赛题 04

数据标注平台

项目概览与计划

目录

1 项目简介： ---------- **3**

1.1 背景： ----------3

1.2 重要性： ----------3

2 目标和范围： ---------- **3**

2.1 目标： ----------3

2.2 范围： ----------4

2.3 预期成果： ----------4

3 项目计划： ---------- **4**

3.1 启动阶段： ----------4

3.2 需求分析与规划阶段： ----------5

3.3 设计阶段： ----------5

3.4 开发阶段： ----------5

3.5 测试阶段： ----------5

3.6 部署和上线阶段： ----------5

3.7 项目收尾阶段： ----------5

4 项目管理策略： ---------- **6**

4.1 团队结构与角色： ----------6

4.2 职责分配与协作： ----------6

4.3 进度管理与监控： ----------7

4.4 风险管理： ----------7

4.5 文档管理： ----------7

5 沟通计划： ---------- **8**

5.1 沟通目标： ----------8

5.2 沟通方式和工具： ----------8

5.3 沟通评率和实践： ----------8

5.4 沟通内容： ----------8

5.5 沟通记录和文档 ----------8

1 项目简介：

1.1 背景：

随着人工智能技术的迅速发展，越来越多的领域需要使用文本和图片数据进行训练和优化，例如自然语言处理、计算机视觉等领域。例如，在自然语言处理领域，机器翻译、情感分析、语音识别等任务都需要使用大量的自然语言数据进行训练；在计算机视觉领域，图像分类、目标检测、人脸识别等任务需要使用大量的图像和视频数据进行训练。而这些数据的标注工作是非常繁琐、耗时的，需要大量的人力和时间投入。

为了解决这个问题，数据标注平台应运而生。这些平台可以为研究者和企业提供一个方便、快捷、高效的标注工具，使得标注过程更加高效、准确和可靠。同时，平台也可以提供多种标注方式和标注规范，以确保标注数据的质量和一致性。此外，平台还可以提供数据管理、质量控制、安全保障等方面的支持，以确保标注过程的稳定性和可靠性。

1.2 重要性：

恒生电子从 2018 年开始涉足人工智能相关领域，目前公司内已有自然语言处理、图像处理、知识图谱等相关人工智能产品。在这个过程中，积累了大量的金融词库、语料及图片标注等数据，目前这些数据分别由不同的系统管理。随着数据量和业务规模的不断的增长，原有的数据标注方式和管理模式已现瓶颈，我们需要搭建统一的标注平台，制定标准的数据标注流程和规范，对标注数据进行统一的管控，并提供版本控制、用户隔离等机制，以提升相关研究人员的工作效率，并保障数据资产安全。

2 目标和范围：

2.1 目标：

针对目前数据标注存在的问题，企业需要搭建数据标注平台。构建一套企业的数据接入、数据标注、标注审核、数据发布的标准流程。对于标注数据平台需要提供标注数据版本管理机制，以达到数据与下游模型版本匹配的效果。同时，对于流程中不同的角色，平台需

要提供差异化的功能。此外，平台还需要对标注数据提供数据隔离机制，不同角色、不同级别的人所能看到数据不同，以保障企业数据资产的安全。

2.2 范围：

功能范围：

信息抽取、文本分类、图像文本、图像分类

技术范围：

开发语言：主要使用 Java 和 JavaScript 进行开发，确保系统的可移植性和高性能。
系统架构：B/S 体系结构：方便用户通过浏览器访问系统，无需额外安装客户端软件。
服务器：使用 Tomcat 或 Jetty 作为后端服务容器。采用 MySQL 数据库，确保数据的持久化存储和高效管理。

存储：支持 T 级别数据量的语料数据管理，保证数据的高可用性和扩展性。

业务范围：
针对金融市场分析、交易文本、财经新闻、金融报告等数据进行标注。

2.3 预期成果：

数据标注平台的开发与部署：成功构建并上线一个功能完善的数据标注平台，该平台能够处理和管理大规模的文本和图像数据。

高效率的数据处理：平台能够高效地导入、管理和导出标注数据，特别是针对金融领域的文本和图像材料。

准确性的标注结果：通过预标注和人工复核提高数据标注的准确性，为后续的分析和模型训练提供高质量的数据集。

用户友好的操作界面：开发直观、易用的用户界面，使得非技术背景的业务人员也能轻松进行数据标注。

提升 AI 模型性能：使用该平台标注的数据用于训练和优化金融领域的 AI 模型，提高模型的准确度和效率。

标准化的标注流程：建立和维护一套标准化的数据标注流程，确保数据质量和标注工作的可持续性。

文档和培训材料：提供完备的使用文档和培训材料，确保用户能够快速掌握平台的使用方法。

数据安全和隐私保护：确保平台符合数据安全和隐私保护的相关标准和法规要求。

系统的可扩展性和可维护性：平台设计需考虑未来的扩展性，易于添加新功能和维护。

3 项目计划：

3.1 启动阶段：

（2023.10.1——2023.10.10）

目标：正式启动项目，明确项目范围和目标。

关键活动：组建项目团队。开展启动会议，确立项目愿景和目标。制定初步的项目时间线和关键里程碑。

输出物：项目启动报告，包括项目目标、团队结构和初步时间线。

3.2 需求分析与规划阶段：

（2023.10.10——2023.10.20）

目标：深入理解和记录项目需求。

关键活动：与利益相关者沟通，收集和整理需求。完成需求规格说明书。制定详细的项目计划和资源分配方案。

输出物：需求规格说明书，详细项目计划。

3.3 设计阶段：

（2023.10.20——2023.11.1）

目标：完成系统的整体设计和架构。

关键活动：设计系统架构和数据库模型。制定用户界面设计草图。确定技术栈和开发工具。

输出物：系统设计文档，用户界面草图。

3.4 开发阶段：

（2023.11.1——2024.1.1）

目标：实现系统功能。

关键活动：编码和实现系统功能。定期进行代码评审和内部测试。更新项目进展和调整计划。

输出物：开发完成的软件，测试报告。

3.5 测试阶段：

（2024.1.1——2024.2.1）

目标：确保系统稳定性和性能。

关键活动：进行全面的系统测试，包括单元测试、集成测试和性能测试。收集测试结果并修复发现的问题。准备测试文档和用户手册。

输出物：测试报告，用户手册。

3.6 部署和上线阶段：

（2024.2.1——2024.3.1）

目标：将系统部署到生产环境并正式上线。

关键活动：配置生产环境。进行最终测试和验收。正式发布系统。

输出物：上线通知，部署指南。

3.7 项目收尾阶段：

目标：完成项目文档，评估项目结果。
关键活动：汇总项目文档。进行项目回顾和评估。安排后续支持和维护计划。
输出物：项目总结报告，维护和支持计划。

4 项目管理策略：

4.1 团队结构与角色：

项目负责人：全面负责项目的规划、执行和交付。确保项目目标按计划实现，并符合预期质量标准。
主要任务：
制定项目计划和时间线。
监控项目进度和预算。
协调团队成员和资源。

技术负责人：负责整个项目的技术方向和技术解决方案的实施。
主要任务：
确定技术架构和开发环境。
指导和协助开发团队实现技术目标。
确保技术解决方案的可行性和效率。
监督代码质量和系统性能。

客户关系经理：作为团队与客户之间的桥梁，确保客户需求和期望得到满足。
主要任务：
与客户沟通，了解并转达他们的需求和反馈。
确保项目成果符合客户的期望。
协调项目团队和客户之间的沟通和会议。
管理客户期望和建立长期合作关系。

4.2 职责分配与协作：

项目负责人（夏飞宇）：负责整个项目的统筹规划、执行和监控，确保项目按时完成并达到预期目标。与技术负责人协作，确保技术实现符合项目需求。与客户关系经理协作，了解客户需求和反馈，并将其整合到项目规划中。管理项目团队，确保所有成员明确职责并高效协作。

技术负责人（杨宇轩、陈昊、方劲）：负责项目的技术实现，包括系统架构设计、开发和测试。向项目负责人报告技术进展和挑战，确保技术方案与项目目标一致。指导开发团队，确保技术实现符合设计规范。与客户关系经理合作，确保技术实现能满足客户的实际需求。

客户关系经理（严才俊）：负责与客户的沟通和关系管理，确保客户需求被准确理解和传达给项目团队。向项目负责人提供客户的反馈和需求，确保项目方向符合客户期望。与技术负责人合作，确保技术解决方案能满足客户需求。定期更新客户关于项目进展和里程碑的达成情况。

4.3 进度管理与监控：

使用 Microsoft Project 进行任务规划和时间线管理：在 Microsoft Project 中创建详细的工作分解结构（WBS），列出所有任务和子任务。为每项任务分配开始和结束日期，确保时间线的准确性。

资源分配和管理：在 Project 中分配资源，包括团队成员和所需的工具或材料。跟踪资源使用情况，确保资源有效分配且无过度分配情况。

进度跟踪和更新：定期（如每周）更新任务的完成情况，记录实际进度与计划进度的对比。使用 Project 的“甘特图”视图跟踪整体项目进度。

里程碑和关键路径分析：标记关键里程碑，确保这些关键点的按时完成。使用关键路径分析来识别可能影响项目交付日期的关键任务。

风险和问题跟踪：在 Project 中记录潜在的风险和遇到的问题，包括风险等级和影响。定期评估这些风险和问题，制定应对策略。

报告和沟通：利用 Project 的报告功能生成项目状态报告，包括进度概览、资源分配和任务完成情况。将这些报告定期（如每月）分享给管理层和项目利益相关者。

4.4 风险管理：

定期识别潜在风险并记录在风险日志中。为每个已识别的风险制定应对策略。

4.5 文档管理：

版本控制和文档组织：使用 GitHub 的版本控制功能来管理所有文档的更改和历史记录。为不同类型的文档（如设计文档、需求文档、用户手册等）创建不同的仓库或在单个仓库中使用不同的文件夹。

文档编写和审阅流程：采用“分支”策略进行文档编写和更新。每次文档更改应在单独的分支上进行，并通过“Pull Request”进行审查。设定清晰的审阅流程，确保每次文档更新都经过至少一名其他团队成员的审阅。

文档格式和标准：统一文档格式，如 Markdown 或其他标准格式，以保持一致性和可读性。在项目的 README 文件中明确文档标准和风格指南。

文档访问和共享：确保所有团队成员都有适当的访问权限，以查看和编辑文档。使用 GitHub Pages 或类似工具将文档发布为易于访问和阅读的格式。

常规维护和更新：定期审查文档，确保内容的准确性和时效性。鼓励团队成员更新过时或不准确的文档。

备份和恢复：利用 GitHub 的自然备份功能来保护文档免遭意外丢失。定期检查备份的完整性和恢复流程。

5 沟通计划：

5.1 沟通目标：

确保信息及时、准确地在项目团队和利益相关者之间传递。
促进项目团队的协作与透明度。
及时解决问题和疑虑，减少误解和冲突。

5.2 沟通方式和工具：

腾讯会议：用于举行项目团队的定期会议和重要讨论。
微信群：日常沟通、快速响应和非正式更新。
项目管理工具：project，用于任务跟踪和进度更新。
电子邮件：传达正式通知和重要的项目文档。

5.3 沟通评率和实践：

项目团队会议（腾讯会议）：每周进行，重点关注项目的当前状态和即将到来的任务。
高级管理层汇报：每月通过电子邮件进行，提供项目进展和关键里程碑的详细汇报。
利益相关者更新:根据需要通过微信或电子邮件进行,特别是在关键决策或重大变更时。

5.4 沟通内容：

项目的当前进展和即将到来的里程碑。任何项目范围、时间线或资源的变更请求。团队成员的贡献和突出成就。

5.5 沟通记录和文档

记录所有正式会议的要点和决策。在项目管理工具中更新任务和进展。保存重要的沟通记录和文件以供未来参考。