2024年秋Java期末大作业

# 1、政策数据库与大模型

（1）**动态追踪**解析每日更新的xml文档（联邦公报）并存储到MySQL数据库，调用大模型API翻译政策文本，形成中英文两个版本的政策文本。基于前述政策文本数据构建milvus向量数据库。

（2）调研、收集美国司法部门发布的与中国科技领域相关的案件信息，包括对企业与个人的处罚等，将这些案件信息整合到MySQL政策数据库中。并将案件信息添加于Milvus向量数据库。

（3）构建B/S模式信息系统，具备三个功能。一是政策文本查询：实现一个用户界面，允许用户基于关键词、政策颁布部门、政策颁布时间等条件查询政策文本。二是政策问答功能：集成大模型，实现对政策文本的问答。用户可以提出问题，如“今天工业安全局发布了哪些文件？是否有与出口管制相关的文件？如果有与出口管制相关的文件，请对相关文件进行总结与摘要。”等。系统需要能够理解问题，并从数据库中检索相关信息，然后使用大模型生成回答。三是案件分析问答：扩展问答功能，允许用户查询和分析美国司法部发布的与中国科技相关的案件。例如，“美国司法部对中国科技企业有哪些最新的处罚措施？”或“分析美国司法部案件对中国科技企业的影响。”

（4）如果有合适的回答结果，允许用户将回答结果一键存为word文件生成简报。

数据源：【<https://www.govinfo.gov/bulkdata/FR/2024/10>】

备注：联邦公报只需要解析存储【Bureau of Industry and Security、Treasury Department、State Department、Justice Department】这四个部门每日更新的所有政策文件即可，需要保存所有字段，具体字段类型为大多数文件的字段的并集。

# 2、新闻数据库与大模型

（1）**动态追踪**解析网易和CNBC每日更新的科技版块新闻并存储到MySQL数据库，利用大模型或者深度学习等方法判断新闻属于哪个科技领域（量子领域、人工智能、半导体…等）作为标签。基于新闻文本数据构建milvus向量数据库。

（2）构建B/S模式信息系统，具备三个功能。一是基于关键词实现对新闻文本的查询，用户可以设置新闻时间。二是新闻查询结果数量可视化，用户选择领域，不同时间的新闻可视化；选择时间，不同领域新闻数量分布情况等等。三是调用大模型实现新闻的问答功能，例如“今天科技板块发布了哪些新闻？是否有与出口管制相关的新闻？如果有与出口管制相关的新闻，请对相关新闻进行总结与摘要。”等。

（3）如果有合适的回答结果，允许用户将回答结果一键存为word文件生成简报。

数据源：【https://tech.163.com/】【https://www.cnbc.com/technology/】

# 3、企业数据库与大模型

（1）给定一个企业名单，包含若干信息字段，基于企查查网站及官方网站完善企业信息（如企业名称、行业、注册地址、主要产品等）并存入MySQL数据库。

（2）竞争对手与类似企业信息的整合。注意除了给定的企业名单，还需要查找企业网站，确定主要产品，并基于此并识别竞争对手以及技术产品类似的国外企业。将这些信息整合到系统中，为用户提供更全面的企业信息。

（3）构建Neo4j图数据库，将完善后的企业信息导入，企业的关系的连接是通过信息字段同属性值实现，例如两家企业行业同属于计算机、通信和其他电子设备制造业，则两家企业通过“计算机、通信和其他电子设备制造业”这一节点连接。基于企业信息构建milvus向量数据库。考虑到图数据库的构建可能较为复杂，任务中不强制要求使用Neo4j，也可以选择以数据表形式进行关联的呈现，实践上可以根据小组成员知识储备情况尽可能完善以提高作业水准。

（4）构建B/S模式信息系统，具备三个功能。一是基于企业名称实现对企业信息的查询，要求使用企业简称模糊查询也可行。二是检索一个企业时，基于图数据库在前端同时显示与之强关联的企业，强关联指两个企业有3个以上节点直接相连，即有3个以上相同的属性值。三是调用大模型实现企业信息的问答功能，例如“结合企业名单，人工智能领域有哪些企业进入该名单了？”等。

数据源：【neo4j教程文档：https://www.w3cschool.cn/neo4j/】

# 4、专利数据库与大模型

（1）给定一个企业名单，采集这些企业的专利申请数据并存储到MySQL数据库，利用大模型或者深度学习等方法判断这些专利属于哪一个领域，加一列作为领域Label。给定两个领域的关键词（人工智能、量子领域…等），检索这两个领域专利近2年的专利申请数据并存储到MySQL数据库。基于前述专利数据构建milvus向量数据库。

（2）构建B/S模式信息系统，具备三个功能。一是基于企业名称/专利领域实现对企业专利信息的查询，要求使用企业简称模糊查询也可行。二是专利查询结果数量可视化，用户选择企业，其不同领域的专利数量情况；选择领域，不同企业/国家的专利申请情况等等。三是调用大模型实现企业专利信息的问答功能，例如“XXX企业今年申请的专利可以分为哪些类别？”、“人工智能领域申请专利的那些企业，哪些企业进入了该名单，哪些企业未进入该名单，请分别提供相应企业名单。”等。

（3）扩展功能：相关专利与论文分析。除了采集名单企业的专利数据，还需要分析与这些企业相关的专利，以及企业主要技术相关的论文。相关专利分析：如果该名单企业没有专利，分析其主要技术的相关专利。查找与该名单企业技术产品相似的专利，包括国内外的专利。论文分析：收集企业本身发表的论文。收集他人发表的与企业主要技术相关的论文。将这些论文信息整合到系统中，为用户提供更全面的企业技术背景。

（4）一键生成简报。如果有合适的回答结果，允许用户将回答结果一键存为Word文件生成简报。

数据源：【https://www.incopat.com/】

【https://www.cnki.net/】

【https://www.webofscience.com/wos/】

# 5、论文/智库报告与大模型

（1）下载一批出口管制主题相关的研究论文和智库报告，将其存入MySQL数据库。调研将论文/智库报告PDF文件向量化的技术路线。

（2）基于前述技术路线，构建B/S模式信息系统，实现对论文、报告的增删改查的基础功能。用户给定主题后，系统基于大模型撰写一份给定主题（收集的论文/报告的子主题，例如“芯片出口管制”等）的报告，要求用户可以自由选择是否连接互联网，如果连接联网，则模型会参考互联网资料以及论文/智库报告，否则只参考论文/智库报告数据。

针对撰写的报告，如果参考了数据库的报告，则需要列出来参考了哪一篇。

（3）如果有合适的回答结果，允许用户将回答结果一键存为word文件导出。

数据源：【http://www.cggthinktank.com/】中国智库

【https://cset.georgetown.edu/】【https://crsreports.congress.gov/】国外智库

【https://www.cnki.net/】【https://www.webofscience.com/wos/】论文库等。

# 6、AI Agent调用小模型预测股票收益率

（1）采集一批股票的历史量价数据，训练一个性能良好的深度学习模型，用于预测指定股票未来3天的收益率。

（2）构建股票量价信息数据库，调用Ricequant或者Tushare等数据源实现股票量价信息数据库信息的实时更新。

（3）构建B/S模式信息系统，用户通过对大模型下达指令，大模型分析指令，并回答用户问题或执行任务。例如“请分析茅台最近一周的股价走势。”、“请预测茅台未来3天的股票收益率。”针对第一个问题，大模型可以基于数据库数据直接回答；第二个问题属于垂直领域的特定任务，大模型预测的性能往往低于自行训练的模型，针对该问题，需要在用户提出指令后，基于Agent自动调用训练的模型实现预测，并将预测结果通过前端反馈给用户。

数据源：【https://www.ricequant.com/welcome/】【https://tushare.pro/】等。

针对所有题目，模型的训练、微调等可以使用任何程序语言和工具，系统的构建需要使用Java，具体来说即构建Java Web系统。大模型通过调用API实现（如通过通义千问、Kimi、OpenAI等API），此外也可以利用LangChain等开发框架。前端使用统一提供的模板，动态交互效果等可根据小组对题目的理解自行设计。

**Milvus数据库资料：**

【https://blog.csdn.net/lsb2002/article/details/132222947】

【GitHub教程：https://github.com/milvus-io/milvus】

【官网手册：https://milvus.io/docs】

不限于以上资料，可参考其他任何文档、视频等。