# 实验报告：TMLR 和 ICLR 2024 论文数据处理与分析


## 张博飞2022201318


## 问题描述

论文列表包含 3139 篇 TMLR 2024 的论文和 7387 篇 ICLR 2024 的论文，共计 10526 篇。目标是完成以下任务：

1. **作者姓名格式转换**：将作者姓名简写，名在前，姓在后。
2. **合并作者字符串**：将作者姓名合并为一个字符串，作者之间用逗号连接，若多于三个作者，用 "et.al" 表示剩余作者。
3. **关键词提取**：为 TMLR 的论文根据摘要提取三个英文关键词（ `keywords` 字段为空）。
4. **代码链接提取**：从摘要中提取代码链接（如 GitHub 链接），存储在新字段 `Code_src` 中；如果没有链接， `Code_src` 字段为空。
5. **摘要扩展**：基于摘要内容扩展一个简单的 `Introduction` ，简要介绍论文的背景、研究问题、方法和主要贡献。
6. **论文主题分析**：根据论文关键词对论文进行聚类分析，并生成 `Topic` 字段，标明研究主题（如 "生成模型"、"图像识别"、"强化学习" 等）。

注释：为清晰展示每一任务的处理结果，本次实验选择分步骤独立处理，不将结果文件合并，也不在之前结果基础上继续执行。保留中间文件和日志信息，以体现实验过程。

# 实验过程与解决方案

## 任务 1 和 任务 2：作者姓名格式转换与字符串合并

### 解决方案

1. **姓名格式转换**：将作者姓名从原格式转换为 "名在前，姓在后"，并简写。
2. **合并字符串**：将转换后的作者姓名合并为一个字符串：
    - 作者之间用逗号连接。
    - 若作者超过三个，用 "et.al" 表示剩余作者。

### 实验方法

- 使用 `zhupuai` 大模型进行批量处理。
- **并行优化**：采用 `concurrent.futures` 进行并行处理，并设置最大并行度为 10。
- 提示设计：
    i. 明确要求返回 JSON 格式结果。
    ii. 提醒输出结果不可包含多余注释或解释。
    iii. 多次强调合并规则，确保超过三个作者时正确使用 "et.al"。
- 使用 `zip` 函数确保论文与作者匹配。

### 实验日志

- **错误日志**：记录多写作者错误，分析发现主要是大模型对复杂姓名格式的误解导致。
- **解析优化**：根据错误日志修正提示，减少错误率。

# 任务 3：关键词提取

## 解决方案

为 TMLR 的论文从摘要中提取三个英文关键词。

## 实验方法

1. 针对 `keywords` 为空的论文，将摘要输入大模型并请求生成关键词。
2. **提示设计**：要求关键词需为英文。
3. **容错处理**：
   - 记录正确日志和错误日志。
   - 错误主要源于摘要中包含 Markdown 格式公式，导致解析困难。
   - 选择直接保留公式的 Markdown 格式。
4. **并行处理**：使用 `concurrent.futures`，并设置最大并行度为 10。

## 实验日志

- 错误率较低，主要集中于公式解析问题。
- 保留部分错误日志以供后续优化。

# 任务 4：代码链接提取

## 解决方案

从摘要中提取代码链接（如 GitHub 链接），存储在 `Code_src` 字段；若未提供链接，则 `Code_src` 字段为空。

### 实验方法

1. 提示设计：
    - 要求返回包含所有代码链接的 JSON 格式结果。
    - 若无代码链接，返回空列表。
2. **容错处理**：
    - 记录正确日志和错误日志。
    - 分析错误日志，发现主要原因是部分摘要提到多个代码链接。
3. 修改提示模板，确保返回完整的代码链接。

### 实验日志

- 错误日志较少，主要为非网址信息导致。
- 修正后效果显著。

## 任务 5：摘要扩展

### 解决方案

基于摘要内容扩展一个简单的 `Introduction`，概述论文背景、研究问题、方法和主要贡献。

### 实验方法

1. 提示设计：
    - 要求返回包含 Markdown 格式公式的 JSON 结果。
    - 直接保留 Markdown 内容，避免复杂的正则解析。
2. **容错处理**：
    - 记录错误日志，发现部分摘要因涉及敏感内容被拒绝处理。

- 使用 GPT-4o 单独生成被拒绝摘要的扩展内容。

### 实验日志

- 两条摘要因敏感内容被拒绝处理，使用替代方法生成扩展内容，确保数据完整性。

## 任务 6：论文主题分析

### 解决方案

根据论文关键词对其进行聚类，生成 `Topic` 字段并标明研究主题。

### 实验方法

尝试两种方案：

1. **方案 1**：
   - 批量提取关键词后，大模型生成若干主题标签（共 106 个）。
   - 使用大模型对标签进行聚类，归为 ML、DL、AI 三类。
   - 每篇论文最多分配 5 个标签，按标签投票决定最终分类。

   **缺点**：
   - 标签数量过多，大模型可能出现幻觉。
   - 多标签影响分类判断。
2. **方案 2**：
   - 在标签聚类为 ML、DL、AI 三类后，直接对论文进行分类。
   - 使用大模型对每篇论文关键词进行聚类，生成单一分类标签。

### 实验日志

- **方案对比**：
  - 方案 1 结果中存在少量未定义分类。
  - 方案 2 更为直接，分类效果更优，且结果一致性更高。

# 实验结果与总结

## 总体总结

通过多次实验和日志记录，完成了 TMLR 和 ICLR 2024 论文数据的处理和分析。实验结果清晰展示了每个任务的处理方法和改进过程。

## 关键点总结

1. **并行优化**：利用大模型的并行处理能力，显著提高了处理速度。
2. **提示改进**：通过日志分析不断优化提示模板，减少解析错误。
3. **容错处理**：记录正确日志和错误日志，确保实验过程透明且可复现。

## 文件列表

1. **batch_requests.jsonl**
   - 该文件包含批量请求数据，通常用于向大模型接口发送请求数据的批处理。
2. **batch_requests2.jsonl**
   - 这是第二个批量请求文件，可能是不同的数据集或请求集合，按类似格式保存。
3. **classified_papers.json**
   - 任务6中间结果
4. **cleaned_papers_with_intro.json**

- 任务5中间结果，包含清洗过的论文数据，并为每篇论文扩展了简短的“Introduction”部分，概述了论文的背景、研究问题、方法和主要贡献。

5. **cleaned_papers_with_introductions.json**
   - 任务5最终结果，包含了不被zhupuai认为非法的摘要
6. **clustered_papers.txt**
   - 任务6中间结果 该文本文件包含了聚类分析结果，论文已根据主题或关键词被分组。
7. **error_log_code.txt**
   - 任务4错误日志
8. **error_log_keywords.txt**
   - 任务3错误日志
9. **error_log_tags.txt**
   - 任务1、2错误日志
10. **logs111.txt**

- 任务5综合日志

11. **model_response_log.txt**

- 任务1、2完整日志

12. **model_response_log_code.txt**

- 任务4完整日志

13. **model_response_log_intro.txt**

- 任务5完整日志

14. **model_response_log_keywords.txt**

- 任务3完整日志

15. **model_response_log_tags.txt**
    - 任务5中间过程日志
16. **pro1_2.ipynb**
    - 任务1、2代码
17. **pro3.ipynb**
    - 任务3代码
18. **pro4.ipynb**
    - 任务4代码
19. **pro5.ipynb**
    - 任务5代码
20. **pro6.ipynb**
    - 任务6代码
21. **papers_with_classes.json**
    - 任务6中间结果
22. **papers_with_code.json**
    - 任务4最终结果
23. **papers_with_intro.json**

- 任务5最终结果

24. **papers_with_keywords.json**

- 任务1、2最终结果

25. **papers_with_tags.json**

- 任务6中间结果（方法1）

26. **paper_metadata_1212_10k.json**
    - 原始文件

# 下一步优化方向

1. **进一步优化提示模板**，减少复杂内容（如公式和特殊字符）导致的解析错误。
2. **探索更高效的并行处理方法**，进一步提高大规模数据处理效率。
3. **标签生成优化**，探索减少标签幻觉的生成方法，提升分类准确性。