

Inteligência Artificial 2

Aula 12 – Parte 2: Experimento de Chunking

Prof. Aléssio Miranda Júnior
alessio@cefetmg.br

CEFET-MG - Campus Timóteo
Dep. Engenharia de Computação

Inteligência Artificial 2

- 1 Da Fonte ao Experimento
- 2 Escolha sem Vazamento
- 3 Clusters: Explorar, não Rotular
- 4 Laboratório

Parte 1: confiar na fonte

- Manifesto: responsável, finalidade, licença, PII, acesso e retenção.
- Extração por página, hash, versão e método.
- Texto auditável, não uma string sem origem.

Parte 2: testar o corte

- Estratégias concorrentes no mesmo corpus.
- Desenvolvimento escolhe; held-out informa o resultado.
- Recuperação, contexto e redundância decidem.

Não perguntamos “qual chunk size é o correto?”. Perguntamos:

Para este corpus e estas consultas, qual configuração recupera evidência citável com contexto útil e pouca redundância?

Controle o experimento: mesmo corpus, modelo de embedding, índice, Top-3 e regras de relevância. Varie apenas tamanho, overlap ou estratégia de chunking.

- **Desenvolvimento:** perguntas e evidências verificadas usadas para escolher tamanho, overlap e estratégia.
- **Held-out:** perguntas separadas antes do experimento; não orientam ajuste algum.
- Após congelar a escolha, execute held-out uma vez e reporte o resultado, inclusive falhas.

Usar a mesma consulta para escolher e celebrar a configuração superestima a qualidade da recuperação.

- **Recall@3**: ao menos uma evidência correta entrou no Top-3?
- **MRR**: em que posição aparece a primeira evidência correta? $1, \frac{1}{2}, \frac{1}{3}$ ou 0.
- **Precisão contextual**: ao ler os trechos, eles realmente sustentam a resposta à consulta?
- **Redundância**: o Top-3 traz três evidências ou cópias quase idênticas causadas por overlap?

Estratégia	Recall@3	MRR	Precisão contextual	Redundância
Fixa				
Recursiva				
Estrutural				

Decisão responsável

Se duas configurações empatam nas métricas, prefira a que preserva proveniência, custa menos e entrega contexto mais diverso.

- Agrupar embeddings sem supervisão pode expor tópicos recorrentes, duplicatas e texto extraído de forma anômala.
- Inspeção exemplos próximos ao centróide; dê nomes provisórios, se úteis.
- O resultado depende do embedding, da normalização e do número de grupos.

Cluster não é verdade de referência

Um grupo não prova que seus textos possuem a mesma intenção, categoria ou qualidade. Ele apenas sugere uma estrutura geométrica a investigar.

- Não forneça páginas esperadas ou rótulos de relevância ao algoritmo de clustering.
- Esses rótulos são supervisão e ficam reservados para avaliar recuperação.
- Não escolha chunking apenas porque a visualização “parece bonita”.

1. Gere índices fixo, recursivo e estrutural com a proveniência da Parte 1.
2. Ajuste a configuração usando somente desenvolvimento.
3. Congele a escolha e reporte Recall@3, MRR, precisão contextual e redundância em held-out.
4. Justifique a escolha, incluindo custo e limitações da extração.

Clínica extra opcional

Faça clustering exploratório somente depois da entrega principal e interprete-o criticamente.

1. Documento confiável começa com manifesto, governança e proveniência.
2. Chunking é uma hipótese mensurável, não uma receita universal.
3. Desenvolvimento escolhe; held-out mede a generalização da escolha.
4. Clusters ajudam a perguntar melhor, mas não substituem evidência rotulada.

Dúvidas?

Email: alessio@cefetmg.br

Aula: Segunda e Quarta-Feira, 7h as 8:40h (Sala 39)

WhatsApp: Dúvidas rápidas

Lembrete

Confira sua Issue no GitLab!

Próxima aula: [Arquitetura Transformer](#)

100% da Aula 12 Concluída!