Week 7 | 课时 3 | 从 section 到 chunk：结构感知切片与 overlap 边界

Table of contents

先尊重文档结构，再处理长度

这一讲处理 Week7 最容易被低估的环节：

chunking 不是把文本切成差不多长，而是把 section、表格、页码和上下文边界转成可索引单元。

进入课时 4 返回 Week7 总览

下载讲义

提供适合离线阅读的 PDF 版和适合批注整理的 Word 版。

PDF 版 · 打印 / 离线阅读 Word 版 · 批注 / 二次整理

这节课解决什么问题

固定长度切片看起来简单，但它会制造很多隐性事故：

- 把标题和正文切开
- 把步骤 1、2、3 切到不同 chunk

- 把表头和数据行拆散
- 把错误码和处理建议拆开
- overlap 过大导致重复噪声

Week7 的默认策略是 section_aware_v1：

heading-aware + section-aware + table-aware + page-aware + token-aware。

参考学习时间

45–55 分钟

学完这一讲，你应该能做到什么

1. 解释为什么固定长度切片不适合作为主路径。
2. 设计 section-aware chunking 的基本规则。
3. 判断 overlap 什么时候有用、什么时候有害。
4. 说明 chunk_strategy_version 和稳定 chunk_id 的作用。

本课产出

- docs/blueprints/week07/chunking-strategy-v1.md
- contracts/data/document_chunk.schema.json

fixed-size vs section-aware

切片方式	优点	风险	Week7 判断
fixed-size	实现简单，长度均匀	容易切断语义、表格、步骤和标题上下文	只能作为 fallback
recursive text splitter	比固定长度更稳	仍可能忽略 parser 结构	可作为辅助
section-aware	尊重文档层级和 page/table 边界	需要更完整 meta-data	Student Core 默认
semantic splitter	可能更贴近语义断点	依赖 embedding，阈值和语言敏感	延后到 Week8+ 评估

1. section_aware_v1 的默认规则

```
1. 先按 parser 输出的 section_path 分组
2. 同一 section 内合并过短 paragraph
3. 表格默认保持完整；超长表格拆分时保留 header
4. 超长 section 再按 token budget 切分
5. overlap 只在同一 section 内出现
6. 每个 chunk 记录 overlap_prev / overlap_next
7. 每个 chunk 继承 source_fingerprint、doc_version、page_no、section_path
```

2. chunk ID 必须稳定

禁止使用随机 UUID、时间戳或数据库自增 ID 作为 chunk 的业务 ID。

推荐公式：

```
chunk_id = sha256(
  doc_id
  + source_fingerprint
  + doc_version
  + chunk_strategy_version
  + section_id
  + chunk_index
  + content_hash
)[:32]
```

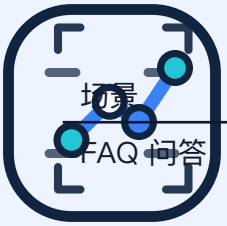
这样才能做到：

- 同一输入重复跑，ID 不变
- 改 `chunk_strategy_version`，chunk ID 变化
- 只改 `parse_run_id`，chunk ID 不变

3. overlap 的边界

Overlap 不是越大越安全。

场景	是否使用 overlap	原因
同一 section 内长段落切分	可以	保留上下文连续性
跨 section	不建议	会污染标题层级
表格	默认不用	容易重复表头和行
错误码列表	谨慎	过大 overlap 会制造重复命中



场景	是否使用 overlap	原因
FAQ 问答	按 Q/A 保持整体	不应把问题和答案拆开

4. chunk 合格的最低标准

检查	要求
非空	content 不能为空
可回源	有 source_fingerprint / doc_version
有结构	有 section_path 或合理 missing reason
可引用	至少能生成一个 evidence anchor
可回归	有 chunk_strategy_version
可过滤	有 content_type / license_tag / pii_flag

本课最重要判断

切片策略不是“为了凑 embedding 输入长度”，而是为了让 Week8 能检索到 **完整、可引用、可质检** 的证据单元。

自检清单

- 我知道 fixed-size chunk 的风险。
- 我能写出 section_aware_v1 的基本规则。
- 我知道 overlap 不能跨 section 滥用。
- 我能解释稳定 chunk ID 为什么重要。

最小行动命令

```
python -m pipelines.parse_normalize.chunking \
  --sections artifacts/week07/sections.json \
  --strategy section_aware_v1 \
  --max-tokens 700 \
  --overlap-tokens 80 \
  --out artifacts/week07/chunks.json
```