实验题目：医学文献的结构化解析实践

1. 实验背景与总体目标

真实的医学 PDF 文献充斥着非结构化的长难句、复杂的排版和图表内容。在利用文献数据构建智能问答系统时，原始的文献难以被直接利用，需从中提取出高价值的知识，并被拆解为结构化的数据形式。

本实验要求大家利用 AI 大模型（如 ChatGPT、Gemini、豆包、DeepSeek 等）、大模型编程接口（OpenRouter）或编程助手（如 GitHub Copilot、Claude Code、Cursor 等），从零开始编写一套 Python 流水线。最终目标是实现文献的：1) 知识建模、2) 信息抽取、3) 知识融合和 4) 智能应用，将这两篇文献（任选一篇即可）转化为可供检索的结构化数据（JSON），为后续构建专病智能体提供高质量的外部知识库。

注意，使用 AI 大模型应用的话，需要每一次交互生成的代码复制到本地运行；使用编程接口的话，需要先编写（或生成）API 调用的代码，实现输入 prompt，输出代码；使用编程助手的话，可以直接在本地编程环境中与 AI 进行交互。

2. 实验流程与任务分解

阶段一：AI 大模型选型与底层数据获取

任务描述：PDF 是人类友好但机器困难的文件格式，包含大量隐藏的排版、换行符和页眉页脚干扰。本任务需要通过与 AI 进行多轮对话，寻找最合适的 Python 第三方库，并编写脚本将系统性硬化症（Systemic Sclerosis，SSc）相关文献 PDF 中的纯文本先提取出来。

具体要求：

1. 利用 AI 对比 PDF 处理的 Python 库（如 PyPDF2、pdfplumber、PyMuPDF 等）的优缺点，选择最适合从复杂的文献中提取纯文本的库。
2. 编写 Prompt（提示词）让 AI 生成读取 PDF 指定范围文本内容的代码（可设置文本范围，如特定章节、特定页码、排除 xx 区域等），并在你电脑中

运行代码，保存好读取的文本文件。

3. 利用 Python 字符串的 replace() 或正则匹配，剔除跨页造成的断词（如 "inter-\nstitial" 需还原为 "interstitial"）以及参考文献角标（如 "disease[12]"）。
4. **将读取和处理后的文本存储在文件中（命名为：ssc_text.txt）**

阶段二：知识建模与自动事件抽取

任务描述：传统的实体抽取仅提取疾病名或药物名，无法呈现临床试验的动态过程。本阶段需要在知识建模和知识抽取中明确加入对事件的建模和对事件的抽取。

具体要求：

1. **事件建模定义：**针对系统性硬化症（SSC），定义具体的临床事件结构。例如，针对文献二，可建模为：{事件类型: "临床干预", 干预手段: "onabotulinumtoxinA", 目标症状: "reduced oral aperture", 结果: "..."}。针对文献一，可建模为：{事件类型: "并发症特征", 疾病: "SSc-ILD", 初始症状: "...", 影像学表现: "..."}
2. **基于规则的事件抽取：**使用 for 循环遍历按句号拆分的文献文本。结合条件分支 if-elif 和字符串的 .find() / .startswith() 操作，让 AI 辅助编写逻辑树。例如，当句子中同时出现 "treatment" 和 "improvement" 时，触发干预事件抽取的逻辑分支。
3. **精准提取带数值的复杂短语：**例如利用字符串切片或循环，提取出文献二中 "16 units of onabotA" 这样的精准用药剂量与实体关联。
4. **将设计出的事件模型存储在文件中（命名为：ssc_event_model.md）。将提取出的所有事件存储在文件中（命名为：ssc_event_data.json）**

阶段三：基于字符串相似度的事件融合

任务描述：在长篇文献的 Abstract、Methods 和 Results 部分，同一个医学事件会被多次以不同维度提及。单纯的罗列会导致数据冗余和冲突。

具体要求：

1. 在知识融合中必须加入对事件的融合。

2. **编写循环比对逻辑：**创建新的结构（如字典的嵌套），将提取出的事件实体（如不同句子中提取到的 OnabotulinumtoxinA 疗效数据）进行遍历比对。
3. **消除数据歧义：**利用字符串相似度比对，将描述同一医学事件的多条记录合并，更新其属性（例如将散落在各处的症状如 "Raynaud phenomenon", "skin swelling" 统一挂载到 "SSc initial symptoms" 这一核心事件节点下）。
4. **将融合后的事件数据存储在文件中（命名为：ssc_event_clean.json）。**

阶段四：结构化输出与智能体验证

任务描述：对于融合后的事件，编写一个简易的终端检索系统，验证数据质量。

具体要求：

1. 编写一个 while True 的交互循环，模拟专病智能体的底层查询：用户在终端输入 "口部张开困难" 或 "ILD 风险因素"，程序通过遍历生成的结构化字典，返回关联的治疗事件或特征数据。当输入 "exit" 时退出。
2. **将交互的过程截图保存**

3. 实验提交物描述

1. **实验报告：根据《实验报告 1:医学文献的结构化解析》模板中要求的内容进行填写和截图展示，写清楚姓名和学号，命名为《姓名-学号-实验报告 1》**
2. **实验过程文件，包括：**
   a) **文本提取后的文件（ssc_text.txt）**
   b) **事件模型描述（ssc_event_model.md）**
   c) **提取出的事件数据（ssc_event_data.json）**
   d) **融合后的事件数据（ssc_event_clean.json）**
   e) **模拟的智能体交互过程截图（agent_sim.png）**
3. **以上报告和文件，打包成压缩文件，命名为：姓名-学号-实验 1.zip，钉钉发送给老师**