01 项目概述

本项目旨在验证文档加载器对文字型 PDF 的解析能力。文字型 PDF 的页面由可提取的文字层构成，PyPDFLoader 读取时直接命中文字层，中文内容完整、页码元数据齐全。

02 目标与范围

覆盖单文件加载、按页切分、元数据读取三个场景。范围不含图片型扫描件，该场景由独立的示例文件演示。

03 技术选型

解析层选用 pypdf；加载层选用 LangChain 的 PyPDFLoader；验证方式为逐页文本抽取与元数据核对。

04 数据准备

语料为本报告自身，共十二章，每章一页。中文段落、数字编号、标点混排，用于检验解析的完整性与顺序。

05 加载流程

PyPDFLoader 构造时传入文件路径，load() 返回 Document 列表，每页一个元素，metadata 携带 source 与 page。

06 元数据说明

source 为加载时的文件路径，page 从 0 开始计数。引用溯源与按页过滤都依赖这两个字段。

07 文字型与图片型

文字型 PDF 含文字层，可直接抽取文本；图片型 PDF 页面是整张图片，没有文字层，抽取结果为空，需要 OCR 或多模态模型。

08 常见问题

中文乱码多因字体嵌入缺失；表格被拆成碎字符串属正常现象；加密 PDF 需在构造参数中传 password。

09 性能表现

按页懒解析，单页开销稳定；千页级文档建议配合目录批量加载与进度条使用。

10 结论

文字型 PDF 走 PyPDFLoader 是标准路径，文字与元数据双通道完整，可直接进入切分与向量化环节。

11 后续计划

补充图片型 PDF 的 OCR 方案对比：PaddleOCR、Tesseract 与多模态模型三条路线。

12 附录

本报告同时作为示例数据使用，任何解析结果均应能还原本段文字的内容与顺序。