服创赛 A15 当前进度汇总

2024.2.25

一、算法

1.数据预处理

根据企业提供的数据集，共有 8763 条计算机类/电子类招聘相关的数据，来源于 boss 直聘，对所给数据进行预处理 (1)id:有用。 (2)公司名称：有 200 多个显示为"gongsi",应置为空 (3)职位名称：可进一步分类，提取出是否实习（实习、全职）、职位类别（前端、后端、运维等） (4)薪资：(最低-最高)*薪数、面议 (5)学历要求：大专、本科、硕士、博士 (6)职位描述：非结构化数据 (7)发布人：没有用，删除(8)最后活跃：没有用，删除 (9)工作地点：提取城市，有一个职位缺失：手动补充 (10)网址：没有用，删除 预处理后的数据中包含 id、公司名称、城市、职位名称、职位类别、薪资、学历要求、是否实习等结构化数据，可直接用于职位知识图谱的构建，以及职位描述等非结构化数据，需要对其进行实体识别与关系抽取。

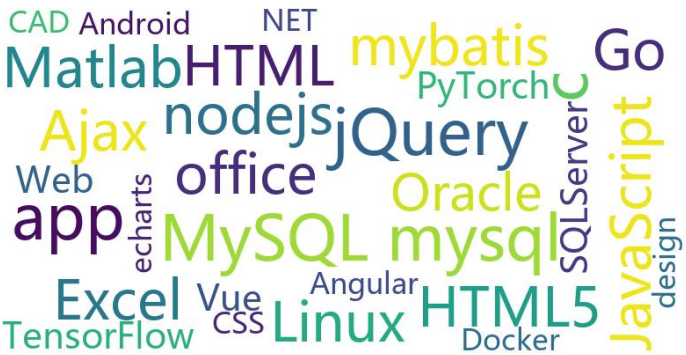
2.本体构建与知识抽取

对于非结构化数据，抽取“专业能力”、“专业知识”、“个人素养”作为顶层概念。其中，“专业能力”主要包含专业工具使用、计算机技能、英语水平等；“专业知识”主要包含专业类型要求和专业知识要求；“个人素养”主要包含学习能力、进取精神、社交能力、工作态度等。

**构建语料库

步骤：Jieba 分词——统计词频——人工选取词频大于 20 的加入语料库。

（1）技能（210 个）



（2）知识（29 个）

（3）素养（30 个）

*（对于上面三类实体的抽取，按理说应该使用 bert-bilstm-crf 模型对使用 BIO 标注的文本进行训练，但考虑到缺少这类数据集，且手工标注极为费力，并且即使标注出来进行训练得到的模型效果也未必较好，另外本题的重点是推荐，此处不应作为重点，所以最终选定“关键词法”提取上述三类实体。由于我们已经选定了计算机类职位，所以只要我们词库的词足够多，就能对大部分职位描述进行提取）
手工标注示意：

能词。技能词可以看成是特定专业领域内的命名实体[2]或者术语[3]。因此，网络招聘文本技能词的抽取任务可以借鉴命名实体识别或者术语抽取的方法。目前，虽然命名实体识别的相关工作很多，但重点都是在识别正式文本中的人名，地名和机构名，而术语抽取的相关工作主要针对特定领域内的术语识别，缺乏通用性和可移植性。其次，大多数命名实体识别或者术语抽取研究均采用人工规则或传统的机器学习方法，基于一个学科或一个领域进行抽取，需要人工设计规则进行特征提取。近年来，许多相关研究也将命名实体识别或术语抽取任务转化为序列标注问题，而最新技术主要基于深度神经网络进行端到端训练以捕获上下文信息。这类方法可以将输入字符转换为输出标签，而无需显式的人工设计特征提取。但是，基于深度学习的序列标注方法仅专注于域内监督学习，这需要大量带标注的数据。对于网络招聘数据而言，人工标注数据既费时又昂贵，只能依靠领域专家手工标注少量语句。因此，很难获得足够的带标注的语句来训练深度神经网络。如何使用少量标注数据来快速、准确地抽取技能词是非常具有挑战的。由此可见，中文网络招聘文本中的技能词抽取研究仍然是一项颇为艰巨的，也是非常具有价值的任务。

的适当水平的能力来表征每个大数据工作族。在国内，詹川[42]根据已有的关于电子商

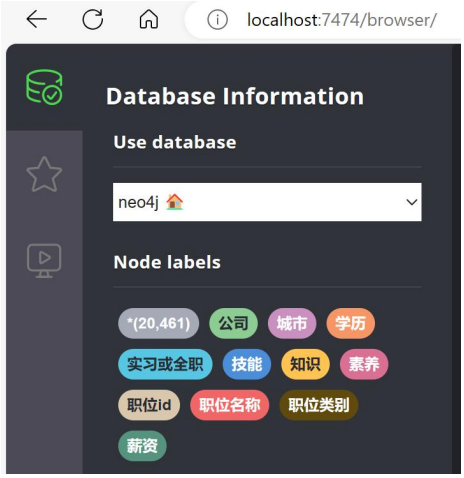
务行业的专业术语构建出该专业的技能词词典，对采集的 66925 条电子商务岗位的招聘信息进行关键词抽取，并对抽取出的高频技能关键词进行分类，分析出电子商务行业中不同类型岗位的通用技能需求和特定岗位技能需求。俞琰[43]等人从前途无忧招聘网站中抓取了 10000 条计算机领域和 30000 条非计算机领域的招聘信息，利用依存

3.知识存储

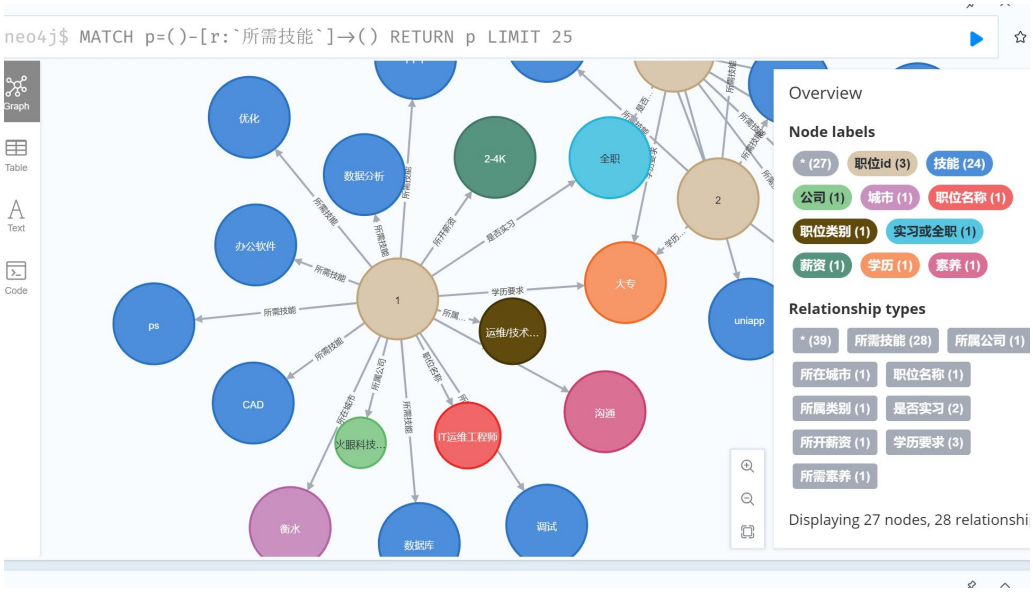
关系抽取完成后，得到一系列具有节点、关系和属性的三元组,使用 Neo4j 图数据库进行实体和关系存储以及可视化。

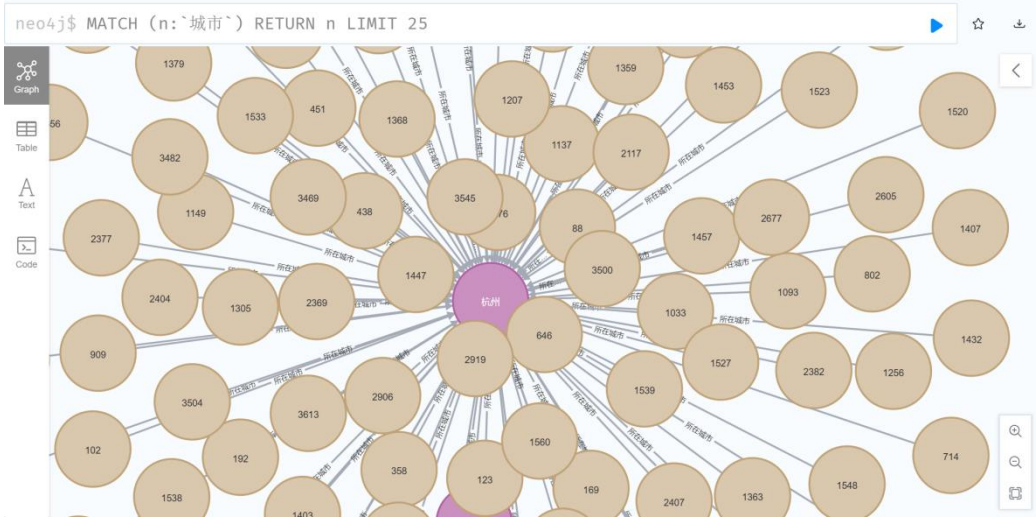
Window+R——>Cmd——>输入 neo4j.bat console

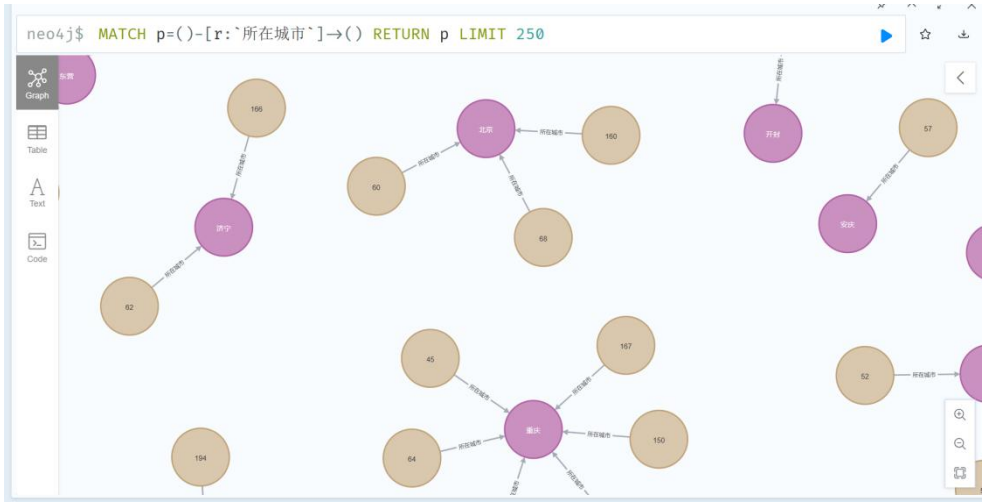
Neo4j 知识图谱概览：对于题目所给数据集，最终得到 2w 多个实体，13w 多个关系。

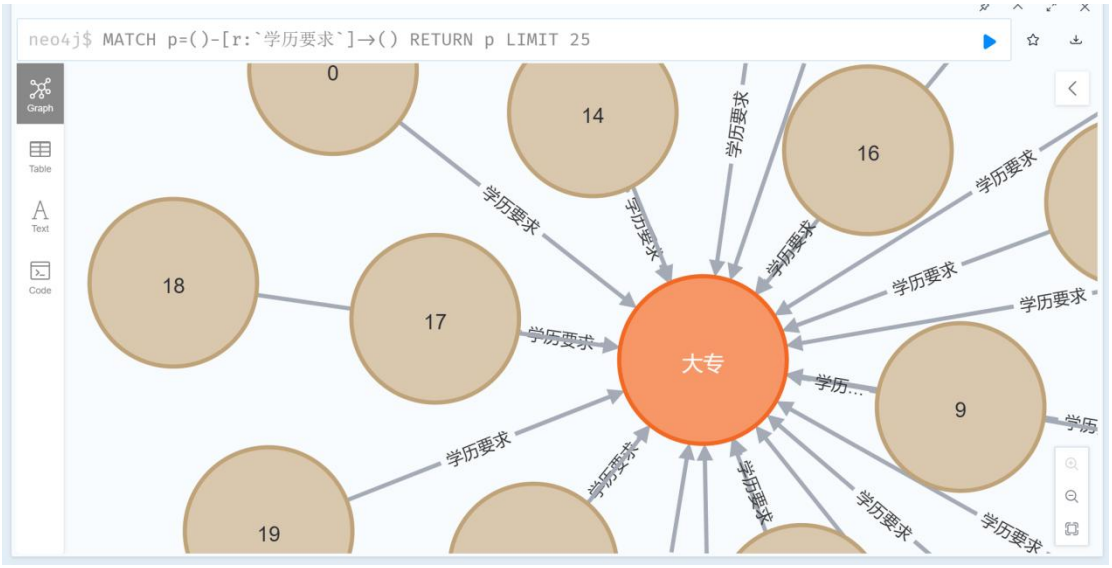


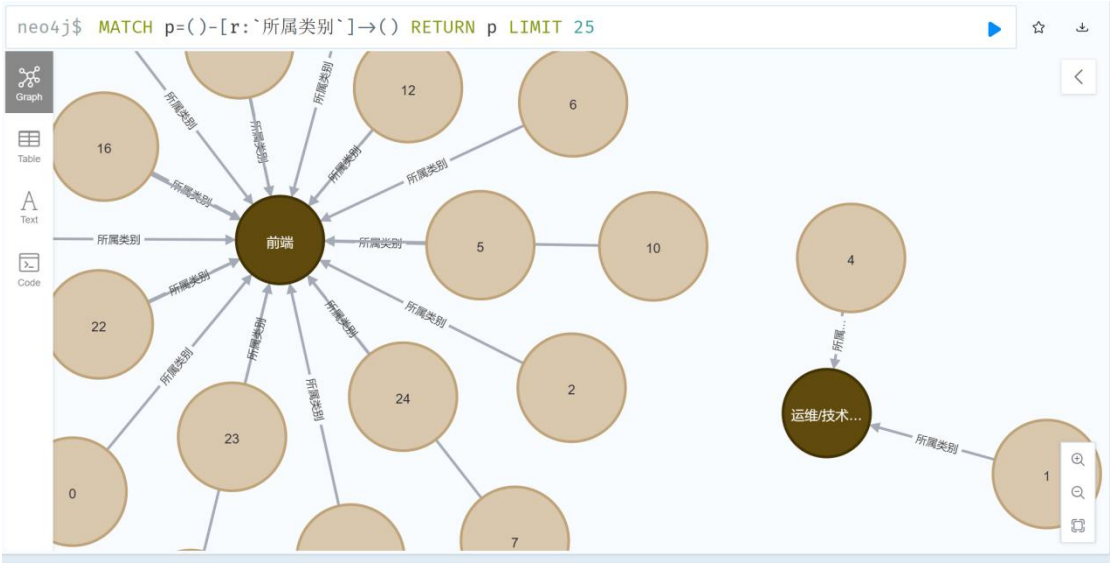
我们建好的知识图谱，展示部分如下

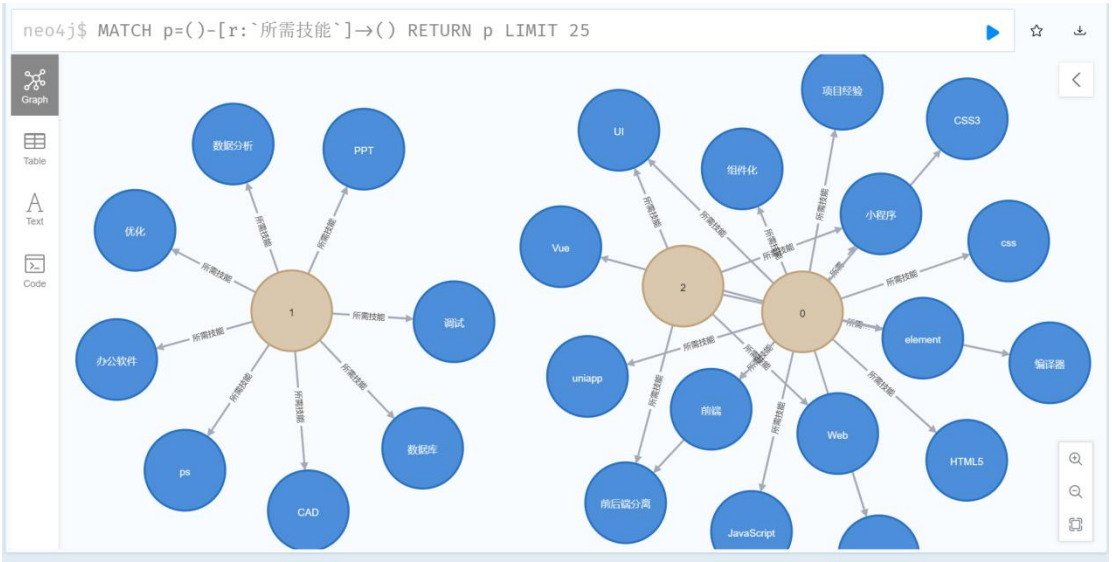


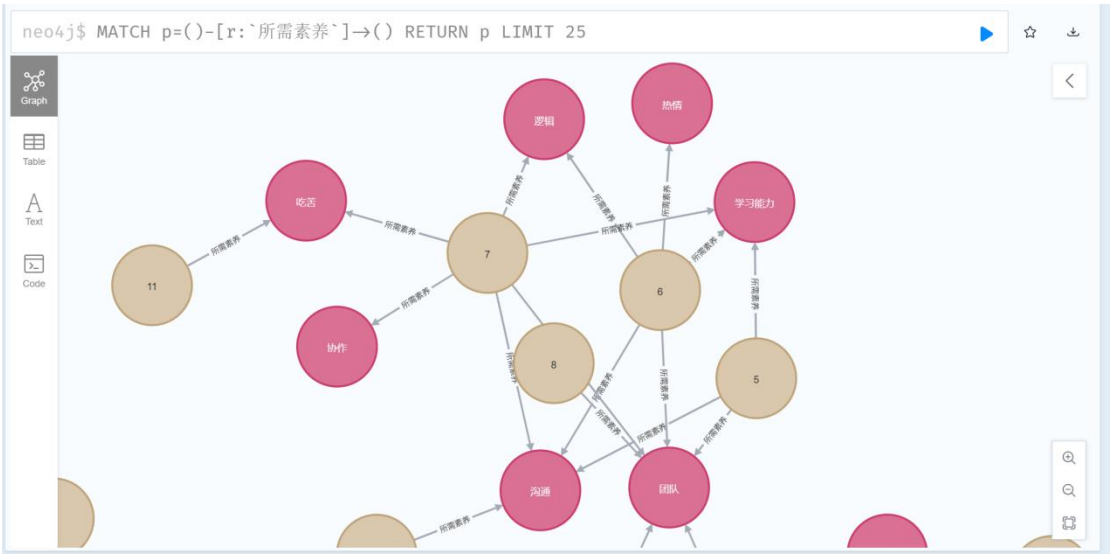


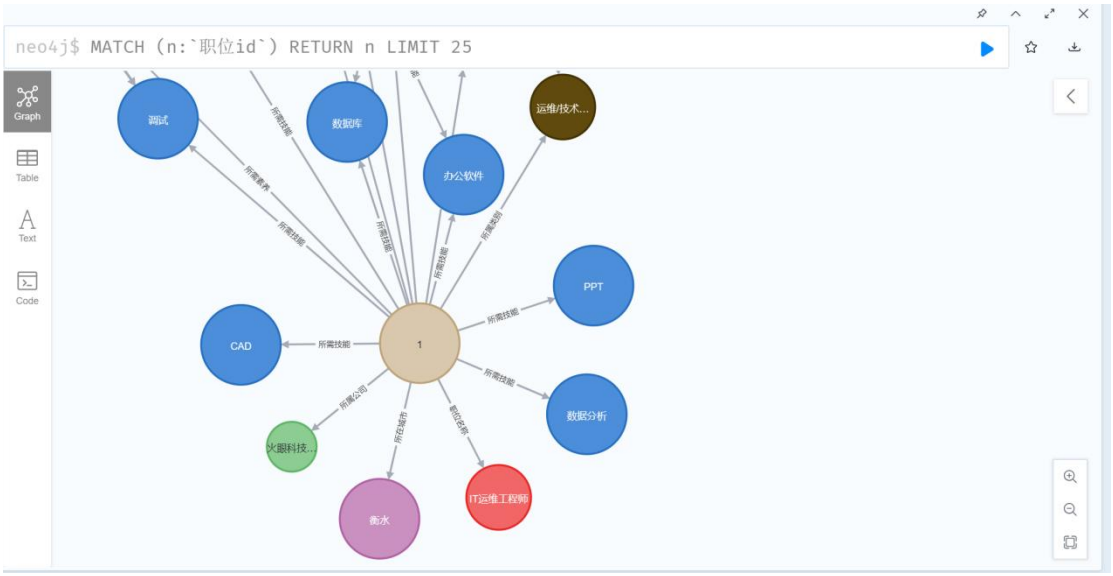


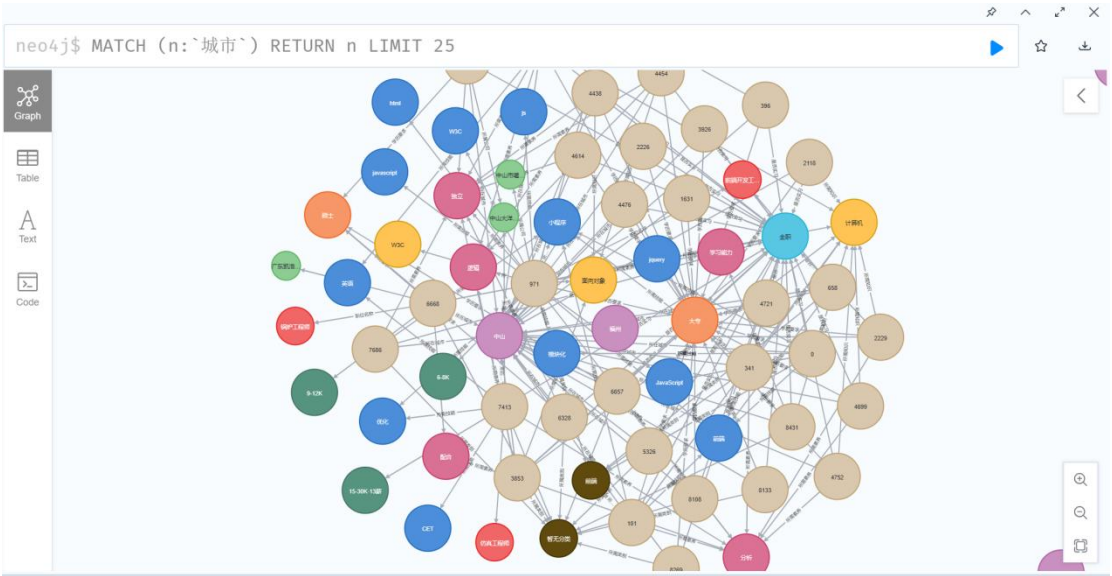

neo4j$ MATCH p=()-[r:`学历要求`]->() RETURN p LIMIT 25
Graph
Table
Text
Code
0
14
16
18
17
大专
9
19
学历要求
学历...
学历


neo4j$ MATCH p=()-[r:`所属类别`]->() RETURN p LIMIT 25
Graph
Table
Text
Code
前端
所属类别
16
12
6
5
10
4
22
24
2
23
0
7
运维/技术...
1


neo4j$ MATCH p=()-[r:`所需技能`]->() RETURN p LIMIT 25
Graph
Table
Text
Code
数据分析
PPT
优化
1
调试
办公软件
ps
CAD
数据库
所需技能
项目经验
UI
CSS3
组件化
小程序
Vue
2
0
css
element
编译器
uniapp
前端
Web
HTML5
前后端分离
JavaScript


neo4j$ MATCH p=()-[r:`所需素养`]→() RETURN p LIMIT 25
Graph
Table
Text
Code
逻辑
热情
吃苦
学习能力
11
7
6
协作
8
5
沟通
团队
所需素养


neo4j$ MATCH (n:`职位id`) RETURN n LIMIT 25
Graph
Table
Text
Code
调试
数据库
运维/技术...
办公软件
PPT
CAD
1
数据分析
火眼科技...
衡水
IT运维工程师
所需技能
所需素养
所属公司
所在城市
职位名称


neo4j$ MATCH (n:`城市`) RETURN n LIMIT 25
Graph
Table
Text
Code

4. 算法剩余任务

（1）**简历解析**（从用户上传的 pdf 或 word 提取其个人信息，包括知识、技能、素养，这种可以继续使用关键词提取；对于姓名、年龄、学校、工作经历等又该如何提取？）

一，尝试了 github 开源的一些模型——效果不理想——继续查找

二，有简历解析 API，我们只需要在后端调用它的接口即可解析简历，目前申请到了 300 份额度(足够，不够再问他要)——效果很好，但不知这样合不合适

（2）推荐算法

学习并使用推荐算法，基于我们上面已经构建的知识图谱。

1，【***Graphsage 图推荐模型***】？？

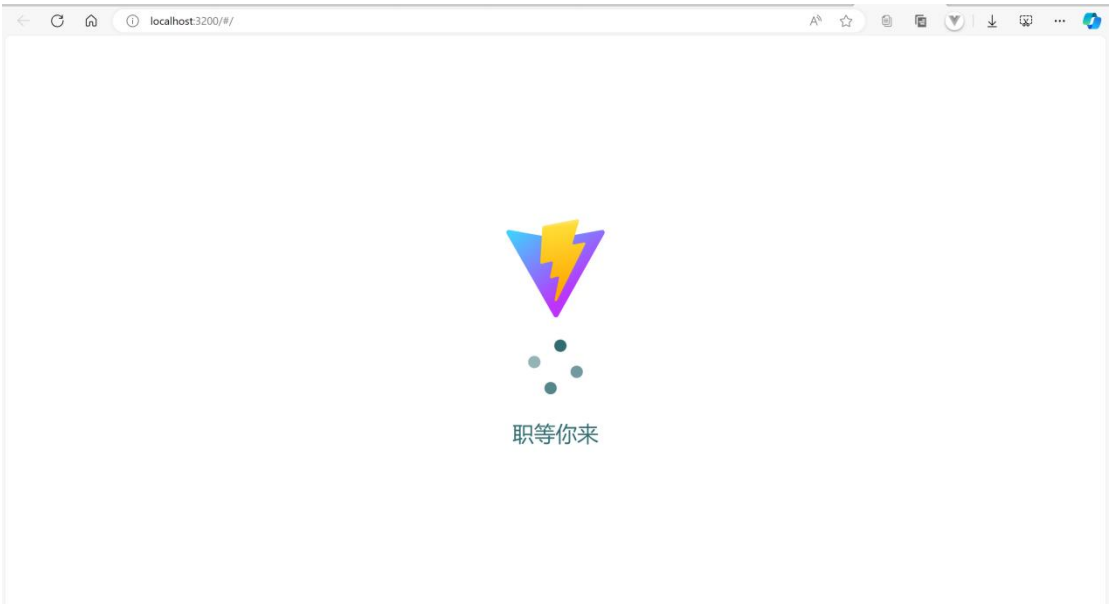
2，协同过滤？

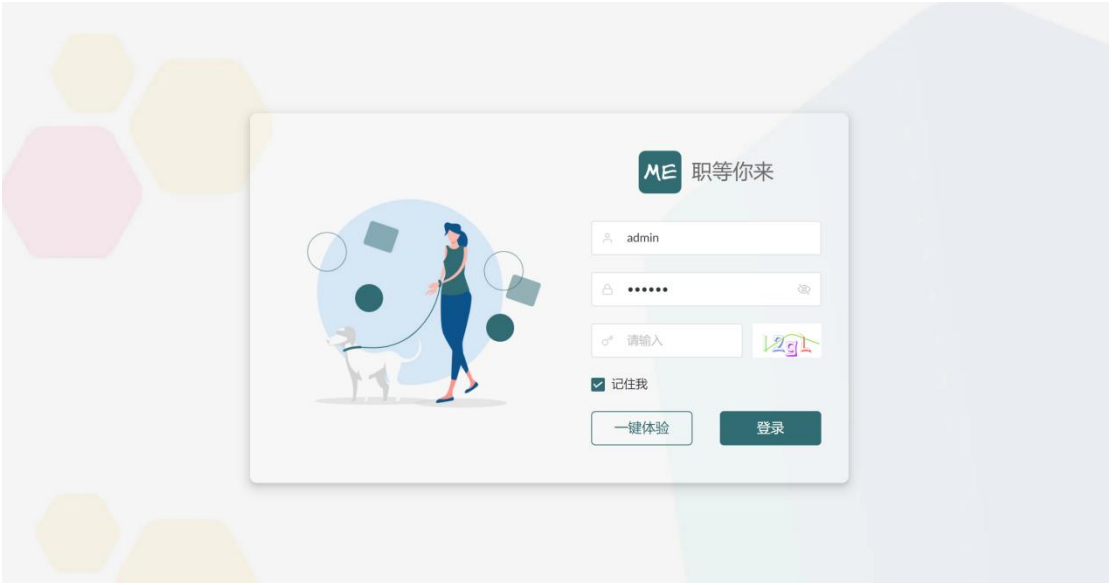
3，基于内容的推荐。？

4，基于深度学习。。。？

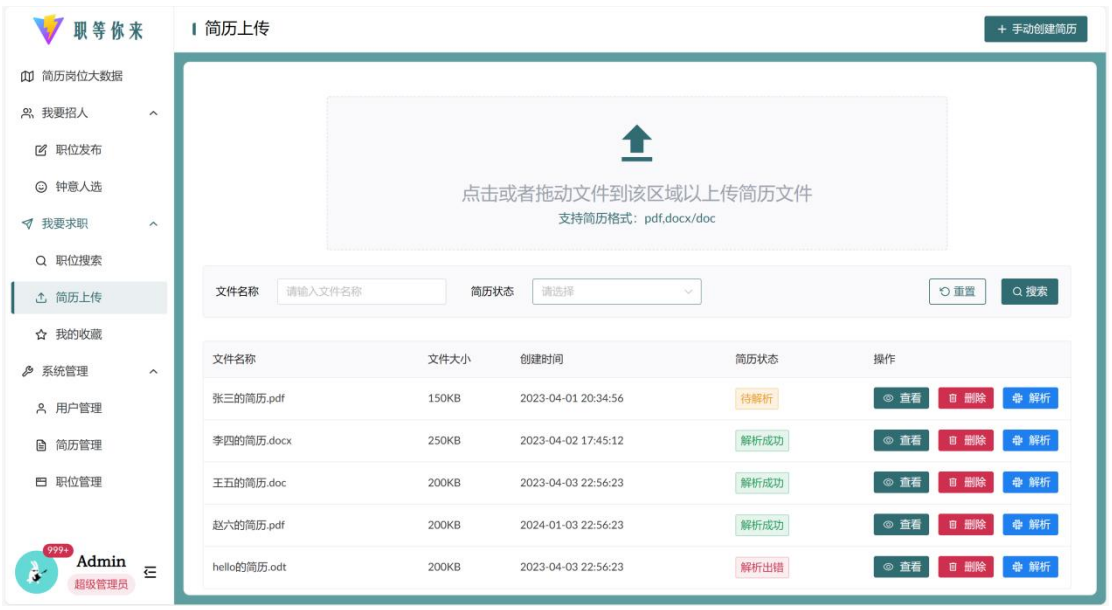
二、前端

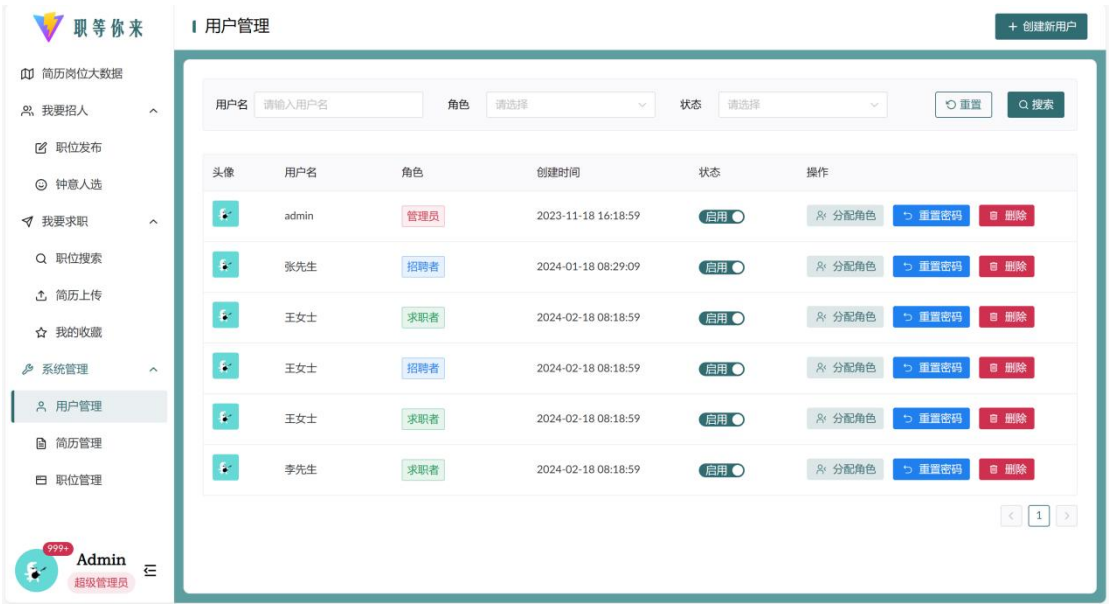
完成了部分界面。




职等你来
admin
••••••
请输入
记住我
一键体验
登录


职等你来
简历岗位大数据
我要招人
职位发布
钟意人选
我要求职
职位搜索
简历上传
我的收藏
系统管理
用户管理
简历管理
职位管理
Admin
超级管理员
请输入职位、公司
搜索
求职类型
不限
全职
实习
城市和区域
全国
北京
上海
广州
重庆
深圳
天津
武汉
成都
杭州
南京
西安
青岛
郑州
苏州
其他城市
更多筛选条件
C++
Java
Python
前端
在校生
应届生
经验不限
1年以内
5-10K
本科
大专
网络工程师
立即沟通
12-20K
3年以上
大专
北京
全职
网络安全
调试
VPN
云计算
计算机
积极
学习能力
沟通
分析
团队
合作
抗压
热情
汉博利特（北京）信息技术有限公司 刘女士
北京通州区金融街园中园6号院53号楼
网络工程师
立即沟通
12-20K
3年以上
大专
北京
全职


职等你来
简历岗位大数据
我要招人
职位发布
钟意人选
我要求职
职位搜索
简历上传
我的收藏
系统管理
用户管理
简历管理
职位管理
Admin
超级管理员
简历上传
手动创建简历
点击或者拖动文件到该区域以上传简历文件
支持简历格式：pdf,docx/doc
文件名称
请输入文件名称
简历状态
请选择
重置
搜索
文件名称
文件大小
创建时间
简历状态
操作
张三的简历.pdf
150KB
2023-04-01 20:34:56
待解析
查看
删除
解析
李四的简历.docx
250KB
2023-04-02 17:45:12
解析成功
查看
删除
解析
王五的简历.doc
200KB
2023-04-03 22:56:23
解析成功
查看
删除
解析
赵六的简历.pdf
200KB
2024-01-03 22:56:23
解析成功
查看
删除
解析
hello的简历.odt
200KB
2023-04-03 22:56:23
解析出错
查看
删除
解析


职等你来
简历岗位大数据
我要招人
职位发布
钟意人选
我要求职
职位搜索
简历上传
我的收藏
系统管理
用户管理
简历管理
职位管理
999+
Admin
超级管理员
用户管理
+ 创建新用户
用户名 请输入用户名
角色 请选择
状态 请选择
重置
搜索
头像 用户名 角色 创建时间 状态 操作
admin 管理员 2023-11-18 16:18:59 启用 分配角色 重置密码 删除
张先生 招聘者 2024-01-18 08:29:09 启用 分配角色 重置密码 删除
王女士 求职者 2024-02-18 08:18:59 启用 分配角色 重置密码 删除
王女士 招聘者 2024-02-18 08:18:59 启用 分配角色 重置密码 删除
王女士 求职者 2024-02-18 08:18:59 启用 分配角色 重置密码 删除
李先生 求职者 2024-02-18 08:18:59 启用 分配角色 重置密码 删除
1


localhost:3200/#/profile
职等你来
简历岗位大数据
我要招人
职位发布
钟意人选
我要求职
职位搜索
简历上传
我的收藏
系统管理
用户管理
简历管理
职位管理
999+
Admin
超级管理员
用户名: admin
修改密码
更改头像
修改头像只支持在线链接，暂时不提供上传图片功能!
个人资料信息
修改资料
照片
姓名 Admin
性别 未知
出生年月 2024-2
城市 南京
电话 15552510062
微信号 hxf15552510062
邮箱 635663114@qq.com
期望职位
修改期望
实习/全职 全职
类型 Java
薪资范围(k) 30 50
工作城市 南京


职等你来
简历岗位大数据
我要招人
职位发布
钟意人选
我要求职
职位搜索
简历上传
我的收藏
系统管理
用户管理
简历管理
职位管理
999+
Admin
超级管理员
教育经历
编辑教育经历
东南大学 985院校 211院校 双一流院校
2021-2025 计算机科学与技术 本科
专业排名: 前1% 主修课程: 数据结构、数据库、C++、Java程序设计、计算机网络 毕设/论文: 暂无 描述: 暂无。
南京大学 985院校 211院校 双一流院校
2025-2027 计算机科学与技术 硕士
专业排名: 前1% 主修课程: 数据结构、数据库、C++、Java程序设计、计算机网络 毕设/论文: 暂无 描述: 暂无。
工作实习经历
编辑工作实习经历
2021-2022 公司名称、职位名称、工作内容、拥有技能等
2022-2025 公司名称、职位名称、工作内容、拥有技能等
项目经历
编辑项目经历
项目1 项目角色、项目描述、项目时间等
项目2 项目角色、项目描述、项目时间等
技能|知识|素养
编辑个人能力
技能
知识
素养
自定义添加
编辑自定义
资格证书、社团组织经历、培训经历和志愿者经历等描述。

三、后端

项目初始化。

四、学习优秀作品

【2023第十二届中国软件杯一等奖】A8-职得智能简历解析系统 宣传视频

1380 0 2023-08-23 00:41:57 未经作者授权，禁止转载

【2023第十二届中国软件杯一等奖】A8-职得智能简历解析系统 宣传视频

1380 0 2023-08-23 00:41:57 未经作者授权，禁止转载

【2023第十二届中国软件杯一等奖】A8-职得智能简历解析系统 宣传视频

1380 0 2023-08-23 00:41:57 未经作者授权，禁止转载

五、企业答疑

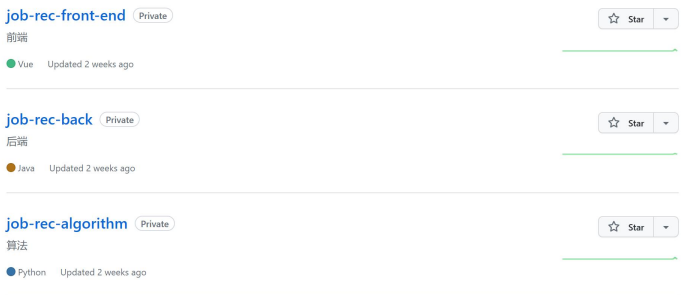
1	问题	A15中的数据集是否已经给全，理论上应该还有一个简历数据集	
	答复	简历没有相关数据集，可以使用学生自有数据集。	2024. 1. 5
2	问题	【问题说明】中"用户上传个人简历"，这个“上传”如何定义，是只需要用户输入信息吗？\n【技术要求与指标】（1）推荐有效性达到80% 以上。这个80%如何定义，题目只给了岗位信息，没有给任何的简历信息、或者推荐信息。	
	答复	企业会根据简历样本做匹配正确性验证。	2024. 1. 5
3	问题	企业要求提交的材料中提到了“软件安装包”，这是否意味着只能开发手机端app，开发web网站可以吗，微信小程序呢？	
	答复	软件安装包是指软件部署程序，各种形式都可以，如果是web的，要提供一下部署包和部署手册，就是如何将你写的程序运行起来。或者提供一个可以访问的地址也行。	2024. 1. 12
4	问题	还是不太清楚：推荐有效性达到80%以上是什么意思？这个有效性是如何验证的？	
	答复	推荐有效性由学生自行测试判断和匹配。比如输入10份简历，匹配出的结果与简历的符合程度人为判断有8份满意。我们也有一个自己的简历库，到时候我们也会测试一下效果。	2024. 1. 12
5	问题	A15中缺少简历数据集，大学生简历数据集属于隐私数据，网上没有公开数据集，爬虫也不能爬取该类数据，属于违法行为，该赛题只提供了岗位信息数据集请问是否会提供相应脱敏处理的简历数据集用于模型的训练？如果不提供是否允许使用大模型生成的简历数据集？	
	答复	B站命题直播回放有相关的解答。	2024. 1. 26
6	问题	请问上传简历的格式有要求吗	
	答复	没有要求。	2024. 2. 1
7	问题	用户上传的简历是图片还是文件呢？	
	答复	文件。	2024. 2. 1

8	问题	请问隐私信息只有授权人员才能解密，授权人员是指什么？管理员吗？	
	答复	需求主要是指意向岗位和专长等方向提取关键词进行匹配。隐私信息只有授权人员解密，是指被授权拥有查看的管理员，管	2024.2.1
9	问题	对于用户个人隐私的加密部分，授权人员是用户本人吗，那么企业有无权限看到匹配到的求职人员简历中人员的隐私信息。	
	答复	用户可以设置自己的隐私部分内容信息是否允许企业查看。	2024.2.1
10	问题	文档中在系统数据库层面对用户的隐私信息进行加密指的是什么？还有只能被授权人员解密中的授权人员指的是谁？是我们设计时让用户本人决定还是？谢谢！	
	答复	隐私信息要在数据库中加密存储，授权人员是自己，被授权人员是指企业方用户，学生用户设定自己的隐私信息是对企业用	2024.2.1
11	问题	“保护简历中个人数据的安全，不侵犯用户隐私”具体指什么，我们有以下猜想:1.在数据库对用户信息进行类似密码的加密，只有授权人才可以获取真实信息，但是授权人是谁，好像不太清楚 2.需要预防企业看到所有用户的简历吗，这是否算侵犯隐私 3.只能确定，用户之间肯定是不能看简历的	
	答复	授权人员指用户自己。企业用户可以查看所有开放的简历信息，隐私字段可以设置企业用户是否可以见。学生用户之间是不能看简历的。	2024.2.1
12	问题	构建职位知识图谱，然后输入简历可以推荐职位，那我们没有构建求职者的知识图谱，那如何为招聘人员推荐求职者呢？	
	答复	简历跟企业招聘的岗位进行匹配。	2024.2.

	问题	“岗位推荐：用户上传个人简历，系统自动分析简历内容，生成推荐职位，用户可以给出推荐是否有效的反馈。如果不满意，可修订简历部分内容，重新进行推荐。”不满意为什么可以修订简历内容，这里修订的主要是哪些内容？	
	答复	不满意是指学生对系统推荐的结果不满意，可以自行调整简历的内容，可能简历写的技能不够完整，导致推荐结果不符合自己的期望。	2024.2.23
	问题	能力评价部分那里，用户上传个人简历，并明确自己期望的职位，系统自动判断用户与期望职位间的契合度。这里职位是指的是具体公司的岗位还是职位大类。简单来说是指比方说是指美团的前端工程师还是整个前端工程师大类	
	答复	职位大类。	2024.2.23
	问题	构建的知识图谱，要包含每个具体职位的信息吗，如公司，薪资，地点等	
	答复	可以包含，自行设定，赛题不做统一要求。	2024.2.23
	问题	职位和岗位是否是同一概念？能力评价中选择的职位、岗位推荐中推荐的岗位，是均指广义的岗位名称还是指某公司的具体某岗位名称？比如，能力评价在选择意向职位时，是选分类后的“前端工程师”、“后端工程师”（只是广义的名称，不是具体某一公司的招聘职位名称），还是选企业1的“后端工程师”、企业2的“后端工程师”（每个企业的招聘岗位名称均会列出）	
	答复	广义的，不具体到某企业。	2024.2.23

六、进度

1.注册 github 账号，加入项目。



2.简历解析、职位推荐、人才推荐

3.完善并丰富前端

4.后端及时跟进

*报名截止时间： 2024 年 3 月 20 日

*初赛作品提交时间：2024 年 4 月 9 日—2024 年 4 月 15 日。