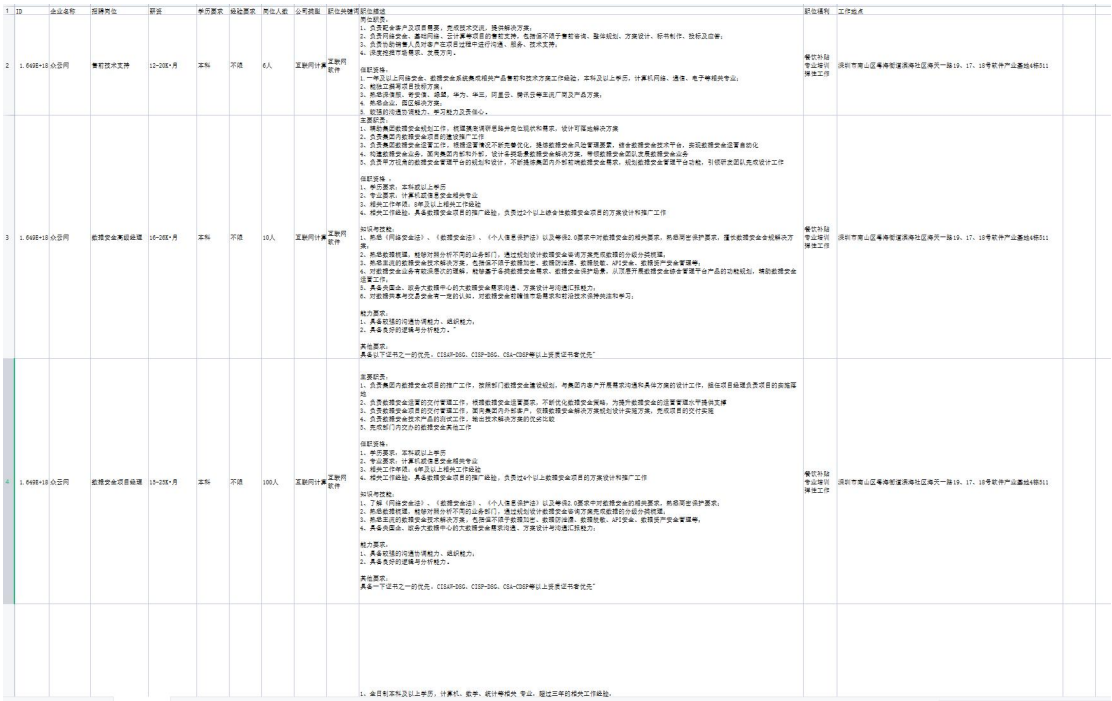
Selenium 是最广泛使用的开源 Web UI（用户界面）自动化测试套件之一。Selenium 支持的语言包括 C+，Java，Perl，PHP，Python 和 Ruby。目前，Selenium Web 驱动程序最受 Python 和 C+欢迎。 Selenium 测试脚本可以使用任何支持的编程语言进行编码，并且可以直接在大多数现代 Web 浏览器中运行。

对于 ID 处理的部分，我们通过读取 json 文件获取 ID 的字典，再用驱动打开相应的网站

无论爬取岗位信息还是求职者信息，我们队伍采用的爬虫技术是调用每一个需要爬取的特征信息的 xpath 路径，提取数据。该技术不仅简单方便，而且所爬取的信息也是十分精准的。

2.1 招聘公司信息爬取

爬取公司岗位招聘的信息，首先进入泰迪内推平台的“找工作”页面（https://www.5iai.com/#/jobList），在该页面我们可以看到多个岗位的招聘信息，本队对岗位的信息爬取包括企业 ID、企业名称、招聘岗位、薪资、学历要求、经验要求、岗位人数、公司类型、职位关键词、职位描述、职位福利和工作地点共 12 个特征信息，我们尽可能多地获取有效的岗位特征信息，以便于后期岗位词云画像的建立和人才匹配。招聘公司信息爬取结果如图：



	ID	企业名称	招聘岗位	薪资	学历要求	经验要求	岗位人数	公司类型	职位关键词	职位描述	职位福利	工作地点
2	1.649E+18	众云网	售前技术支持	12-20K*月	本科	不限	6人	互联网计算软件	互联网软件	岗位职责： 1、负责客户及项目前期售、完成技术交流，提供解决方案； 2、负责网络安全、基础网络、云计算等项目的售前支持，包括但不限于售前咨询、整体规划、方案设计、标书制作、投标及应答； 3、负责协助销售人员对客户在项目过程中进行沟通、报告、技术支持； 4、完成经理下达需求，发现万元。 任职资格： 1、一年及以上网络安全、数据安全系统集成相关产品售前和技术方案工作经验，本科及以上学历，计算机网络、通信、电子等相关专业； 2、能独立编写项目技标方案； 3、熟悉深信服、奇安信、绿盟、华为、华三、阿里云、腾讯云等主流厂商及产品方案； 4、熟悉企业、园区解决方案； 5、能强的沟通协调能力，学习能力及责任心。	餐饮补贴 专业培训 绩生工作	深圳市南山区粤海街道高新社区海天一路19、17、18号软件产业基地4栋311
3	1.649E+18	众云网	数据安全高级经理	16-28K*月	本科	不限	10人	互联网计算软件	互联网软件	岗位职责： 1、熟悉集团数据安全规划工作，根据集团实现研究路并完出现状和需求，设计可落地解决方案 2、负责集团内数据安全项目的建设推广工作 3、负责集团数据安全运营工作，根据运营情况不断完善优化，提高数据安全风险管理质量，结合数据安全技术平台，实现数据安全运营自动化 4、构建数据安全业务，面向集团内部和外部，设计并规划集团数据安全解决方案，带领数据安全团队完成数据安全业务 5、负责学习领先的数据安全管理平台的规划和设计，不断提高集团内外部对数据安全需求，规划数据安全管理平台功能，引领安全团队完成设计工作 任职资格： 1、学历要求：本科或以上学历 2、专业要求：计算机或信息安全相关专业 3、经验工作年限：5年及以上相关工作经验 4、相关工作经验：具备数据安全项目的推广经验，负责过2个以上统合性数据安全项目的方案设计和推广工作 知识与技能： 1、熟悉《网络安全法》、《数据安全法》、《个人信息保护法》以及等保2.0要求中对数据安全的相关要求，熟悉国家保护要求，擅长数据安全合规解决方案； 2、熟悉数据梳理、敏感对象分析不同的业务部门，通过规划设计数据安全标准方案完成数据的分级分类梳理。 3、熟悉至少四款数据安全技术解决方案，包括但不限于数据加密、数据防泄漏、数据脱敏、API安全、数据资产安全管理等。 4、对数据安全业务有较深度的了解，能够基于各项数据安全需求、数据安全保护现象，从顶层开展数据安全综合管理平台产品的功能规划，辅助数据安全运营工作。 5、具备资源、服务大数据中心的大数据安全需求沟通，方案设计与沟通汇报能力。 6、对数据资产与交易安全有一定的认识，对数据安全和隐私计算等新和前沿技术保持关注和学习。 能力要求： 1、具备较强的沟通协调能力，组织能力， 2、具备良好的逻辑与分析能力，” 其他要求： 具备以下证书之一的优先：CISAP-DSG、CISP-DSG、CSA-CSSP等以上资质证书者优先”	餐饮补贴 专业培训 绩生工作	深圳市南山区粤海街道高新社区海天一路19、17、18号软件产业基地4栋311
4	1.649E+18	众云网	数据安全项目经理	15-25K*月	本科	不限	100人	互联网计算软件	互联网软件	岗位职责： 1、负责集团内数据安全项目的推广工作，按照部门数据安全建设规划，与集团内客户开展需求沟通和具体方案的设计工作，组织项目经理负责项目的实施落地 2、负责数据安全运营的交付管理工作，根据数据安全运营要求，不断优化数据安全策略，为提升数据安全的运营管理水平提供支撑 3、负责数据安全项目的交付管理工作，面向集团内外部客户，依据数据安全解决方案规划设计实现方案，完成项目的交付实现 4、负责数据安全技术产品的迭代工作，输出技术解决方案的优化功能 5、完成部门内交办的数据安全类相关工作 任职资格： 1、学历要求：本科或以上学历 2、专业要求：计算机或信息安全相关专业 3、相关工作年限：5年及以上相关工作经验 4、相关工作经验：具备数据安全项目的推广经验，负责过4个以上数据安全项目的方案设计和推广工作 知识与技能： 1、了解《网络安全法》、《数据安全法》、《个人信息保护法》以及等保2.0要求中对数据安全的相关要求，熟悉国家保护要求； 2、熟悉数据梳理，能够对数分析不同的业务部门，通过规划设计数据安全标准方案完成数据的分级分类梳理。 3、熟悉至少四款数据安全技术解决方案，包括但不限于数据加密、数据防泄漏、数据脱敏、API安全、数据资产安全管理等。 4、具备资源、服务大数据中心的大数据安全需求沟通，方案设计与沟通汇报能力。 能力要求： 1、具备较强的沟通协调能力，组织能力， 2、具备良好的逻辑与分析能力， 其他要求： 具备以下证书之一的优先：CISAP-DSG、CISP-DSG、CSA-CSSP等以上资质证书者优先”	餐饮补贴 专业培训 绩生工作	深圳市南山区粤海街道高新社区海天一路19、17、18号软件产业基地4栋311
										1、全日制本科及以上学历，计算机、数学、统计等相关 专业，超过三年的相关工作经验。		

2.2 求职者信息爬取

爬取求职者招聘的信息，首先进入泰迪内推平台的“找人才”页面（https://www.5iai.com/#/moreResume），在该页面我们可以看到多位求职者的求职信息，本队对求职者的信息爬取包括求职者 ID、姓名、性别、年龄、个人简历、工作经验、期望职位、工资薪水、期望行业、到岗时间、工作地区、关键词列表、更新时间和技能奖章共 14 个特征信息，我们尽可能多地获取有效的求职者特征信息，以便于后期求职者词云画像的建立和岗位匹配。求职者信息爬取结果如图：

	ID	姓名	性别	年龄	个人简介	工作经验	期望职位	工资薪水	期望行业	到岗时间	工作地区	关键词列表	更新时间	技能奖章
2	1.65E+18	刘女士	女	22岁	统计专业学	1年工作经	数据分析师	3000-5000	不限	时间待议	湖北省武汉	统计专业	spssmysqlpython	
3	1.649E+18	王先生	男	24岁	1.曾经在学	无经验	数据分析师	5000-7000	不限	随时到岗	广东省广州	市白云区		
4	1.648E+18	特先生	男	1岁	awdaw	1年工作经	数据挖掘工	4000-9000	金融/游戏	1周后到岗	北京市北京	潮男靓仔		
5	1.649E+18	王先生	男	29岁	啥也不会	10年以上工	Hadoop大数	35000-390(	互联网/金	时间待议	台湾省台湾	鸡你太美芝士雪豹		
6	1.649E+18	张先生	男	24岁	白胡子是世	无经验	数据分析师	50000-550(	不限	时间待议	广西壮族自治区	南宁市马山县		
7	1.649E+18	张先生	男	1岁	1	2年工作经	数据分析师	9000-1400(	媒体	时间待议	天津市天津市	和平区		
8	1.649E+18	额先生	男	1岁	1	无经验	自然语言处	4000-7000	互联网	1周后到岗	北京市北京市	东城区		
9	1.649E+18	张女士	女	21岁	本科毕业，	无经验	数据分析师	7000-8000元·月		随时到岗	广东省广州市	天河区		
10	1.649E+18	蔡先生	男	6岁	法律改革	4年工作经	计算机视觉	6000-1000(	网络设备	1周后到岗	福建省莆田市	涵江区		
11	1.648E+18	蔡女士	女	1岁	再看一眼就	6年工作经	其他	3000-6000元·月			山东省菏泽市	曹县		
12	1.648E+18	傻女士	女	25岁	我TM就是一	5年工作经	数据挖掘工	3000-5000	游戏	2周后到岗	北京市北京市	西城区		
13	1.648E+18	练先生	男	1岁	精通...	2年工作经	数据挖掘工	10000-150(	电子商务	随时到岗	江苏省南京市	玄武区		
14	1.647E+18	丽女士	女	31岁	善于交际，	3年工作经	数据挖掘工	7000-1200(	数据服务/	2周后到岗	北京市北京市	东城区		
15	1.647E+18	赵女士	女	24岁	泰迪杯挑战	无经验	数据分析师	3000-6000	医疗健康/	随时到岗	浙江省杭州	泰迪杯挑战赛	二等奖	
16	1.645E+18	林先生	男	23岁	熟悉Pythor	无经验	数据分析师	4000-6000	互联网	时间待议	广东省广州	数据分析	数据挖掘	
17	1.646E+18	中先生	男	1岁	awdawd	无经验	数据分析师	3000-4000	不限	随时到岗	天津市天津市	和平区		
18	1.646E+18	什先生	男	24岁	熟练使用p	2年工作经	机器学习工	3000-5000	互联网	1周后到岗	天津市天津	1		
19	1.645E+18	李先生	男	21岁	非常厉害，	10年以上工	其他	10000-14000元·月			吉林省四平	鸡你太美		
20	1.64E+18	陈先生	男	22岁	掌握pythor	无经验	数据挖掘工	6000-10000	不限/互联	1周后到岗	广东省中山	第七届数维杯大学生数学建模三等奖十届泰迪杯数据挖掘挑战赛三等奖全国大学生数学建模比赛省三等奖		
21	1.627E+18	张先生	男	23岁	熟悉VR游戏	无经验	数据分析师	50000-550(	互联网/游	随时到岗	河南省郑州	深度学习网	页逆向分析	
22	1.644E+18	王先生	男	34岁	熟练掌握p	3年工作经	自然语言处	9000-1300(	互联网/信	随时到岗	广东省广州	阳光开朗大	男孩	
23	1.641E+18	段女士	女	22岁	爱好	3年工作经	数据挖掘工	8000-1000(	金融/电子	1周后到岗	河北省邯郸	你好邯郸		
24	1.64E+18	文女士	女	23岁	无	无经验	数据分析师	9000-1100(	金融/互联	随时到岗	浙江省杭州	市西湖区		
25	1.64E+18	毕先生	男	1岁	啥都会	10年以上工	数据分析师	9000-1400(	互联网	随时到岗	北京市北京市	西城区		
26	1.638E+18	节先生	男	5岁	啥都会	10年以上工	图像处理工	50000-55000元·月			天津市天津市	河西区		

3. 数据预处理

3.1 文字列表元素化处理

3.1.1 公司招聘信息的文字列表元素化处理

公司招聘信息的文字元素化处理主要针对：招聘岗位、职位关键词、工作地点这三大方面，因为这几个特征爬取的信息都是一段文字描述，对于我们后面对人才的期望职位匹配来说没有什么实际的作用，因此我们将这些文字信息化为一个个元素，将他们放入空列表中。由此，后期在匹配过程中的复杂度将大大降低，只需要核对公司招聘的这三大类别的各个列表中的元素与求职者相对应匹类别的各个列表中的元素，从而计算岗位匹配度和求职者满意度。

招聘岗位	职位关键词	工作地点
['售前技术支持']	['互联网', '软件']	['深圳']
['数据安全高级经理']	['互联网', '软件']	['深圳']
['数据安全项目经理']	['互联网', '软件']	['深圳']
['数据挖掘工程师']	['互联网', '大数据']	['山东']
['数据管理']	['互联网', '大数据']	['山东']
['算法工程师']	['互联网', '大数据']	['山东']
['数据标注员']	['大数据', '互联网']	['山东']
['会计实习生']	['互联网', '人工智能']	['广东']
['技术服务工程师']	['互联网', '人工智能']	['广东']
['大数据分析师（BI）']	['互联网', '软件']	['广州']
['自然语言处理工程师']	['自然语言']	['广州']
['爬虫工程师']	['Python', '爬虫框架']	['广州']
['Python', 'django后台开发工程师']	['Python', 'django', '后台开发']	['广州']
['团队运营管理实习生']	['运营']	['广州']
['技术文章撰写实习生']	['技术文章撰写']	['广州']
['创新俱乐部成员']	['产品设计']	['广州']
['解决方案']	['互联网', '软件']	['湖北']
['产品总监']	['互联网', '软件']	['湖北']
['大数据架构师']	['互联网', '软件']	['湖北']
['大数据系统分析师']	['互联网', '软件']	['湖北']
['AI类数据标注']	['互联网', '软件']	['四川']
['大数据开发工程师']	['互联网', '软件']	['四川']
['数据产品经理']	['互联网', '软件']	['四川']
['数据工程师']	['互联网', '软件']	['四川']
['高级大数据开发工程师 Hadoop Hive']	['互联网', '广告']	['北京']
['数据标注工程师']	['互联网', '大数据']	['福州']
['数据挖掘工程师']	['互联网', '大数据']	['广州']
['python工程师']	['互联网', '大数据']	['天河']
['数据实施实习生']	['互联网', '软件']	['广东']

3.1.2 求职者信息的文字列表元素化处理

求职者信息的文字元素化处理主要针对：期望职位、期望行业、工作地区这三大方面，原理同公司招聘信息的文字元素化处理一样。

期望职位	期望行业	工作地区
['数据分析师', '数据挖掘工程师'	['互联网'	['北京']
['数据分析师', '其他']	['不限']	['湖北']
['自然语言处理工程师', '图像处理	['游戏', '	['广西']
['其他', '机器学习工程师']	['不限', '	['湖南']
['数据分析师', '数据挖掘工程师'	['不限']	['广东']
['数据挖掘工程师', '图像处理工程	['金融', '	['北京']
['Hadoop大数据开发工程师', '其他	['互联网',	['台湾']
['数据分析师', '数据挖掘工程师'	['不限']	['广西']
['数据分析师']	['媒体']	['天津']
['自然语言处理工程师']	['互联网'	['北京']
['数据分析师']	['不限']	['广东']
['计算机视觉工程师']	['网络设备	['福建']
['其他']	['不限']	['山东']
['数据挖掘工程师', '算法工程师'	['游戏']	['北京']
['数据挖掘工程师', '数据分析师'	['电子商务	['江苏']
['数据挖掘工程师', '自然语言处理	['数据服务	['北京']
['数据分析师']	['医疗健康	['浙江']
['数据分析师', '数据挖掘工程师'	['互联网'	['广东']
['数据分析师']	['不限']	['天津']
['机器学习工程师', '数据挖掘工程	['互联网'	['天津']
['其他']	['不限']	['吉林']
['数据挖掘工程师', '数据分析师'	['不限', '	['广东']
['数据分析师', '数据挖掘工程师'	['互联网',	['河南']
['自然语言处理工程师', '数据分析	['互联网',	['广东']
['数据挖掘工程师', '图像处理工程	['金融', '	['河北']
['数据分析师']	['金融', '	['浙江']
['数据分析师', '数据挖掘工程师'	['互联网'	['北京']
['图像处理工程师', '计算机视觉工	['不限']	['天津']

3.2 信息数字化处理

3.2.1 学历信息数字化处理

不同的公司的不同岗位对求职者的学历要求不同，在这里为了统一标准，无论是对公司的招聘信息还是求职者的信息，我们都将学历要求分为 6 大类：不限：0，大专：1，本科：2，硕士：3，博士：4，技工：5。为了方便后期的匹配计算，我们将每一类用 0-6 的数字来划分类别，在后期的计算岗位匹配度和求职者满意度时，两者数字核对相同将增大匹配度，相反则会降低匹配度。而且，数字的匹配核对比文字的核对将更加精确高效。学历要求数据处理如图：

学历要求
2
2
2
2
2
2
1
1
1
2
2
2
2
2
2
0
2
2
2
2
1
2
2
2

3.2.2 最低薪资和最高薪资的数字化处理

对我们的公司招聘岗位信息还是求职者信息，给出的薪资和求职者的期望薪资都是一个区间范围，区间范围在后期的匹配计算中是不方便的，难以进行。因此，在这里我们将薪资范围一分为二：最低薪资和最高薪资。在匹配时，将公司招聘的最低薪资、最高薪资分别同求职者的最低期望薪资和最高期望薪资比较，达到适配度的计算。

最低薪资	最高薪资
12000	20000
16000	26000
15000	25000
5000	8000
3500	6000
5000	8000
2500	7000
3500	7000
5000	10000
4500	7000
2000	4000
2000	4000
2000	4000
2000	4000
2000	4000
2000	4000
30000	40000
20000	35000
25000	50000
25000	40000
3500	8000
3500	8000
3500	8000
3500	8000
20000	35000
6000	8000
6000	8000
4500	6000

最低薪资	最高薪资
50000	55000
3000	5000
3000	8000
50000	55000
5000	7000
4000	9000
35000	39000
50000	55000
9000	14000
4000	7000
7000	8000
6000	10000
3000	6000
3000	5000
10000	15000
7000	12000
3000	6000
4000	6000
3000	4000
3000	5000
10000	14000
6000	10000
50000	55000
9000	13000
8000	10000
9000	11000
9000	14000
50000	55000

3.1.3 经验要求、岗位人数和工作经验数字化处理

这里处理的数据包括的是公司招聘信息的经验要求、岗位人数和求职者信息的工作经验，因为这些特征信息爬取的数据都是数字+文字单位，这里为了简洁和后期匹配的高效性，我们在各个类别统一单位后，直接将单位去掉，仅留下数字。

经验要求	岗位人数	1	工作经验
0	6	2	2
0	10	3	1
0	100	4	10
0	5	5	2
0	6	6	0
0	10	7	1
0	100	8	10
0	2	9	0
0	5	10	2
0	1	11	0
0	0	12	0
0	0	13	4
0	0	14	6
0	0	15	5
0	0	16	2
0	0	17	3
0	1	18	0
0	1	19	0
0	1	20	0
6	2	21	2

3.3 公司 ID 和求职者 ID 数据处理

每一个公司职位页面和求职者信息详情页面均有属于他们各自的 ID，在获取这些 ID 后，我们使用了 drop_duplicates（）方法，达到去重的目的，避免一个岗位或求职者多次重复出现，让 ID 信息更加准确规范化。

3.4 处理思路分析

对 result1-1.csv 进行清洗和处理：
去除重复的招聘信息 id；
对于空值或无意义的数据进行删除或替换；
对员工数量、学历、岗位经验等分类数据进行数字化处理，便于后续计算。
对 result1-2.csv 进行清洗和处理：
去除重复的求职者 id；
对于空值或无意义的数据进行删除或替换；

将预期岗位和技能进行关键词提取和分词处理，便于后续计算。

计算岗位匹配度：
对于每条招聘信息，根据预期岗位和技能对求职者进行筛选，筛选出符合条件的求职者；
对于每个符合条件的求职者，计算其与招聘信息的岗位匹配度，可以使用技能匹配度和职位匹配度等指标进行计算；
对于每条招聘信息，将所有符合条件的求职者的岗位匹配度求和，得到该招聘信息的总岗位匹配度。

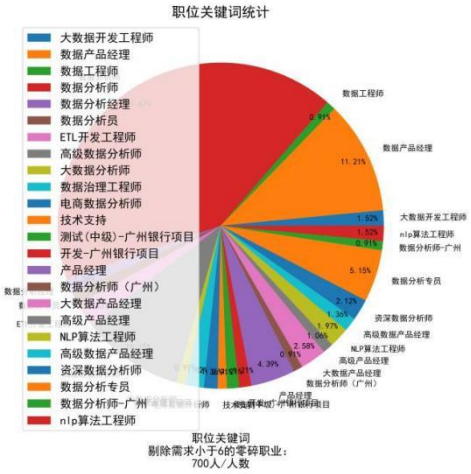
计算求职者满意度：
对于每位求职者，根据学历、岗位经验等条件对招聘信息进行筛选，筛选出符合条件的招聘信息；
对于每个符合条件的招聘信息，计算其与求职者的求职者满意度，可以使用薪资、公司类型、工作地点等指标进行计算；
对于每位求职者，将所有符合条件的招聘信息的求职者满意度求和，得到该求职者的总求职者满意度。

4. 招聘与求职信息分析

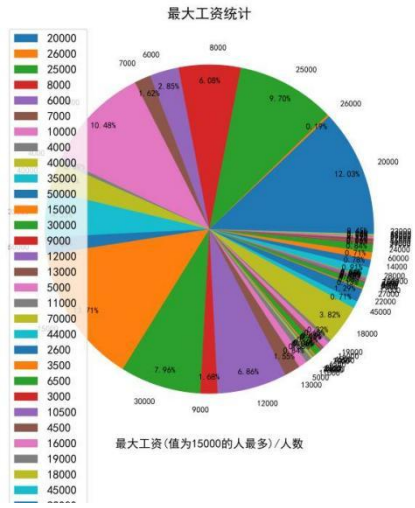
4.1 招聘信息画像的建立

为了能够更好地帮助求职者了解到公司在市场上对人才需求的倾向性，我们建立了招聘信息的画像，更加直观地表现公司招聘的特征信息，让求职者一目了然。本队使用词云建立招聘信息画像，效果如图：

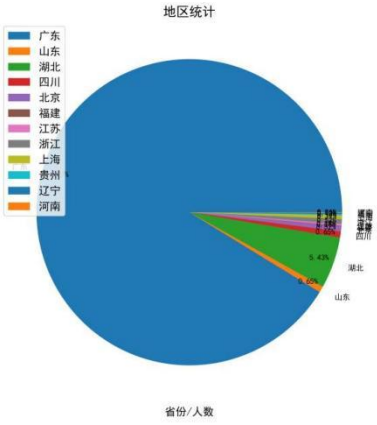
从词云图中，我们可以很直观地了解到公司的学历要求，给出的工作薪资，岗位，招聘人数等等众多信息，对于找工作的求职者来说非常直观。



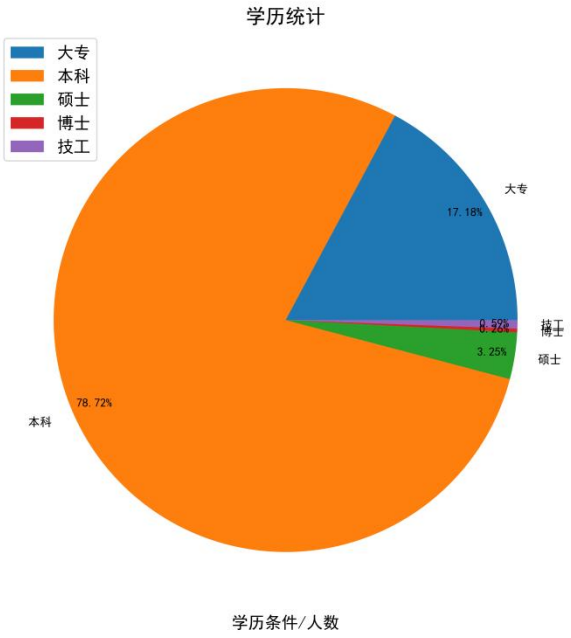
以上是岗位关键词的饼图。从该饼图成分分析，招聘的岗位类型主要集中在数据分析。其中数据分析师的岗位占比最多。主要的行业是数据分析处理，大数据的应用.



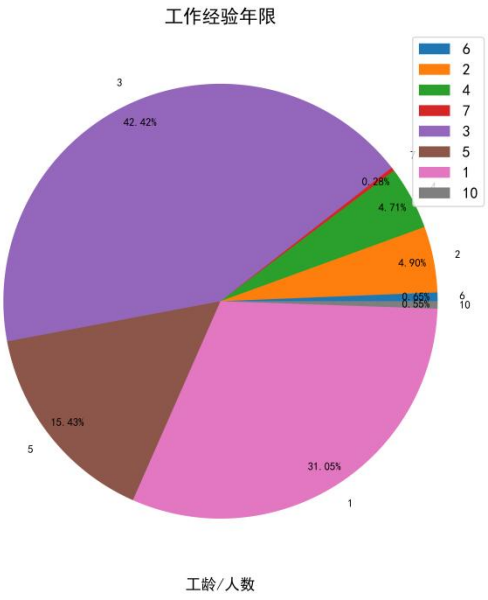
以上是企业最高工资饼图。从统计数据来分析，各个公司岗位的最大薪资分布较为平均，差异较大。其中 15000 元/月的占比最大。占比 13%



以上是企业的工作地区，从该饼图成分分析，企业的工作地区主要集中于广东省。第二个集中点在湖北省。南方城市居多



从该企业的学历要求统计可以看出，学历要求最多的是本科学历，大专和本科学历要求占据了绝大多数，也有公司岗位对学历有像硕士、博士这样的更加专业、更加高精尖的学历要求



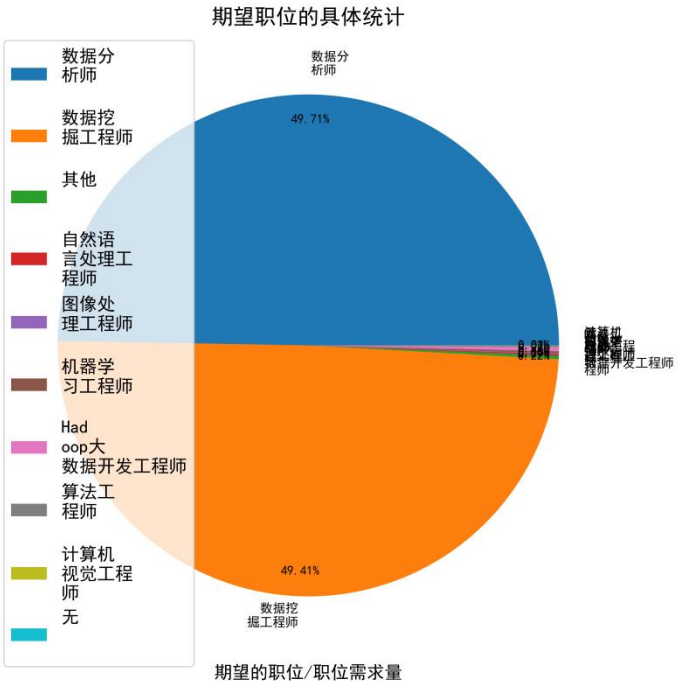
从企业对求职者工作经验年限的统计图可以看出，工作经验为三年的占比最多其次是一年的和五年工作经验的，可以发现，公司岗位更倾向于有工作经验者，且要求工作经验较为丰富，但同样，市场对工作经验较少者容纳度较高，也有大量的工作岗位愿意招聘工作经验相对较少者

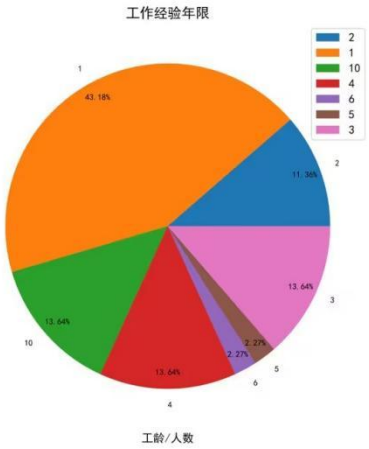
从该求职者的期望职位数量的统计图中可以看出，大部分求职者的期望职位数量为 2 个，期望职位数量为 1 个和 3 个以上的占少数，由此分析可见，求职者同时求职多个职位，这样将提高求职的成功率，让更多求职者找到适合自己的岗位

4.2 求职者信息画像的建立

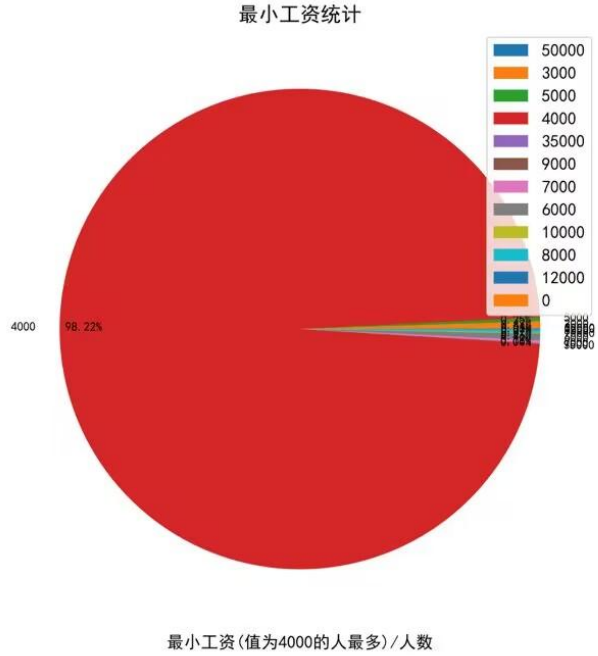
同样地，为了能够更好地帮助公司了解求职者个人信息、技能，我们建立了求职者信息的画像，更加直观地表现求职者的特征信息，让公司岗位招聘人员一目了然，更加快速了解每一位求职者。本队使用词云建立求职者信息画像，效果如图：

从词云图中，我们可以很直观地获取到每一位求职者的多个个人信息，包括性别、年龄、期望职位、期望行业等等众多信息，对于找招聘人才的公司来说非常直观。

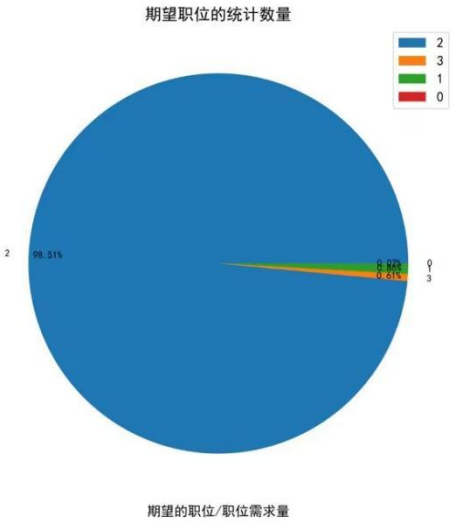




该图为求职者的工作经验年限图，可以从该图的分布看出，工作年限只有一年的求职者占大多数，占比百分之四十多，工作年限为 2、3、4 和 10 年的占比基本相同，均为百分之十几，工作年限为 5、6 年的占比为最少的两位。由此可见，在招聘市场上，大多数为工作经验较少的求职者



从该求职者的最小工资统计图可以发现，最小工资为 4000 元/月的求职者占绝大多数，其次是 3000 元/月的求职者，而最小工资达到 10000 元以上的占极少数，因此可以预测市场上最小工资的平均数应为 6000 元左右，该项数据可以作为公司岗位给出最低薪资的重要参考以及建议求职者在提出底薪要求时的重要参考。



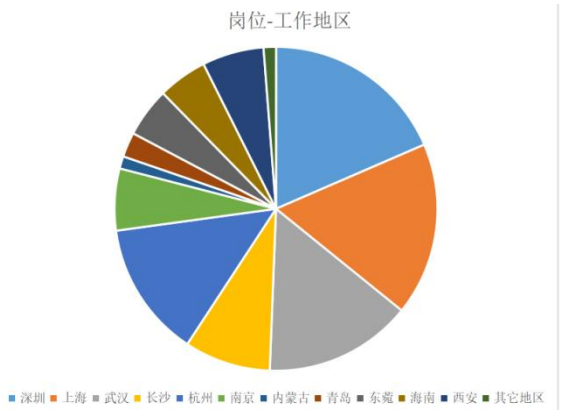
从该求职者的期望职位数量的统计图中可以看出，大部分求职者的期望职位数量为 2 个，期望职位数量为 1 个和 3 个以上的占少数，由此分析可见，求职者

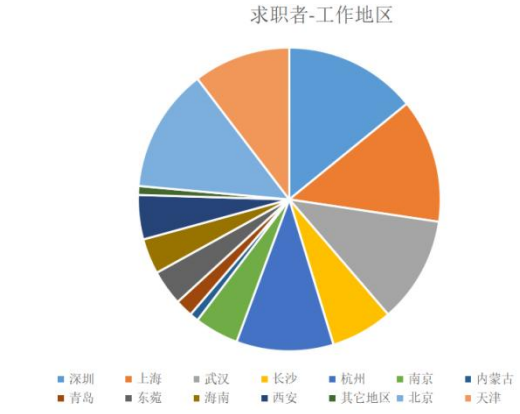
同时求职多个职位，这样将提高求职的成功率，让更多求职者找到适合自己的岗位

4.3 招聘市场趋势分析

4.3.1 工作地区数据分析

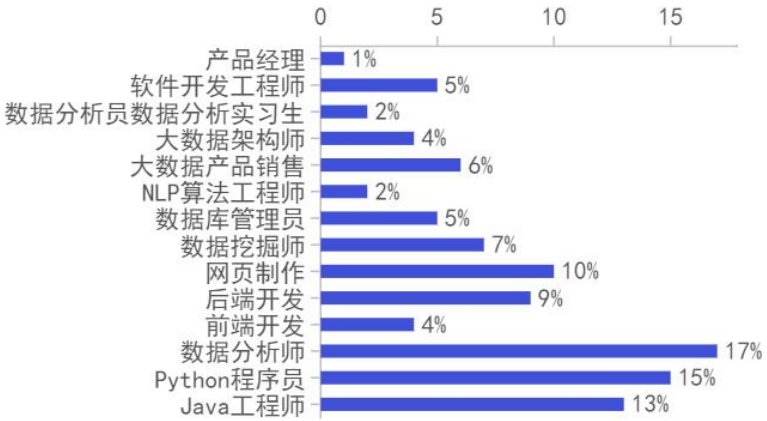
本队为公司工作岗位和求职者期望的工作地区的数据制作了如下图的两张饼状图，可以发现，对于公司和求职者来说，深圳是选择工作最多的地方，因此对于在深圳招聘或者求职的公司和人来说，成功率会更高一些。同样，上海、武汉、杭州、长沙等这些也都是招聘人才和求职者找工作的热门城市。





4.3.2 工作职位数据分析

我们将相对比较热门的十几个岗位的数据拿来分析，制作了如下图所示的条形图，从该图的数据分析中，我们可以直观清晰地了解到市场上哪些岗位招聘人才的需求量最大，从而更好地对市场的人才需求进行预测。



4.4 模型评估及优化

在这里，本队只对工作地区和热门职位进行了市场的预测分析，除此之外，还可以对岗位的学历要求，求职者的学历要求，公司规模，工作薪资等其他方面加以综合分析，这样会更有利于更加精确地对招聘求职市场进行预测。

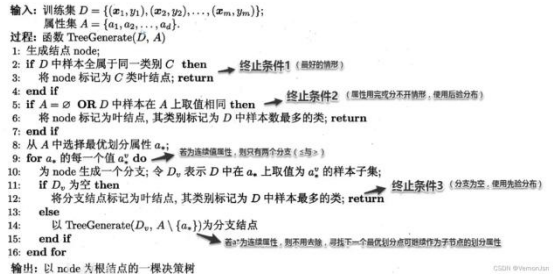
其中，使用决策树算法建立模型，并使用了 sklearn 库中的 DecisionTreeClassifier 等模块，对求职者们的求职岗位、求职地区等进行了预测，并对该模型进行评估。首先对岗位、经验、地区等属性进行了 One-hot 编码处理，并且划分训练集和测试集，使用决策树算法建立相应的模型，对数据进行训练。

4.5 应用模型介绍

决策树是一种分类和回归的监督学习算法。它将数据集分成不同的决策区域，每个决策区域对应一个分类或回归输出。在决策树中，每个内部节点表示一个属性或特征，每个叶子节点表示一个分类或回归输出。其训练过程是递归的，它从根节点开始，根据一个属性或特征将数据集分成不同的子集，然后对每个子集递归地应用相同的过程，直到所有数据都被分类或回归为止。

决策树算法有多种不同的实现，如 ID3、C4.5、CART 等。在训练过程中，这些算法采用不同的方法选择最佳属性或特征来划分数据集，例如信息增益、信息增益比、基尼指数等。

决策树算法具有易于理解、可解释性强、适用于高维数据等优点。它也可以处理缺失数据和噪声数据。然而，决策树算法容易过拟合，需要进行剪枝等预处理技术以提高其泛化能力



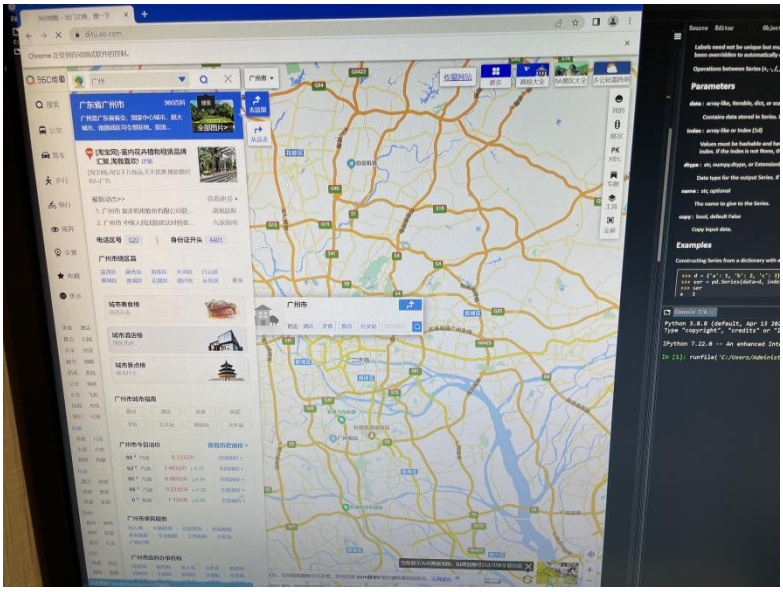
5. 岗位匹配度和求职者满意度模型的构建

5.1 数据预处理

岗位匹配度和求职者满意度的计算过程中，使用到的数据包括岗位的公司 ID、学历要求、工作经验要求、工作地区、最低薪资和最高薪资、招聘岗位和职位关键词，求职者的求职者 ID、学历、工作经验、期望行业、预期最低薪资和

预期最高薪资、工作地区。

特别的，在预处理的过程中，为了针对企业地址模糊，难以确定省份的问题，我们借助 selenium 来驱动浏览器访问 360 地图，并自动搜索爬取了模糊地址对应的详细省份。



公司 ID 和求职者 ID 均采用 ASCII 码统一格式；公司的学历要求和求职者学历统一用 0-5 将“不限”、“大专”、“本科”、“硕士”、“博士”、“技工”六个类别；将公司的经验要求和求职者的工作经验统一用 0-10 将工作经验分为“不限”、“1 年”、“1-3 年”、“3-5 年”等分为八个类别。 将岗位最低薪资、最高薪资和求职者的预期最低薪资、预期最高薪资均以元为单位，然后去掉单位文字，仅留下数字；工作地区以元素的形式放入列表。

数据处理完成,将求职者和招聘信息用 pandas 的 merge 方法创建笛卡尔积，并且将两张数据表合并，这将大大减少后续数据运行的时间复杂度。

5.2 评分模型的建立

本队建立的两个评分模型：

公司匹配度计算模型：匹配度 = 【学历 *0.3* 额外权值 + 经验 *0.3 * 额外权值 + 薪资 *0.3 +行业 * 0.1】* 岗位；

求职者满意度计算模型：满意度 = 【学历 * 0.1 * 额外权值 + 经验 * 0.1 * 额外权值 + 薪资 *0.4 +行业 * 0.4】* 岗位；

在招聘和求职的过程中，企业和求职者首先关注的应该是所对应的岗位，所以在模型中我们将岗位定为决定匹配度/满意度是否存在的重要因素。

在评分之前，我们将所有的评分项的分数均初始为 0.0，在将两者的各种对应特征信息对比，例如：遍历求职者的期望职位的列表，如果能够找到与公司职位关键词列表中相同的元素，那么该项评分为 1，反之得分为 0。得出该类别的评分为 0 还是 1，再带入到对应的匹配度计算模型或满意度计算模型，乘于分配的权重比例，就可以得到该项的评分，如果该企业/求职者的条件高于要求的阈值，将适当为该项添加额外的值数。当计算完所有类别的权重评分后，最后进行相加，即可得到岗位匹配度或求职者满意度。

5.3 匹配度和满意度结果的输出

公司岗位匹配度：对每一个公司进行多位求职者的匹配度计算，然后以降序的顺序排列，对每一个公司岗位均进行该操作。输出结果如图：

1	企业ID	求职者ID	匹配度
2	1.648E+18	1.648E+18	0.66
3	1.648E+18	1.648E+18	0.45
4	1.648E+18	1.648E+18	0.36
5	1.648E+18	1.647E+18	0.39
6	1.648E+18	1.64E+18	0.3
7	1.648E+18	1.641E+18	0.72
8	1.648E+18	1.462E+18	0.76
9	1.648E+18	1.529E+18	0.3
10	1.648E+18	1.462E+18	0.7
11	1.648E+18	1.462E+18	0.66
12	1.648E+18	1.484E+18	0.52
13	1.648E+18	1.501E+18	0.4
14	1.648E+18	1.476E+18	0.33
15	1.648E+18	1.469E+18	0.3
16	1.648E+18	1.638E+18	0.52
17	1.648E+18	1.463E+18	0.6
18	1.613E+18	1.65E+18	0.6
19	1.613E+18	1.649E+18	0.3
20	1.613E+18	1.644E+18	0.39

求职者满意度：对每一位求职者进行多个岗位的满意度计算，然后以降序的顺序排列，对每一位求职者均进行该操作。输出结果如图：

求职者ID	企业ID	满意度	企业名称
1.65E+18	1.65E+18	0.22	森羽网络
1.65E+18	1.65E+18	0.15	森羽网络
1.65E+18	1.65E+18	0.12	森羽网络
1.65E+18	1.65E+18	0.13	森羽网络
1.64E+18	1.65E+18	0.1	森羽网络
1.64E+18	1.65E+18	0.24	森羽网络
1.46E+18	1.65E+18	0.22	森羽网络
1.53E+18	1.65E+18	0.1	森羽网络
1.46E+18	1.65E+18	0.2	森羽网络
1.46E+18	1.65E+18	0.22	森羽网络
1.48E+18	1.65E+18	0.14	森羽网络
1.50E+18	1.65E+18	0.1	森羽网络
1.48E+18	1.65E+18	0.11	森羽网络
1.47E+18	1.65E+18	0.1	森羽网络
1.64E+18	1.65E+18	0.14	森羽网络
1.46E+18	1.65E+18	0.2	森羽网络
1.65E+18	1.61E+18	0.2	奇之
1.65E+18	1.61E+18	0.1	奇之
1.64E+18	1.61E+18	0.13	奇之
1.65E+18	1.60E+18	0.22	极能信息
1.65E+18	1.60E+18	0.15	极能信息
1.65E+18	1.60E+18	0.12	极能信息
1.65E+18	1.60E+18	0.13	极能信息
1.64E+18	1.60E+18	0.1	极能信息

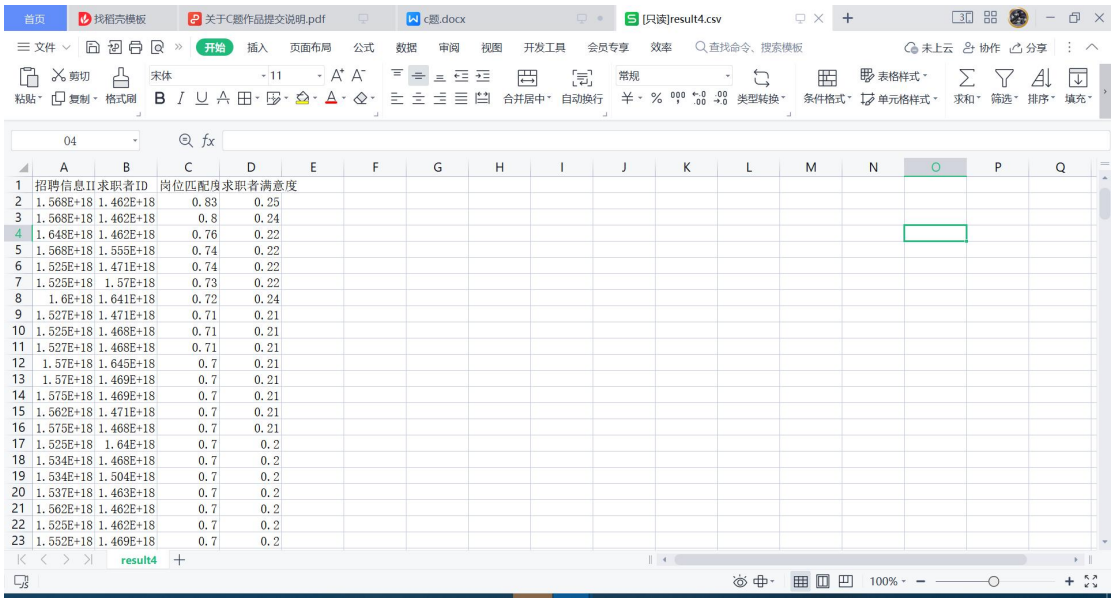
6.招聘求职双向推荐模型的建立

6.1 岗位匹配的计算分析及模型

读取 result3-1.csv 和 result3-2.csv 两个文件，合并招聘信息和求职者信息，使用交叉连接的方式，得到一个包含所有可能匹配的招聘信息和求职者的数据表。根据相应的匹配度和满意度。将求职者按照招聘信息的剩余岗位数进行排序，依次向求职者发出 offer。对于每个招聘信息，选取接受 offer 的求职者中满意度最高的一个，作为该招聘信息的签约人。计算所有岗位的签约人数之和和所有拟聘岗位人数之和，得到履约率指标。

6.2 岗位和求职者的选择算法分析及模型

快速排序是一种高效的排序算法，它的基本思想是选取一个基准元素，将数组分为两个子数组，一个子数组中的所有元素都比基准元素小，另一个子数组中的所有元素都比基准元素大，然后对这两个子数组递归地进行排序。基于快速排序算法使用两个字典对每个招聘信息还需招聘的人数和每个求职者已经接到的 offer 数，使用列表对签约结果的集合进行统计。构建配对模型，通过对 dataframe 进行遍历，将求职者-招聘信息一一对应，进行签约。同时统计剩余的岗位人数，来判定流程是否结束，进入下一个循环。



招聘信息II	求职者ID	岗位匹配度	求职者满意度
1.568E+18	1.462E+18	0.83	0.25
1.568E+18	1.462E+18	0.8	0.24
1.648E+18	1.462E+18	0.76	0.22
1.568E+18	1.555E+18	0.74	0.22
1.525E+18	1.471E+18	0.74	0.22
1.525E+18	1.57E+18	0.73	0.22
1.6E+18	1.641E+18	0.72	0.24
1.527E+18	1.471E+18	0.71	0.21
1.525E+18	1.468E+18	0.71	0.21
1.527E+18	1.468E+18	0.71	0.21
1.57E+18	1.645E+18	0.7	0.21
1.57E+18	1.469E+18	0.7	0.21
1.575E+18	1.469E+18	0.7	0.21
1.562E+18	1.471E+18	0.7	0.21
1.575E+18	1.468E+18	0.7	0.21
1.525E+18	1.64E+18	0.7	0.2
1.534E+18	1.468E+18	0.7	0.2
1.534E+18	1.504E+18	0.7	0.2
1.537E+18	1.463E+18	0.7	0.2
1.562E+18	1.462E+18	0.7	0.2
1.525E+18	1.462E+18	0.7	0.2
1.552E+18	1.469E+18	0.7	0.2

为了优化模型，我们定义了火热度，它可以衡量一个职位是否被多人选择，当一个职位被很少的人选择时，证明他的 offer 很容易被忽略，所以我们优先考虑他。

6.3 双向推荐模型对于未来应用的一些展望

对于该模型的优化，我们还可以实现通过输入求职者信息，来对改求职者可能会前往的公司、所从事的岗位进行预测。基于决策树算法，将员工信息划分训练集和测试集。导入机器学习模块，对求职者的信息处理分析，建立关系矩阵对数据进行训练。从而对其未来的去向进行预测

总结

这个项目是一个非常有挑战性的过程，需要掌握一系列的技能和知识，同时也需要具备一定的耐心和毅力。在我参加数据挖掘比赛的过程中，我遇到了很多的困难和挑战，但也获得了很多的收获和成长。

首先，在数据的爬取的过程中，我们尝试过 request、selenium、urllib 等模块来爬取数据，期间出现了各种各样的问题，比如接口不对，数据丢失等问题，最终我们完整学习了 selenium 模块的爬虫技术，采用 css，Xpath 等组件爬取所需要的信息。尤其在对于 ID 处理的部分，我们通过读取 json 文件获取 ID 的字典，再用驱动打开相应的网站。

另外，数据的理解和预处理是非常关键的。在比赛中，我们需要对数据进行详细的分析，包括数据的缺失情况、异常值处理等等。只有对数据进行了充分的理解和预处理，才能正确地对数据进行训练和预测。

其次，比赛让我接触了很多不同的数据挖掘算法和技术。通过参加比赛，我学习了很多新的算法和技术，如协同过滤算法等。同时，我也发现了一些常见算法的优缺点，如决策树、随机森林等。这些经验对我的数据挖掘实践非常有帮助。

在模型的建立中，我们参考了协同过滤算法，制定相应的满意度和匹配度的模型，将 dataFrame 转换为笛卡尔积并进行合并处理。编写相应的函数。

最后，模型的建立和优化也是我们需要重点关注的。建立双向匹配的模型时，将求职者按照招聘信息的剩余岗位数进行排序，依次向求职者发出 offer，计算所有岗位的签约人数之和和所有拟聘岗位人数之和，得到履约率指标。

总的来说，参加数据挖掘比赛是一项非常有挑战性的工作，需要我们不断学习和实践，掌握数据处理、特征工程和模型建立等技能，才能在比赛中取得好的成绩。在这个过程中，我们也会获得很多的收获和成长，提高自己的数据分析能力和解决问题的能力。比赛也让我认识到了团队合作的重要性。在比赛中，我与队友合作，共同研究问题，分析数据，实现算法，并且不断调整和改进。

参考文献

[1]周志华，《机器学习》，清华大学出版社.
[2]Pang-Ning Tan,Michael Steinbach,Anuj Karpatne,Vipin Kumaar,《数据挖掘导论》，机械工业出版社