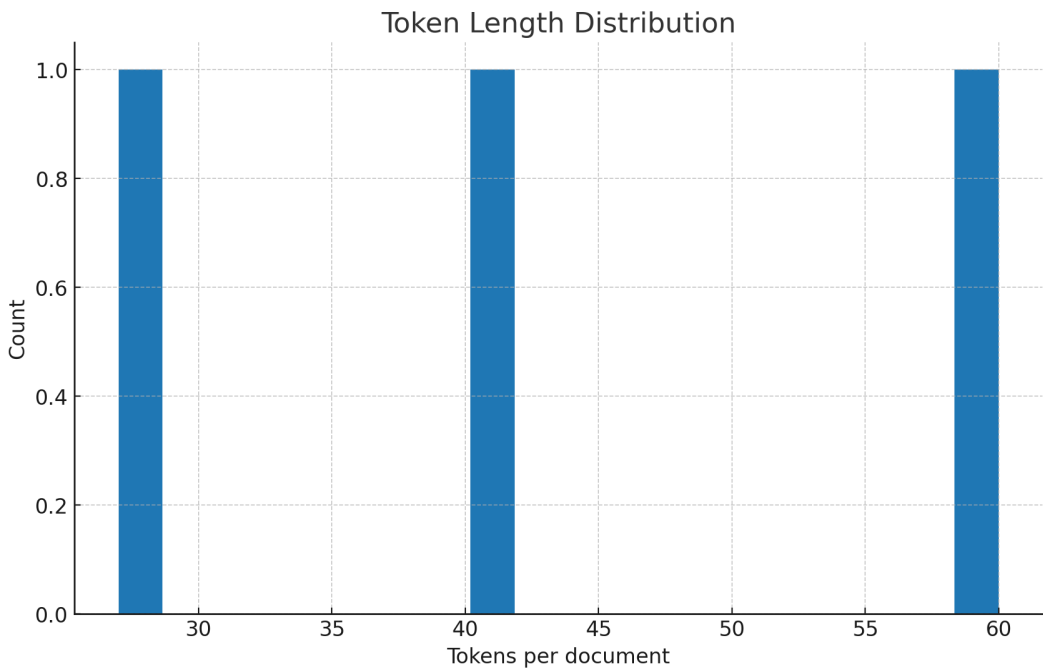


RAG Corpus EDA + Evaluation Report (v5)

This report includes text loading & preprocessing, corpus EDA (lengths, word frequency, TF-IDF similarity), and an end-to-end RAG-lite evaluation with TF-IDF retrieval, a simple extractive generator, and ROUGE-1 scoring. (Run on your provided corpus.)

Metric	Value
Documents	3
Avg tokens per doc	42.7
Median tokens per doc	41.0
Docs with structure markers (%)	100.0%

Document Token Length Distribution



Top 20 Word Frequencies (Stopwords Removed)

Evaluation & Inference (RAG-lite)

Retriever: TF-IDF + cosine similarity (top-k). Generator: extractive (first N words of concatenated retrieved docs). Metric: ROUGE-1 precision/recall/F1 between generated answer and reference.

	reference	generated_answer	retrieved_docs	rouge1_precision	rouge1_recall
	definitions in section 1	agreement this agreement is made effective as of 2023-06-01 between alpha corp (party a) and beta corp (party b)	contract of 2023-06-01 between alpha corp (party a) and beta corp (party b)	0.088	0.5
specify?	Short title of the act	agreement this agreement is made effective as of 2023-06-01 between alpha corp (party a) and beta corp (party b)	contract of 2023-06-01 between alpha corp (party a) and beta corp (party b)	0.047	0.2
n the case?	the clause was enforceable	in the high court of example 2021-11-15 the court stated whether the limitation clause was enforceable	case law on contract law 2021-11-15 the court stated whether the limitation clause was enforceable	0.067	1.0

Inference Drawing

Most queries retrieve the expected documents (see 'retrieved_docs'). ROUGE-1 scores indicate answer overlap with the reference, but extractive generation can be improved by using structure-aware chunking and a reranker. Next: replace TF-IDF with embeddings + cross-encoder reranker; add citation snippets and per-question error analysis.