

ST7003 Procesamiento Natural del Lenguaje

Lecture03 – Tokenizers



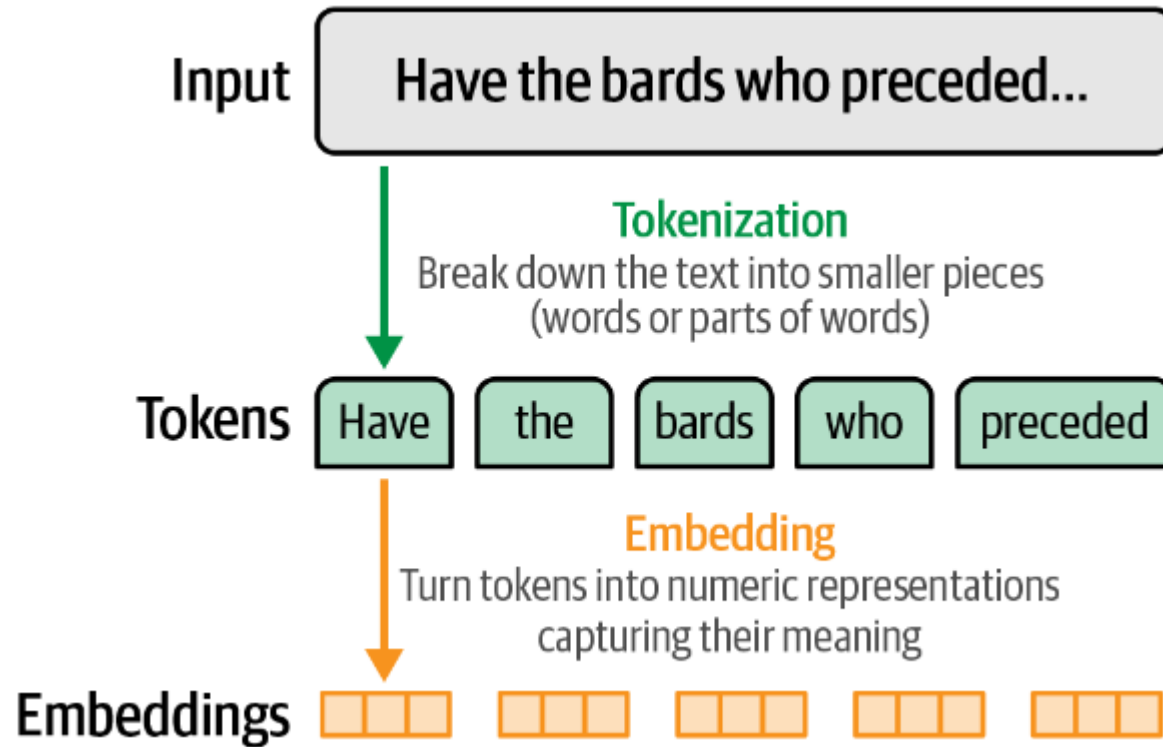
Contenido

1. Tokenizers

2. BPE

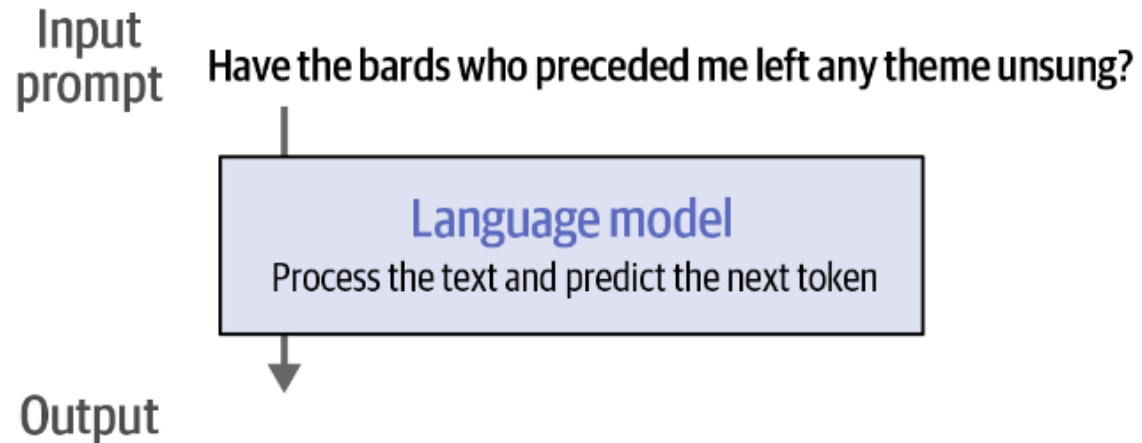


Embeddings



LLM Tokenization

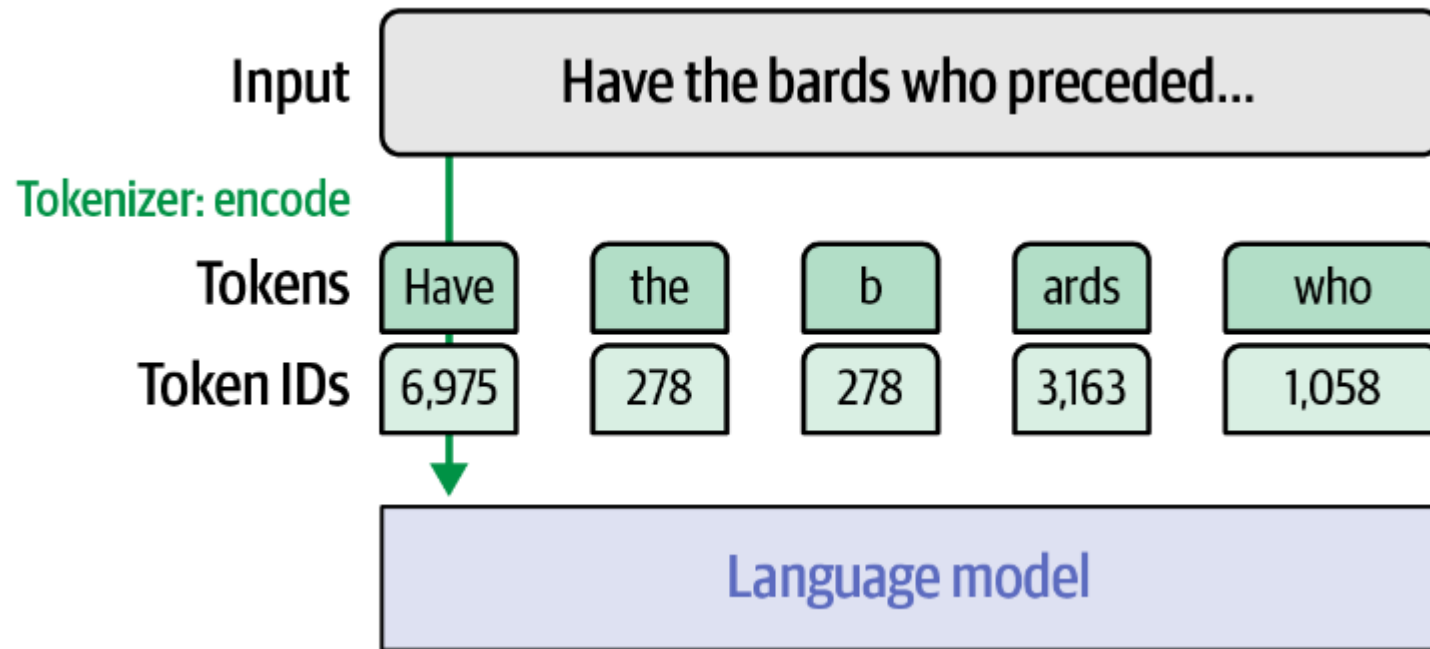
How tokenizers prepare the inputs to the Language Model



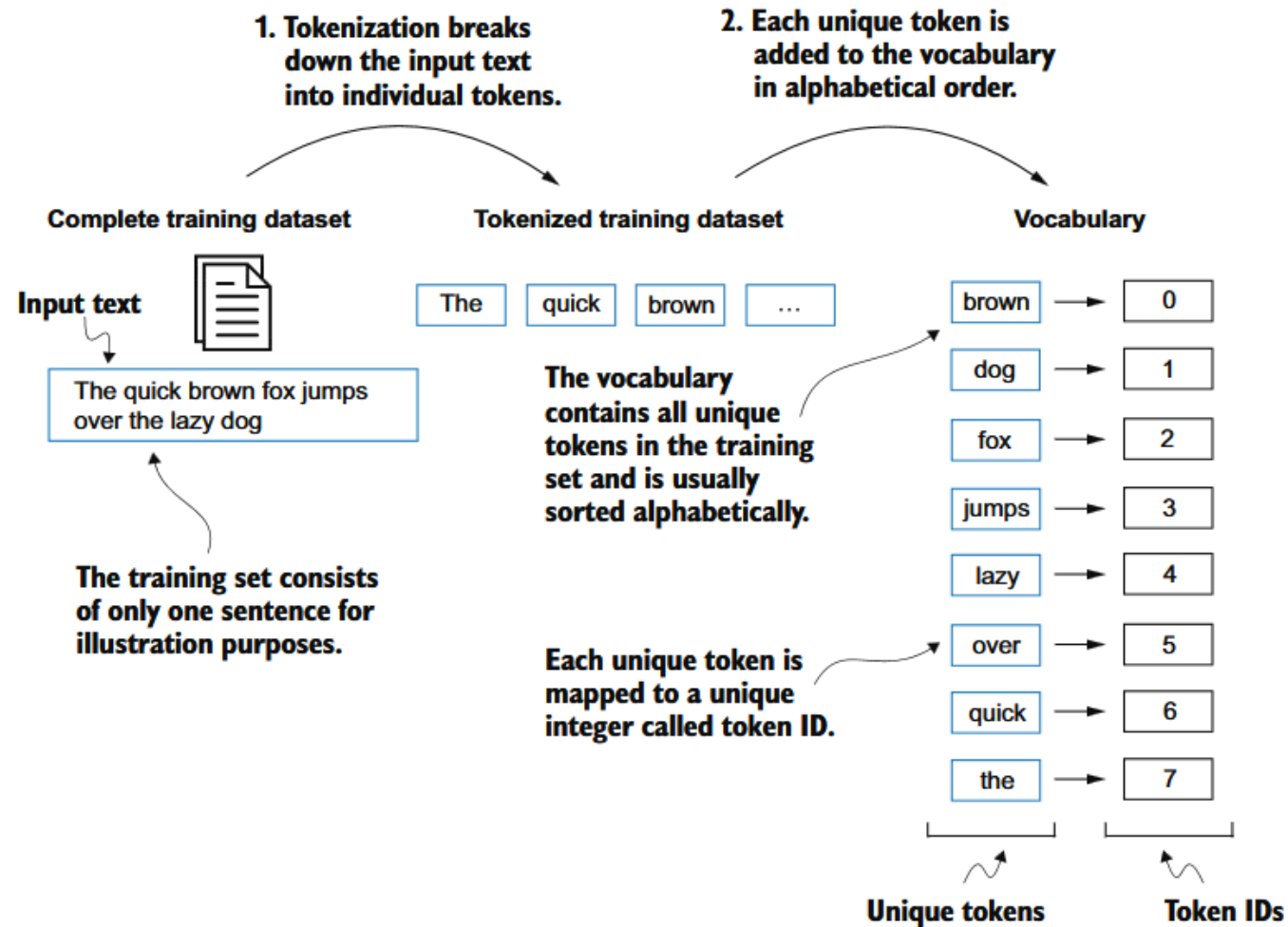
<https://platform.openai.com/tokenizer>



LLM Tokenization



LLM Tokenization



LLM Tokenization

This is how the tokenization broke down the input token:

1. Some tokens are complete words (e.g., Write, an, email).
2. Some tokens are parts of words (e.g., apolog, izing, trag, ic).
3. Punctuation characters are their own token.

Text Have the 🎵 bards who preceded...

Word tokens

Have	the	🎵	bards	who	preceded	...
------	-----	---	-------	-----	----------	-----

Subword tokens

Have	the	🎵	bard	s	who	preced	ed	...
------	-----	---	------	---	-----	--------	----	-----

Character tokens

H	a	v	e		t	h	e		🎵		b	a	r	d	s	...
---	---	---	---	--	---	---	---	--	---	--	---	---	---	---	---	-----

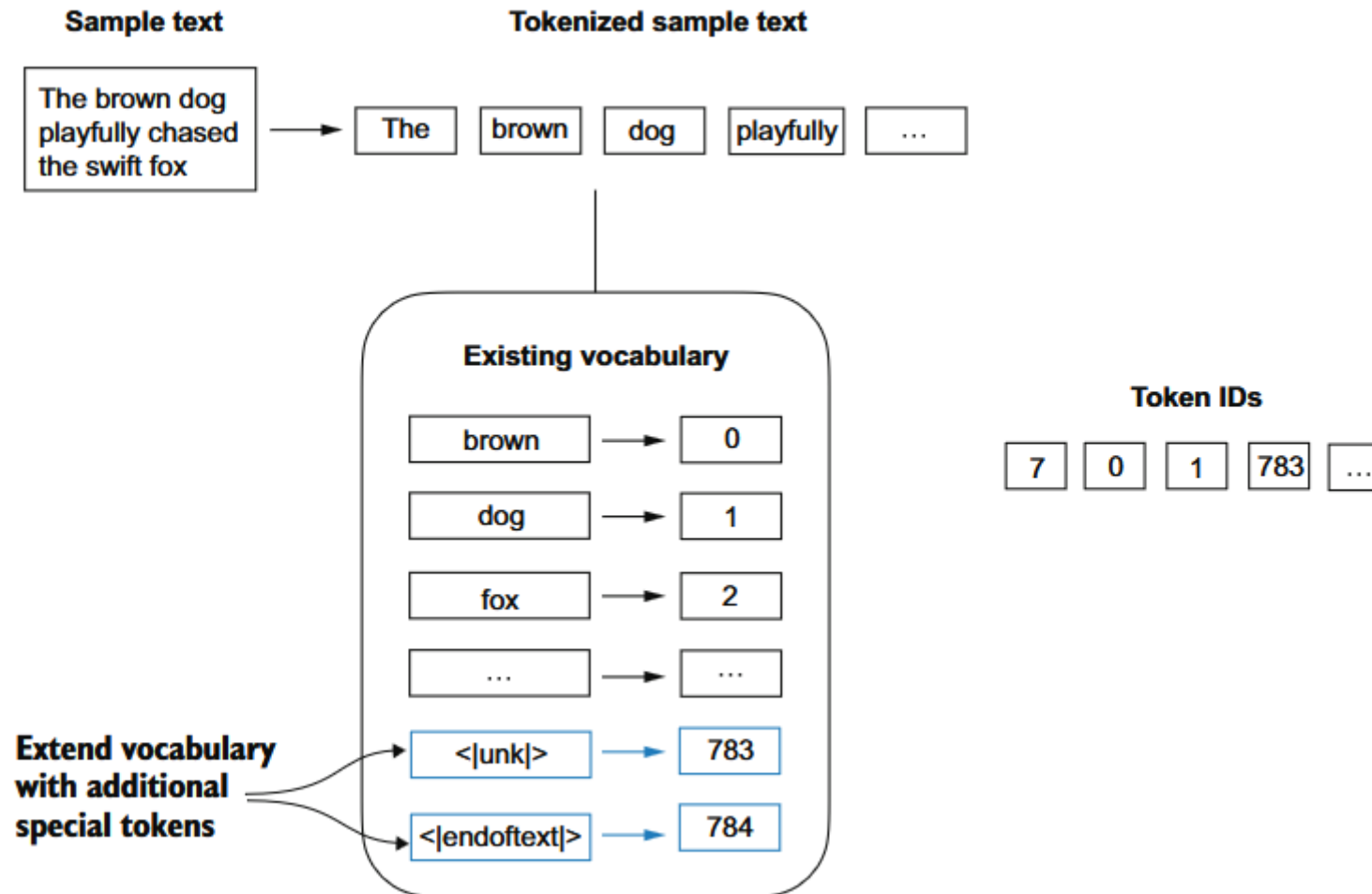
H a v e <space> t h e <space> 🎵 <space>

Byte tokens

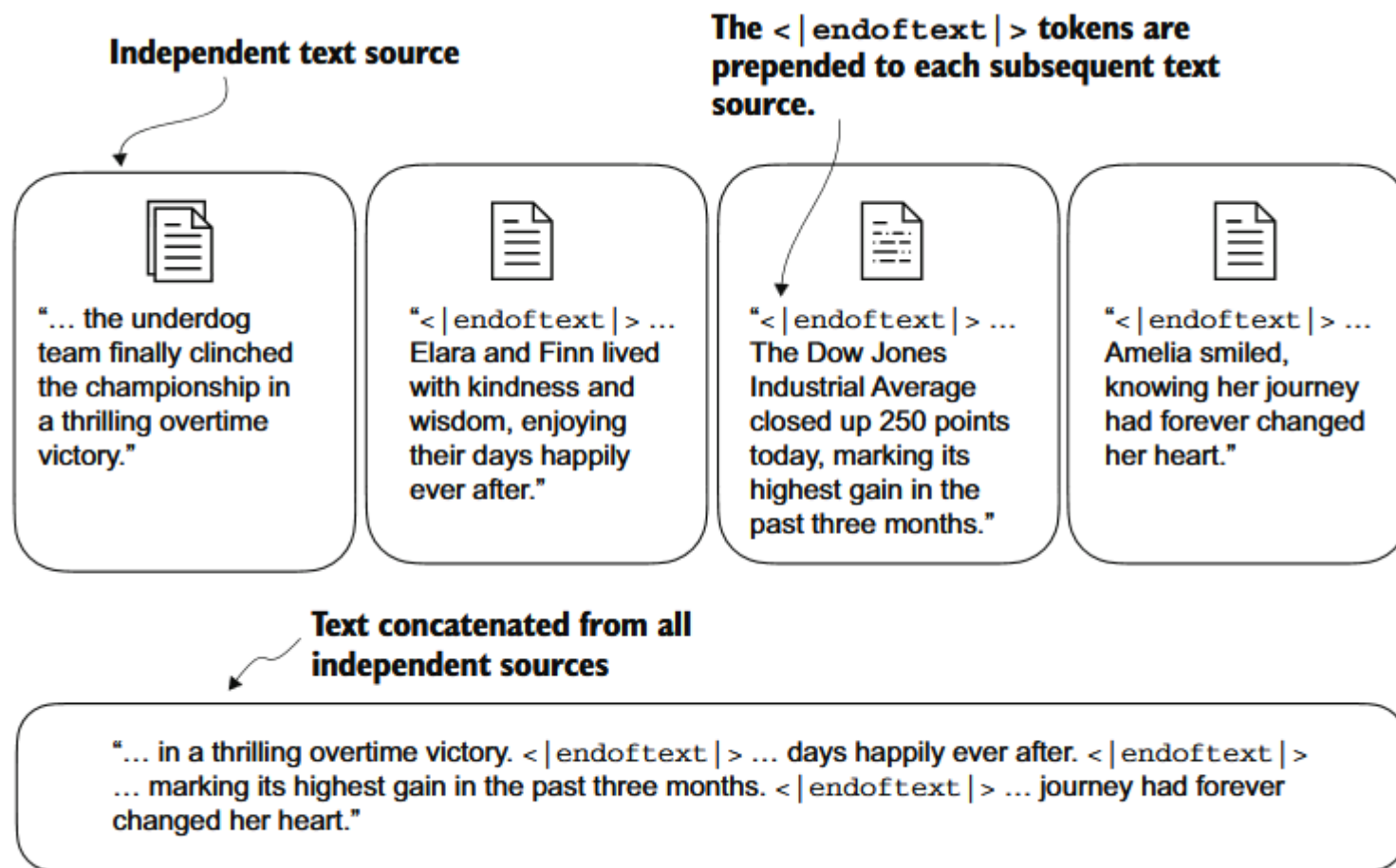
0	0	0	0	0	0	0	0	0	1	1	1	1	0	
1	1	1	1	0	1	1	1	0	1	0	0	0	0	
0	1	1	1	1	1	1	1	1	1	0	0	1	1	
0	0	1	0	0	1	0	0	0	1	1	0	1	0	
1	0	0	0	0	0	1	0	0	0	1	1	0	0	...
0	0	1	1	0	1	0	1	0	0	1	1	1	0	
0	0	1	0	0	0	0	0	0	0	1	1	0	0	
0	1	0	1	0	0	0	1	0	0	1	0	1	0	



LLM Tokenization



LLM Tokenization



Byte Pair Encoding - BPE

GPT-2

Tokenization method: Byte pair encoding (BPE), introduced in “**Neural machine translation of rare words with subword units**”.

Vocabulary size: 50,257

Special tokens: <|endoftext|>

English and CAP ITAL IZ ATION

◆◆◆◆◆◆

show █ t o k e n s F a l s e N o n e e l i f == >= e l s e : t w o t a b s : " " T h r e e t a b s : " "

12 . 0 * 50 = 600

Neural Machine Translation of Rare Words with Subword Units

Rico Sennrich and Barry Haddow and Alexandra Birch

School of Informatics, University of Edinburgh

{rico.sennrich,a.birch}@ed.ac.uk,bhaddow@inf.ed.ac.uk

GPT-4

- The GPT-4 tokenizer represents the four spaces as a single token. In fact, it has a specific token for every sequence of whitespaces up to a list of 83 whitespaces.
- The Python keyword **elif** has its own token in GPT-4. Both this and the previous point stem from the model's focus on code in addition to natural language.



Tokenizer parameters

📌 Vocabulary Size

- How many tokens the tokenizer can use.
- Common choices: **30K, 50K, 100K+ tokens** (larger vocabularies becoming more common).
- Larger vocabularies reduce sequence length but increase memory usage.

📌 Special Tokens

- Extra tokens added to handle specific cases.
- Common examples:
 - **<s>** → Beginning of text token.
 - **</s>** → End of text token.
 - **<pad>** → Padding token for sequence alignment.
 - **<unk>** → Unknown token for unseen words.
 - **<cls>** → Classification token (e.g., for BERT).
 - **<mask>** → Masking token for masked language models.
- **Custom tokens** can be added for domain-specific models (e.g., Galactica's **<work>** and **[START_REF]**).



Tokenizer parameters

Capitalization Handling

- Should the tokenizer preserve **capitalization**, or convert everything to **lowercase**?
- **Pros of keeping capitalization:** Maintains important information (e.g., "Apple" vs. "apple").
- **Cons:** Increases vocabulary size (separate tokens for "Hello" and "hello").
- **Trade-off:** Some tokenizers store only lowercase forms, using case markers instead.



Byte Pair Encoding – BPE

1. Identify frequent pairs

- In each iteration, scan the text to find the most commonly occurring pair of bytes (or characters)

2. Replace and record

- Replace that pair with a new placeholder ID (one not already in use, e.g., if we start with 0...255, the first placeholder would be 256)
- Record this mapping in a lookup table
- The size of the lookup table is a hyperparameter, also called “vocabulary size” (for GPT-2, that's 50,257)

3. Repeat until no gains

- Keep repeating steps 1 and 2, continually merging the most frequent pairs
- Stop when no further compression is possible (e.g., no pair occurs more than once)

Decompression (decoding)

- To restore the original text, reverse the process by substituting each ID with its corresponding pair, using the lookup table

Token ID	Byte Value	Character Representation
0	0x00	NULL (NUL)
1	0x01	Start of Heading (SOH)
...
32	0x20	Space ()
65	0x41	'A'
97	0x61	'a'
128	0x80	Extended ASCII
...
255	0xFF	Extended ASCII



BPE Example

- Suppose we have the text (training dataset) “the cat in the hat” from which we want to build the vocabulary for a BPE tokenizer

- **Iteration 1**

1. Identify frequent pairs

In this text, `th` appears twice (at the beginning and before the second `e`)

2. Replace and record

Replace `th` with a new token ID that is not already in use, e.g., 256

the new text is: `<256>e cat in <256>e hat`

the new vocabulary is

```
0: ...  
...  
256: "th"
```



BPE Example

- **Iteration 2**

1. Identify frequent pairs

In the text `<256>e cat in <256>e hat`, the pair `<256>e` appears twice

2. Replace and record

replace `<256>e` with a new token ID that is not already in use, for example, 257.

The new text is:

```
<257> cat in <257> hat
```

The updated vocabulary is:

```
0: ...  
...  
256: "th"  
257: "<256>e"
```



BPE Example

- **Iteration 3**

1. Identify frequent pairs

In the text `<257> cat in <257> hat`, the pair `<257>` appears twice (once at the beginning and once before “hat”).

2. Replace and record

Replace `<257>` with a new token ID that is not already in use, for example, 258.

The new text is:

```
<258>cat in <258>hat
```

The updated vocabulary is:

```
0: ...  
...  
256: "th"  
257: "<256>e"  
258: "<257> "
```



BPE Example

- To restore the original text, we reverse the process by substituting each token ID with its corresponding pair in the reverse order they were introduced
- Start with the final compressed text: <258>cat in <258>hat
- Substitute <258> → <257> : <257> cat in <257> hat
- Substitute <257> → <256>e: <256>e cat in <256>e hat
- Substitute <256> → “th”: the cat in the hat





Summary

Trained tokenizer

Tokens	
Token ID	Token
0	!
1	"
...	...
50,257	

Language model

Token embeddings	
0	
1	
...	...
50,257	