Lecture - 7 : Tokenization

Data preparation and sampling

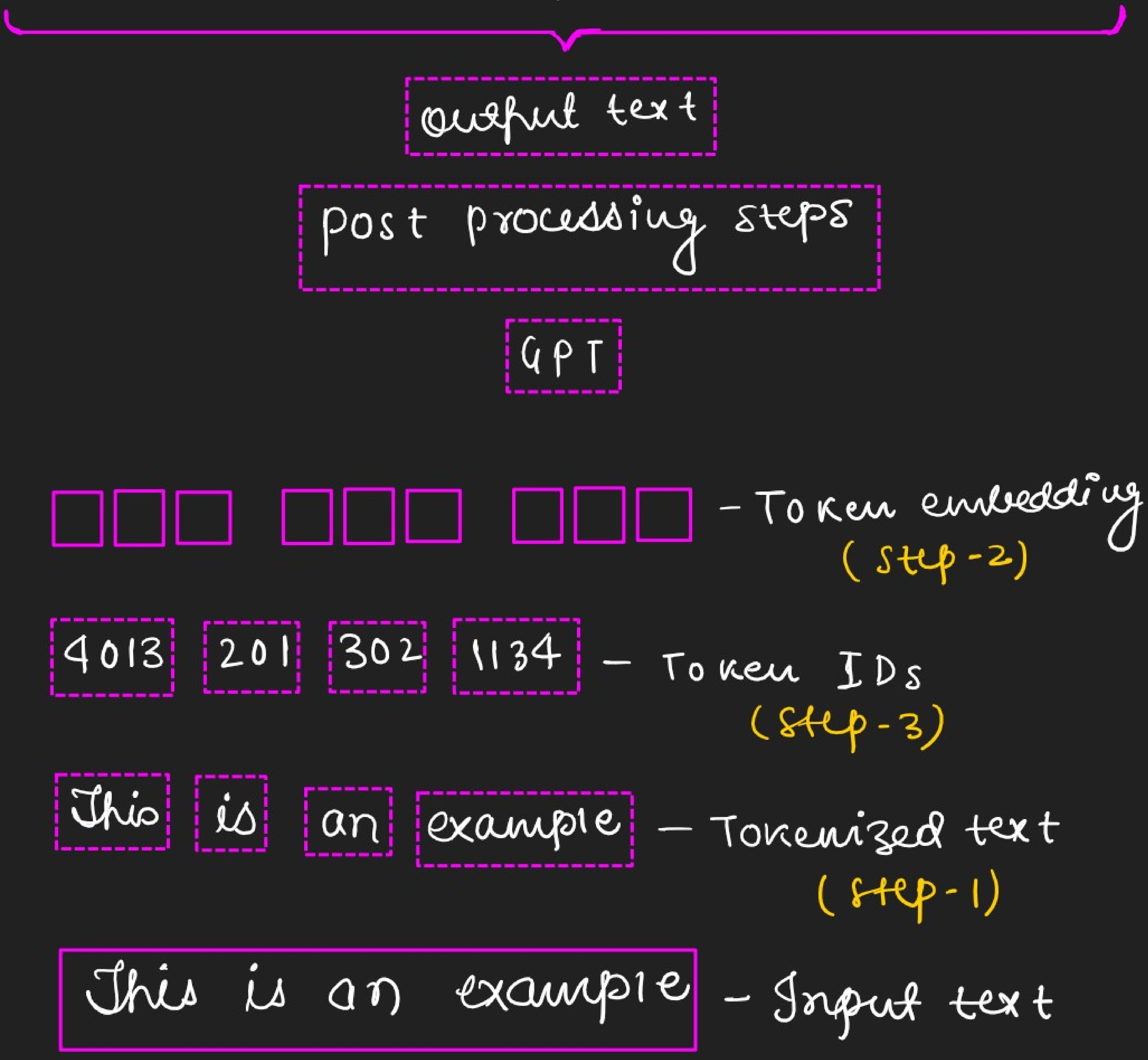
→ we will look at "Data preparation and sampling" in this lecture.

1. How do you prepare input text for training LLMS?

Step-1: (a.) Splitting text into individual word and subwords & token

Step-2: (b.) convert token into token IDs.

Step-3: (c.) Encode token IDs into vector representation.



output text

post processing steps

GPT

□□□ □□□ □□□ - Token embedding (step-2)

4013 201 302 1134 - Token IDs (step-3)

This is an example - Tokenized text (step-1)

This is an example - Input text

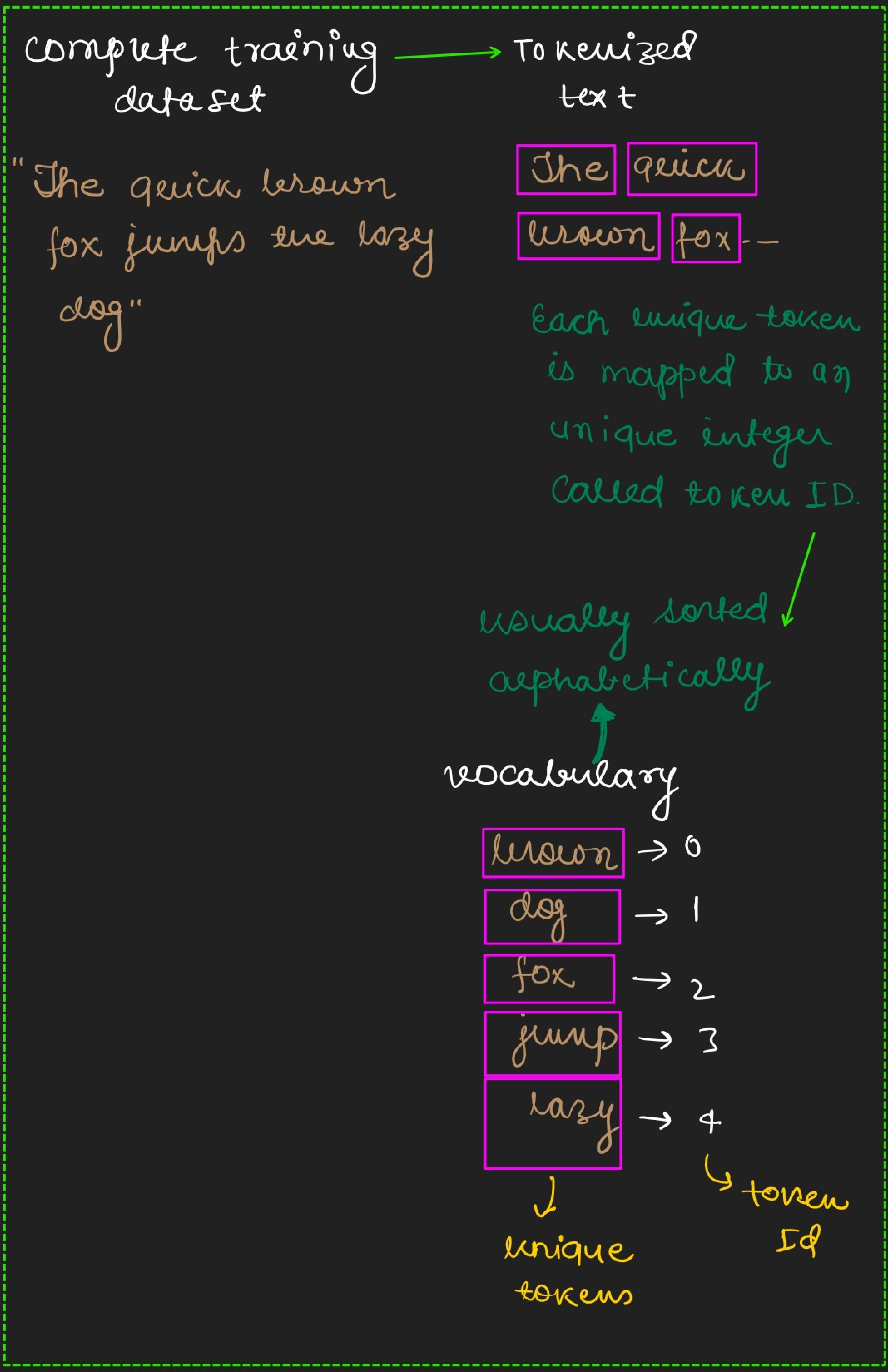
2. Let's us look at Step-1: Tokenizing Text

(a.) Dataset used: "The verdict" by Edith Wharton.

(b.) Download and load in python.

(c.) Tokenize the short story.
↳ use python's regular expression liberary.

(d.) convert Token into Token IDs


compute training dataset → Tokenized text

"The quick brown fox jumps the lazy dog"

The | quick | brown | fox --

Each unique token is mapped to an unique integer called token ID.

usually sorted alphabetically

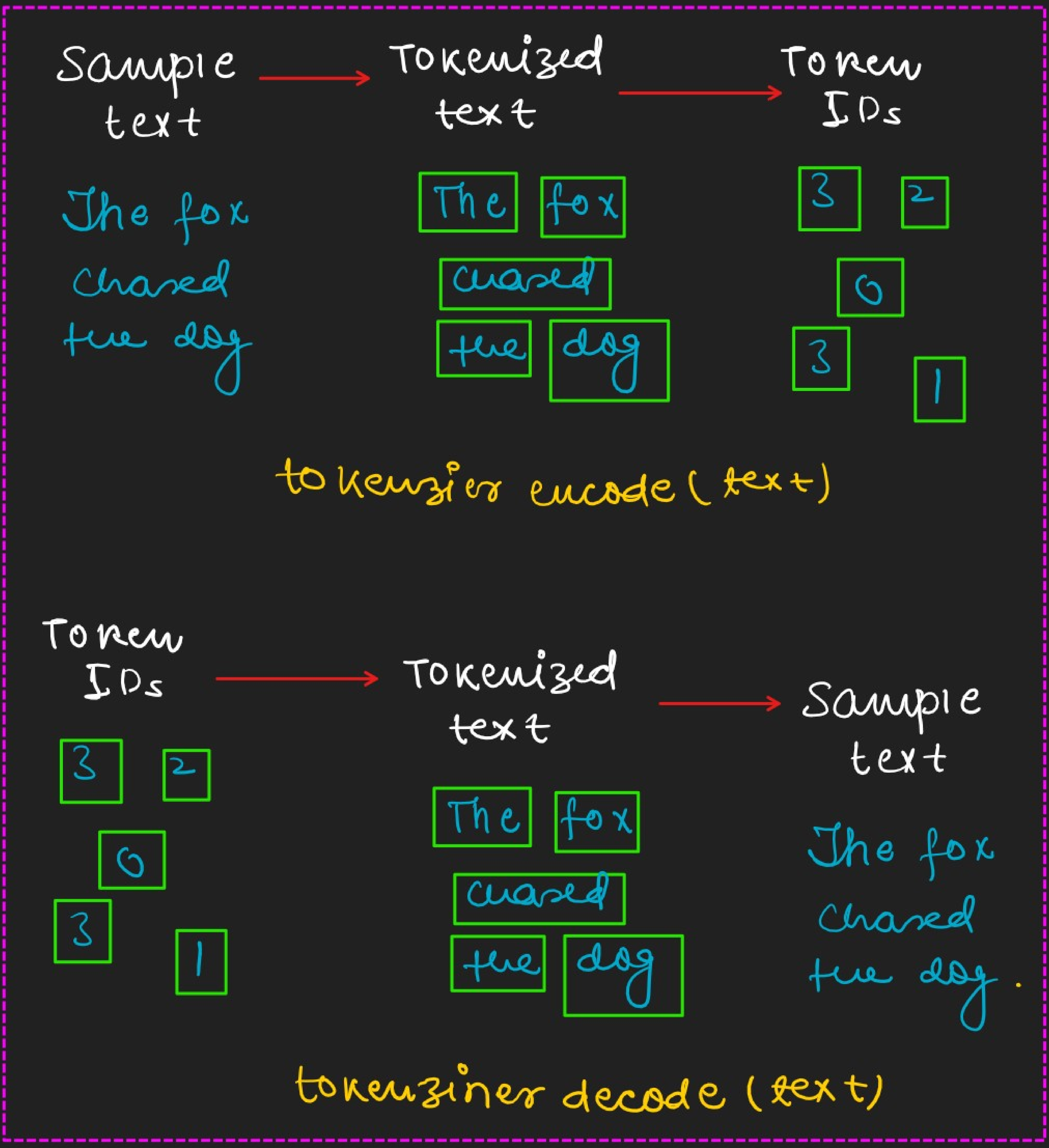
vocabulary

brown → 0
dog → 1
fox → 2
jump → 3
lazy → 4

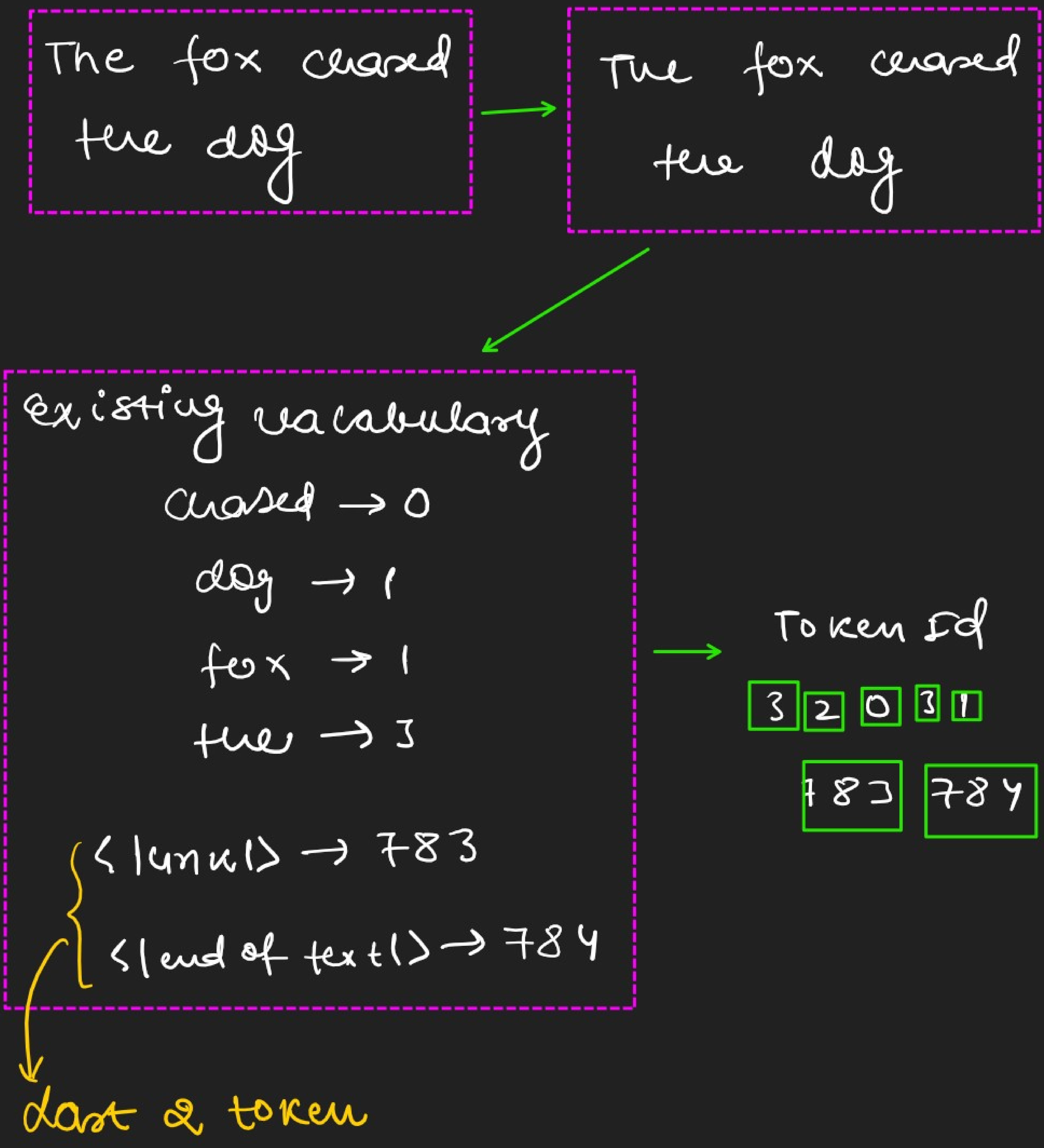
unique tokens

token Id

(e.) Implement the token class in python.



(f.) Adding special text token



more details on <|end of text|>

- * when working with multiple text sources, we add <|end of text|> token b/w these text.

- * These <|end of text|> token acts as markets, sigeingling the start of end a particular segment.