第十一届“泰迪杯”数据挖掘挑战赛——

C 题：泰迪内推平台招聘与求职双向推荐系统构建

一、问题背景

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。

二、解决问题

1. 招聘信息爬取

网络爬虫，抓取对应网址中的数据，具体字段应该包括问题2中所需要的一些信息。

从泰迪内推平台（https://www.5iai.com/#/index）的“找工作”页面和“找人才”页面，爬取所有招聘与求职信息并整理，依据招聘信息 ID 记录每条招聘信息并保存为“result1-1.csv”文件，求职信息则依据求职者 ID 记录并保存为“result1-2.csv”文件，涉及的招聘信息 ID 和求职者 ID 均来自网址路径后端的数字串，如图 1 所示。（模板文件见附件 1 中的 csv 文件）

图 1　某招聘信息网页

间接告知了需要采集的数据字段

对1中的数据做数据预处理，特征工程等

2. 招聘与求职信息分析

应用问题 1 的招聘信息与求职信息构建画像：根据采集的企业招聘信息，从招聘岗位、学历要求、岗位需求量、公司类型、薪资待遇、岗位技能、企业工作地点等多个方向建立招聘信息画像；根据采集求职者求职信息，从预期岗位、薪资需求、知识储备、学历、工作经验等多个方向建立求职者画像。

3. 构建岗位匹配度和求职者满意度的模型

硬性指标，例如：
1. 学历
2. 工作经验
3. 期望的最低工资
4. 岗位给的最高工资低于求职者的最低工资要求

在招聘和求职过程中，企业面对多位优质求职者，将会考虑求职者能力要求、技能掌握等多方面，岗位匹配度是体现求职者满足企业招聘要求的匹配程度；同样，求职者对于多种招聘信息，也会依据自身条件和要求，选取符合自己心意的岗位，因此求职者满意度指标可客观体现求职者对企业招聘岗位的满意程度。对于不满足岗位最低要求的求职者，企业可定义其岗位匹配度为 0。同样，对于不满足求职者最低要求的岗位，求职者可定义其求职者满意度为 0。

根据问题 2 的招聘信息与求职者信息，构建岗位匹配度和求职者满意度的模型，基于该模型，为每条招聘信息提供岗位匹配度非 0 的求职者，将结果进行降序排序存放在“resul3-1.csv”文件中，以及为每位求职者提供求职者满意度非 0 的招聘信息，将结果进行降序排序存放在“result3-2.csv”文件中。（模板文件见附件 1 中的 csv 文件）

4. 招聘求职双向推荐模型

假设招聘流程如下：设某岗位拟聘 n 人，泰迪内推平台向企业推荐岗位匹配度非 0 的 n 位求职者发出第一轮 offer，求职者如果收到多于 1 个岗位的 offer，则求职者选取满意度最高的岗位签约，每个求职者只允许选择 1 个岗位签约。第一轮结束后，平台根据当前各招聘信息的剩余岗位数，向后续被推荐求职者发出第二轮 offer，如此继续，直到招聘人数已满或者向所有拟推荐求职者均已发出 offer 为止。

在上述招聘流程中，由于条件优秀的岗位求职者都愿意去，而条件优秀的求职者各岗位都愿意录用，很难做到履约率达到百分之百，因此履约率高低是评价平台的推荐系统优劣的重要指标。这里的履约率定义为：

履约率=所有岗位的签约人数之和/所有拟聘岗位人数之和

请为平台设计招聘求职双向推荐模型，使得履约率指标达到最高。并将招聘岗位与求职者签约成功的结果存放在“resul4.csv”文件中。

三、附件说明

附件 1 是问题 1、问题 3 和问题 4 的模板文件，文件均为 csv 文件，采用 ANSI 编码。

result1-1.csv：从泰迪内推平台爬取的招聘信息，文件参考表 1 格式。

表 1　result1-1.csv 样例

序号	招聘信息 ID	企业名称	招聘岗位	其他招聘信息
1	12356983056	中*企业	数据分析师	⋮
2	35679125799	联*信息有限公司	人工智能开发师	⋮
…	…	…	…	…

result1-2.csv：从泰迪内推平台爬取的求职信息，文件参考表 2 格式。

表 2　result1-2.csv 样例

序号	求职者 ID	姓名	预期岗位	其他招聘信息
1	9864267664	刘**	数据分析师	⋮

2	2565887924	张**	数据挖掘工程师、数据分析师	：
…	…	…	…	…

result3-1.csv：该文档存储每条招聘信息中岗位匹配度非 0 的求职者，需将结果进行降序排序，具体字段名和样例见表 3。

表 3　result3-1.csv 样例

招聘信息 ID	求职者 ID	岗位匹配度
1461579387123138560	1365879387125688560	0.87
1461579387123138560	5761579356823138358	0.81
1461579387123138560	1259734889723772246	0.72
1461579387123138560	5761579356823138358	0.70
1461579387123138560	3655671218465889104	0.64
1461579387123138560	1979608497675601812	0.51
…	…	…

result3-2.csv：该文档存储每位求职者满意度非 0 的招聘信息，需将结果进行降序排序，具体字段名和样例见表 4。

表 4　result3-2.csv 样例

求职者 ID	招聘信息 ID	公司名称	求职者满意度
1365879387125688560	1461579387123138560	公司 V	0.85
1365879387125688560	1659723460008465808	公司 T	0.81
1365879387125688560	1957497826300498700	公司 H	0.76
…	…	…	…

result4.csv：根据履约率最高的模型，提供招聘岗位签约成功后的求职者 ID。该结果需对招聘信息 ID 进行排序，并对每个招聘信息的数据按岗位匹配度降序排序，具体字段名和样例见表 5。

表 5　result4.csv 样例

招聘信息 ID	求职者 ID	岗位匹配度	求职者满意度
1461579387123138560	1365879387125688560	0.87	0.91
1461579387123138560	5761579356823138358	0.81	0.86
1461579387123138560	1259734889723772246	0.72	0.90
1897843625703822037	5761579356823138358	0.84	0.83
1897843625703822037	3655671218465889104	0.80	0.86
1897843625703822037	1979608497675601812	0.76	0.78
1897843625703822037	1368492782664320473	0.71	0.88
…	…	…	