**خطر تقسیم‌بندی نادرست:** اگر ما متن را به صورت تصادفی یا بر اساس یک اندازه ثابت (مثلاً هر ۵۰۰ کاراکتر) تقسیم کنیم، این خطر بسیار بزرگ وجود دارد که یک "ماده" از "تبصره" مربوط به خودش جدا شود. اگر این اتفاق بیفتد، وقتی کاربر سوالی در مورد آن تبصره بپرسد، سیستم ما ممکن است کانتکست اصلی (ماده) را پیدا نکند و پاسخ اشتباه یا ناقصی بدهد.

**راه حل چیست؟** ما باید از یک روش تقسیم‌بندی (Chunking) هوشمند استفاده کنیم که تلاش می‌کند ساختار معنایی متن را حفظ کند. به جای اینکه فقط متن را بشکنیم، ما آن را به **"قطعات معنادار"** تقسیم می‌کنیم.

برای این کار، ما از ابزار داخلی و قدرتمند LlamaIndex به نام SentenceSplitter استفاده خواهیم کرد.

**تحلیل دقیق مشکل: چرا کد فعلی کافی نیست؟**

ابزار SentenceSplitter که ما در کد قبلی استفاده کردیم، یک ابزار **عمومی** است. این ابزار متن را بر اساس نشانه‌های نگارشی (مانند نقطه و پاراگراف) و یک اندازه ثابت (مثلاً ۱۰۲۴ توکن) تقسیم می‌کند. این ابزار هیچ درکی از ساختار منطقی و قانونی یک آیین‌نامه ندارد.

### راه حل: تقسیم‌بندی هوشمند مبتنی بر ساختار (Structure-Aware Chunking)

برای حل این مشکل، ما باید از یک روش تقسیم‌بندی سفارشی استفاده کنیم که ساختار منطقی آیین‌نامه شما را درک می‌کند. ما این کار را در دو مرحله انجام خواهیم داد:

1. **استخراج کل متن:** ابتدا، کل متن ۱۴ صفحه PDF را به صورت یک رشته متنی واحد و طولانی استخراج می‌کنیم.
2. **تقسیم‌بندی با قوانین مشخص:** سپس، به جای استفاده از SentenceSplitter ، ما خودمان این متن طولانی را بر اساس یک الگوی مشخص تقسیم می‌کنیم. الگوی ما، شروع هر "ماده" جدید خواهد بود. ما به برنامه می‌گوییم: "هر چیزی که از ابتدای یک 'ماده' شروع می‌شود تا درست قبل از 'ماده' بعدی، یک قطعه کامل و معنادار است."

این روش تضمین می‌کند که هر ماده و تمام تبصره‌های زیرمجموعه‌اش، همیشه در یک قطعه واحد باقی می‌مانند.