新闻批量处理系统

1. 项目背景和目标

本系统为新闻批量处理系统，主要用于处理光伏与储能行业相关新闻，包括信息筛选归类、去重以及内容提炼转写功能。系统支持多语言新闻处理，能够根据预设规则进行归类和处理，最终输出结构化、专业化的新闻摘要信息。

2. 功能需求

2.1 信息筛选归类

2.1.1 功能描述

- 系统接收JSON格式的新闻数据文件
- 提取文件中的标题、时间、URL和内容字段，并去除无效字符
- 创建26种语言的关键词库
- 匹配关键词数量小于等于5个的筛选，不要删除，归类为“人工复核”分类并进入后面的流程进行处理输出，再对关键词大于5个的文章进行分类
- 根据预设关键词库把匹配到的关键词以及标题发给AI结合分类规则定义综合分析对新闻进行自动归类。
- 处理后的内容用大模型或者翻译API翻译成英文输出到本地JSON数据库文件.

说明：

需要读取完整关键词，比如Industrial and commercial energy storage,不能只读取energy storage

2.1.2 新闻分类规则

系统将新闻分为以下7类：

分类	定义	关键词示例
政策动态	涉及政府或官方机构发布的政策、法规、规划、补贴调整、能源目标设定或战略合作等内容，通常与宏观能源转型或行业指导相关。	政策、规划、法规、机制、目标、补贴、税收、市场机制、能源转型、审批流程

项目动态	涉及具体能源项目的规划、审批、建设、并网或调试等进展，通常聚焦于项目本身的实施细节。	项目、投资、建设、开工、规模、示范、商业模式、储能、电力站、合作
企业动态	涉及特定企业的投资、并购、技术研发、产品发布或市场策略等企业行为。	并购、投资、合作、收购、出口、供应链、战略合作
电力市场	涉及能源市场的运行数据、交易结构、收入变化、市场份额或行业趋势，通常与市场经济行为或电力系统运营相关。	市场机制、电力调度、电价、容量市场、清洁能源优先、灵活性服务
经贸环境	涉及国际贸易、投资、产业链合作或经济环境对能源行业的影响，通常聚焦于跨国或跨地区经济活动。	贸易、关税、跨境投资、市场准入、出口政策
市场分析	聚焦于能源行业或相关领域的市场运行状况、趋势预测及经济行为，强调通过数据和趋势分析揭示市场规模、增长率、需求变化、竞争格局、装机容量、价格波动或行业结构变化等特征，通常不涉及具体项目实施、政府政策制定或单一企业行为。	市场规模、市场趋势、需求、竞争格局、装机量、价格趋势
其他	与以上分类不相关的部分判定为其他分类	

- 新闻分类页面管理以上分类。

2.2 咨询去重

2.2.1 功能描述

- 系统维护一个本地JSON数据库文件，用于存储归类后的新闻内容
- 数据库保存最近七天的内容，超过七天的数据自动删除
- 系统对新入库的内容进行去重分析，识别内容相似度达到90%的新闻
- 相似度高的内容在Web界面进行聚类显示(分别区分展示七天内重复的新闻和今日导入的重复新闻)，展示项有标题、时间、URL和内容字段以及导入数据的日期时间五个字段。
- 用户可以手动对重复内容进行删除操作：

1. **去重逻辑**

1.今日导入的数据例如200条，首先这200条要进行重复性匹配，内容重复度大于90%的列出来在web界面显示，web界面有一个删除重复数据按钮，点击按钮可以把发布时间相对晚的所有重复新闻删除。比如以下重复新闻，A的发布时间早于B,把B删除。

- 标题：A
- 发布时间：2025年5月1日
- 导入时间：2025年5月1日
- URL：xxxx
- 正文：xxxx
- 标题：B
- 发布时间：2025年5月2日
- 导入时间：2025年5月2日
- URL：xxxx
- 正文：xxxx

以上新闻相似度95%

2. 今日导入的数据例如200条，过去六天导入的总数据为1000条。把今日导入数据与过去六天的数据进行内容重复度匹配，内容重复度大于90%的列出来在web界面显示，web界面有一个删除重复数据按钮，点击按钮可以把今日导入的所有重复新闻删除。比如以下重复新闻，C和D重复，D是今日导入的数据，把D删除。

- 标题：C
- 发布时间：2025年5月1日
- 导入时间：2025年5月1日
- URL：xxxx
- 正文：xxxx
- 标题：D
- 发布时间：2025年5月2日
- 导入时间：2025年5月2日（今日）
- URL：xxxx
- 正文：xxxx

以上新闻相似度95%

注意事项：数据库只保存最近七天的内容，超过七天的数据自动删除。

2.2.2 处理规则

- 采用内容相似度分析，阈值设置为90%
- 相似内容以聚类方式展示，便于用户判断和处理
- 支持用户手动删除操作

2.3 咨询提炼转写

2.3.1 功能描述

- 用户可在信息筛选归类界面选择新闻并点击提炼转写按钮
- 系统自动执行工作流，对选中的新闻进行提炼转写
- 转写后的内容存入第二个JSON本地数据库
- 用户可在第三个界面查看提炼转写后的内容

2.3.2 摘要提取规则

原文概括要求：

- 对新闻内容进行详细概括，确保准确反映原文信息
- 数据与细节完整性：
  - 所有数据使用英文格式单位（MW、MWh、GW等），单位与数字间不需要空格
  - 确保核心信息和关键数据点完整呈现
- 概括精度要求：
  - 严格基于原文内容总结，保持信息准确性
  - 仅保留原文中提及的背景分析内容
- 避免预测性和推测性内容：
  - 删除未基于原文的预测性陈述
  - 所有数据必须带有明确的时间节点与来源

优化表达要求：

- 专业术语优化：
  - 使用光伏和储能行业的标准术语、简写和标准
- 精炼与紧凑表达：
  - 优化语句结构，确保表达简洁而详细
  - 内容分为三段：

1. 第一段：政策背景与目标
2. 第二段：市场动态与项目进展
3. 第三段：市场影响与未来预期

- 移除非原文内容：
  - 移除专家观点或未经证实的技术数据推测

标题改写要求：

- 结构与内容：
  - 前半句：明确主体或背景
  - 后半句：聚焦影响内容或核心机制
- 语言风格：
  - 专业、规范，符合能源行业表达习惯
  - 避免夸张、情绪化修饰语
  - 控制在25-35字之间

输出格式要求：

1. 原始概括：基于原文内容的完整概括
2. 标题：精炼的标题，突出新闻核心内容
3. 优化后表达：
   - 简洁、连贯，符合行业规范
   - 每段核心要点清晰，信息密度高
   - 输出语言要求： 最后的输出的文本内容为为中英文

语言与透明度要求：

- 语言要求：
  - 科学严谨，符合中文语法规范
  - 表达准确、专业，不偏离原文
- 数据与时间节点：
  - 所有数据必须带有明确的时间节点和来源

技术说明：建议调用AI大模型实现摘要提取功能。

3. 前端界面需求

提供四个主要界面：

1. **信息筛选归类界面**：
    - 显示提取和归类后的新闻内容
    - 提供文件上传功能
    - 显示分类结果
    - 提供提炼转写功能入口
2. **咨询去重界面**：
- 聚类显示相似度高的新闻内容
- 提供手动删除操作功能
- 显示相似度分析结果
3. **咨询提炼转写界面**：
- 显示经过提炼转写后的新闻内容
- 提供查看原文功能
4. 配置管理页面
- 分类和关键词管理
- 提示词管理

4. **数据管理**

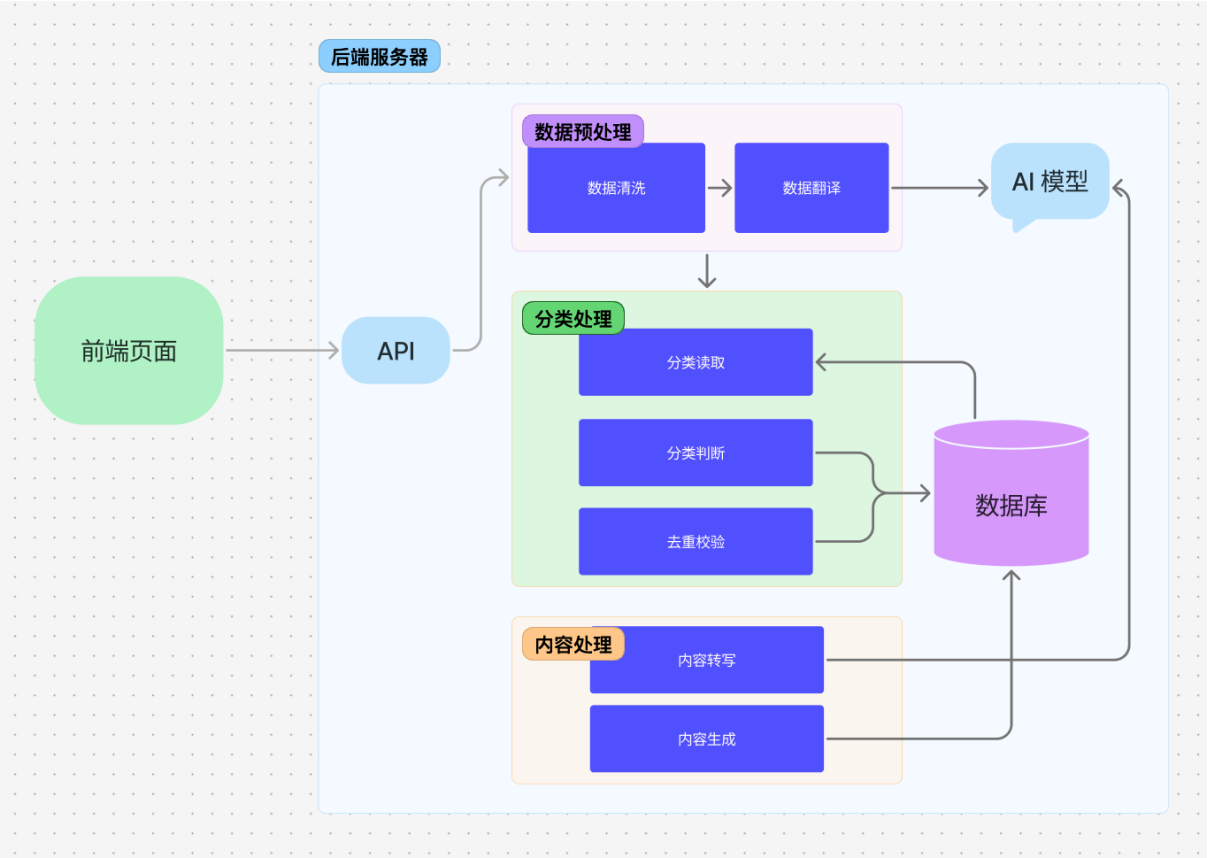
使用Nosql数据库存储数据：

1. **归类集合**：
    - 存储归类后的新闻内容
    - 保存最近七天数据，超时自动删除
    - 包含标题、时间、URL、内容、分类等字段
2. **提炼转写集合**：
- 存储经过提炼转写后的内容
- 包含原始概括、优化后表达、改写标题等字段

3. 技术架构

- 前端：React

- 后端：Python
- 数据处理：NLTK/SpaCy（文本处理）
- AI翻译：大语言模型API
- 数据存储：MongoDB（存储新闻数据和处理结果）
- **系统流程图：**



4. 项目实施计划

- **阶段一：后端系统开发 (2周)**：开发后端service，整合业务逻辑提供前端使用。
- **阶段二：前端页面开发 (2周)**：开发前端页面，集成后端service。
- **阶段三：系统测试 (1周)**：对系统进行全面测试。

5. 待确定部分

1. **部署方式：云端部署 or 本地部署？**

2. 大模型：使用哪个大模型api？
3. 文本输出：中英文？

6. 沟通机制

- 每周进行项目进度评审
- 及时反馈系统分类准确率和翻译质量

7. 交付物与验收标准

- **交付物:**
  - 完整的系统代码
  - 系统部署文档
  - 关键词库维护指南
  - 用户操作手册

8. 项目预算与开发周期

- 项目预算：11000元
- 开发周期：5周