**关于知识库项目中数据预处理的经验总结与方法论沉淀**

**数据清洗**

核心目标：将来源各异、格式不统一的原始文档，处理为干净、标准化的纯文本或结构化数据，为后续的向量化处理奠定基础。

**无用信息污染（页眉/页脚/页码）**

* 场景：技术方案书，包含大量重复、与核心知识无关的页码。
* 解决方案：
* 正则表达式批量删除；
* 手动将word的页码、页脚等关闭/删除；
* 工具与技术：python re库

**表格格式/单位统一**

* 场景：设备点位表（IO点表）格式不同意，表头、单位有差异：
* 点位编号（Tag Name）
* 点位描述（如“1号熔铸炉东侧温度”）
* 物理量程（如0-1000℃）
* 单位（℃, %, m）
* 所属设备
* 数据采集频率
* 解决方案：手动校准
* 工具与技术：无

**客户隐私信息屏蔽**

**其他**

* 文档编码格式需要统一，工具：notepad--

**解析与切片：**

**PDF\word统一格式化为txt**

适合复杂文档（技术手册、规格功能说明书、审计报告等），与上下文、逻辑强相关的内容；

1. pdf（转word）\word先将图表转为excel数据（调用基于ApachePOI的接口）；
2. 统一转markdown；
3. 图片转存至minio库，用生成的url替换路径链接；
4. 所有文档都转为纯文本格式后，手动分段；
5. 父段：[主题XXX]  
   子段：/n拆分
6. 导入dfiy知识库，父子分段；

**文本切片（Chunking）**

**问题：**

直接按固定长度（如512字）切割，会把一个完整的操作步骤或概念切到两个块里，导致检索信息不完整。

**方法：**

* 初级： 使用递归字符文本分割器，尝试不同的chunk\_size（e.g., 200-1000）和 chunk\_overlap（e.g., 50-100）。这是Dify等平台的默认方法。
* 高级： 使用语义分割器。利用NLP模型识别段落、标题（# ## ###）、句子边界进行切割。你可以说你探索过这种方法，但由于计算资源和小公司场景，最终选择了基于规则的分割，这非常真实。
* 自定义规则： 针对公司文档特点定制规则。例如，你们的文档都用 “步骤1：”开头，你就可以按这个模式来分割，保证每个操作步骤的完整性。

**结论：**

强调没有银弹，最终采用的策略是：优先按标题结构分割，其次按段落，最后再按固定长度兜底，并设置了重叠字符以确保上下文连贯。