**专业实践重点专题实践周报（2023.7.10-2023.7.14）**

杜国豪，面壁智能

1. 本周工作总结及收获

本周我主要进行了舆情检查接口的编写，并使用Python从大量PDF和Excel文件中抽取题目，提取并转换为JSON格式，以用于训练模型。在这个过程中，我积累了以下的工作总结和收获：

成功完成了舆情检查接口的编写。我设计了一个高效的接口，可以接收PDF和Excel文件，并自动提取其中的题目信息，将其转换为JSON格式，方便后续的模型训练使用。

使用Python编写了数据抽取和转换的代码。我熟练运用了Python中的PDF解析库和Excel读取库，成功从大量文件中提取了题目信息，并进行了有效的数据清洗和格式转换。

提取的题目数据经过验证和检查，确保了数据的准确性和完整性。我花费了大量时间和精力对提取的数据进行了验证和校对，以确保训练模型时的准确性。

掌握了JSON格式的使用和转换。通过将题目信息转换为JSON格式，我能够更好地管理和组织数据，便于后续的模型训练和分析。

1. 本周工作存在的问题

在本周的工作中，我也遇到了一些问题和困难：

大量文件的处理耗时较长。由于需要处理的PDF和Excel文件较多，导致数据的抽取和转换耗费了较多的时间和计算资源。在下一步的工作中，我需要进一步优化代码和算法，提高处理效率。

部分文件的格式不规范。在处理文件的过程中，我遇到了一些格式不规范的PDF和Excel文件，导致数据提取过程中出现了一些错误和异常情况。下一步我将加强对文件格式的检查和处理，提高程序的鲁棒性。

1. 下一步工作计划

基于本周的工作总结和问题反馈，我制定了下一步的工作计划：

优化数据处理代码和算法。针对大量文件处理耗时较长的问题，我将进一步分析代码和算法，寻找优化的空间，提高数据处理的效率和速度。

完善文件格式处理逻辑。针对文件格式不规范的情况，我将进一步完善代码逻辑，增加对各种格式的容错处理，确保数据提取的准确性和稳定性。

进一步验证和校对数据。在下一步的工作中，我将对提取的题目数据进行进一步的验证和校对，确保数据的准确性和完整性，为模型训练提供可靠的数据源。

开始模型训练和评估。一旦数据准备工作完成，我将开始进行模型训练和评估，利用提取的题目数据来训练模型，并进行性能评估和优化。

1. 其他（可另起标题）

暂无