Analytics Vidhya

Chunking & Tokenization

Instructors

Prashant Sahu

Manager - Data Science, Analytics Vidhya

Ravi Theja

Developer Advocate Engineer, LlamaIndex

What is Chunking?

Chunking refers to breaking text into "Chunks" or smaller segments, which are then vectorized and stored, boosting the efficiency of NLP tasks.

How to decide the size of Chunks?

1. **Fixed size Chunks**
   Characters, Sentences or Paragraphs
2. **Precision in Small Chunks**
   Enhances Match and Accuracy
3. **Noise in Large Chunks**
   Reduces Retrieval Accuracy
4. **Balance in RAG**
   Balance between comprehensiveness and precision.

How do you choose the right chunk size for your use case??

Chunking Demo

Tokenization

What is Tokenization?

The process of dismantling the sentences, paragraphs and articles into smaller chunks is called tokenization .

Sentence Tokenization	Word Phrases (N-Gram Tokenization)	Individual Word (Uni-Gram Tokenization)	Character Level Tokenization

Tokenization in LLMs

Subword Tokenization

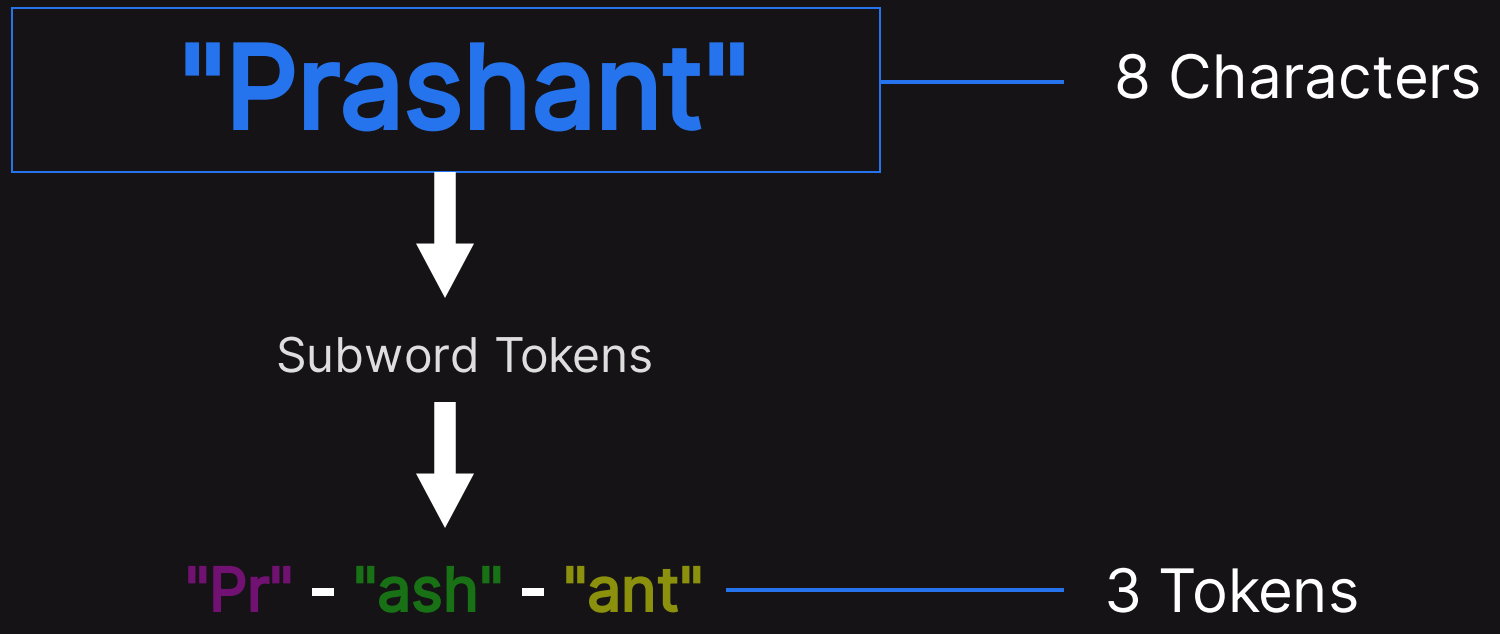
What is Subword Tokenization?

Subword tokenization breaks the words into known subparts from the dictionary, the model can handle words it hasn't seen before..

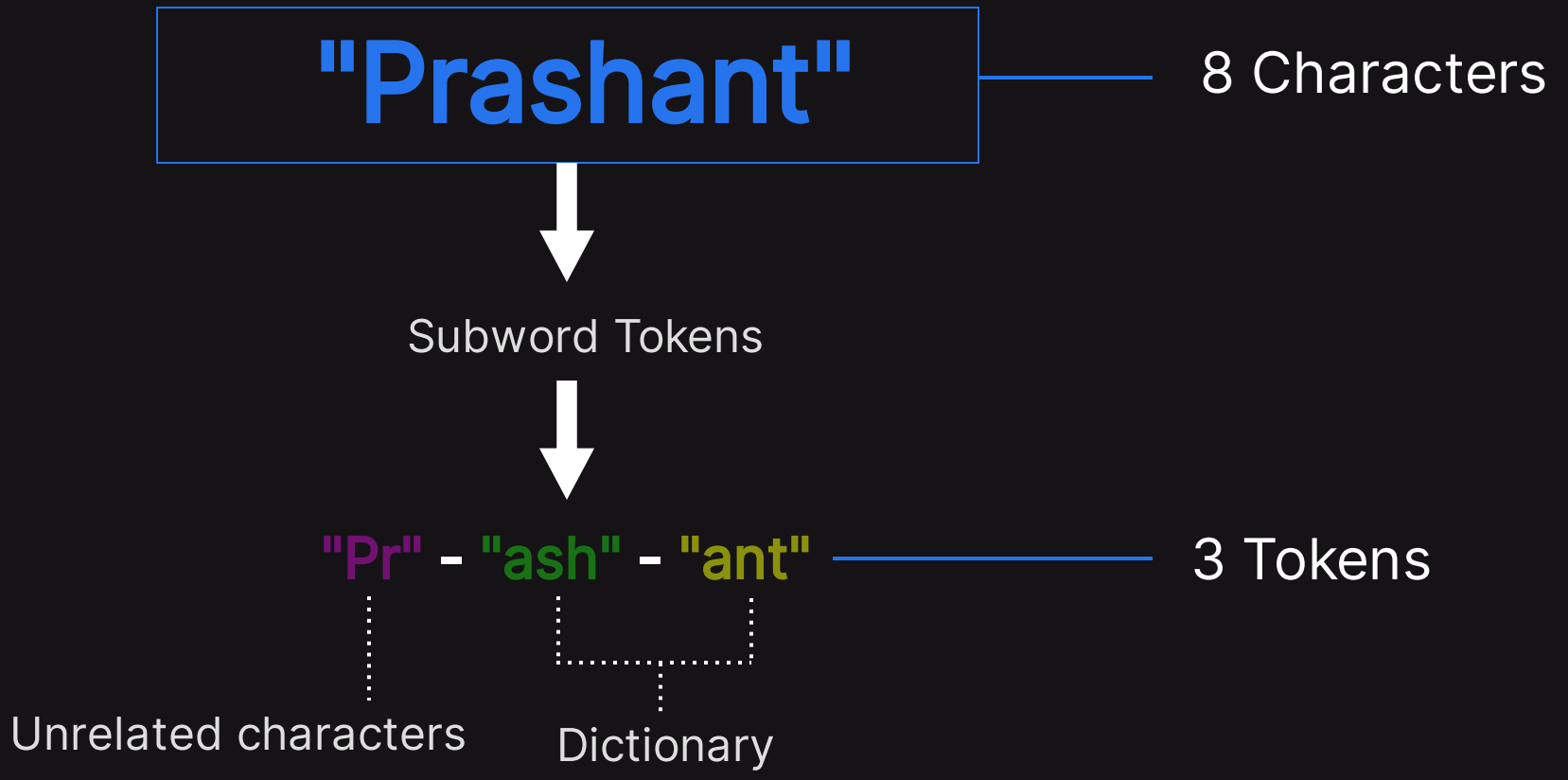
Subword Tokenization

Subword Tokenization

Subword Tokenization



Subword Tokenization



Types of Subword Tokenization

3 Types

Byte Pair Encoding (BPE)

WordPiece

SentencePiece

Types of Subword Tokenization

3 Types

Byte Pair Encoding (BPE)

- Merges the most frequent pairs of characters or subwords.
- Widely used in GPT and GPT-2, GPT-3.5 & GPT-4 models

WordPiece

SentencePiece

Types of Subword Tokenization

3 Types

Byte Pair Encoding (BPE)

- Merges the most frequent pairs of characters or subwords.
- Widely used in GPT and GPT-2, GPT-3 & GPT-4 models

WordPiece

- Uses a probabilistic approach to merge subwords.
- Used in BERT and DistilBERT models

SentencePiece

Types of Subword Tokenization

3 Types

Byte Pair Encoding (BPE)

- Merges the most frequent pairs of characters or subwords.
- Widely used in GPT and GPT-2, GPT-3.5 & GPT-4 models

WordPiece

- Uses a probabilistic approach to merge subwords.
- Used in BERT and DistilBERT models

SentencePiece

- Uses BPE or unigram models, processing text as Unicode character sequences.
- Used in T5 and AlBERT models

Tokenization in Open AI platform

Why is it important to know the total number of tokens in a chunk or document ??

Why is it important to know the total number of tokens in a chunk or document ??

1

Embedding Model

Important to understand the context limit, or maximum token capacity of these models

2

Text Generation Model

Models have a token limit for both context window and output responses

- **(Input) Context Limit/ Window** : Maximum number of tokens that the model can process as input.
- **(Output) Response Synthesis Limit** : Maximum number of tokens that the model can generate as output.

Embedding Models

Model Name	Input Context Limit
OpenAI: text-embedding-3-small and large	8191 Tokens
AWS: amazon.titan-embed-text-v2:0	8192 Tokens
Cohere: embed-english-v3.0	512 Tokens
Voyage: voyage-large-2	16000 Tokens
mxbai-embed-large-v1	512 Tokens
UAE-Large-V1	512 Tokens
bge-large-en-v1.5	512 Tokens
gte-large-en-v1.5	8192 Tokens

As of : 21-May-2024

Analytics Vidhya

Text Generation Models

Model Name	Context Window	Response synthesis Limit
GPT-3.5-turbo	16385 Tokens	4096 Tokens
GPT-4o	128000 Tokens	4096 Tokens
Claude 3 Opus	200K Tokens	4096 Tokens
Gemini 1.5 Pro	1M Tokens	8192 Tokens
Llama 3	8K Tokens	N/A
Mistral-7B-v0.2	32K Tokens	N/A
Microsoft: Phi-3	128K Tokens	N/A
Gemma-7B	8192 Tokens	N/A

As of : 21-May-2024

Analytics Vidhya

Pricing: Embedding Models

Model Name	Input Context Limit	Price per Million tokens
text-embedding-ada-002	8191 Tokens	$0.10
text-embedding-3-small	8191 Tokens	$0.02
text-embedding-3-large	8191 Tokens	$0.13
Cohere: embed-english-v3.0	512 Tokens	$0.1
Voyage: voyage-large-2	16000 Tokens	0.12

As of : 21-May-2024

Pricing: Text Generation Models

Model Name	Context Window	Price per Million tokens (Input)	Price per Million tokens (Output)
GPT-3.5 Turbo	16385 Tokens	$0.5	$1.5
GPT-4o	128K Tokens	$5.00	$15
Claude Opus	200K Tokens	$15	$75
Gemini 1.5 Pro	1M Tokens	$7	$21

As of: 21-May-2024

Thank You