What is Tokenization?

Tokenization is the process of **breaking text into smaller parts**, usually **words or sentences**. These smaller parts are called **tokens**.

Think of it like **cutting a paragraph into words** using spaces and punctuation as scissors.

Example:
Text: "I love natural language processing."
Tokens: ["I", "love", "natural", "language", "processing", "."]

Tokenization in Space-Delimited Languages

Some languages (like **English**, **Hindi**, **Spanish**, etc.) use **spaces** to separate words. These are called **space-delimited languages**.

In these languages, tokenization is **easier** because you can split the text wherever there is a space.

Example:
Text: "I am learning NLP"
Tokens: ["I", "am", "learning", "NLP"]

But it's not always perfect!
For example:

- "U.S.A." – It should be one token, but if we split by space and punctuation, it may get wrongly split into "U", ".", "S", ".", "A", ".".

So, **tokenization tools** often use rules or machine learning to handle such special cases.

Tokenization in Unsegmented Languages

Some languages, like **Chinese**, **Japanese**, or **Thai**, do **not use spaces between words**. These are called **unsegmented languages**.

So, tokenization becomes much harder, because we **don't know where one word ends and another begins**.

Example (Chinese):

Text: "**我喜**欢自然语言处理" (means "I like natural language processing")

There are **no spaces**. So, how do we break it?

To tokenize this, we need:

- Dictionaries (to check known word combinations)
- Machine learning models (to guess correct word boundaries)

Correct Tokens: ["**我**", "**喜**欢", "**自然**", "语言", "处理"]

Sentence Segmentation

Sentence segmentation is about **splitting a paragraph into sentences**.

Usually, we look for **sentence-ending punctuation**, like ., !, or ?.

Example:
Text: "He is smart. She is kind!"
Sentences:

1. "He is smart."
2. "She is kind!"

This seems easy, but sometimes it's tricky.

Sentence Boundary Punctuation – Challenges

Just using punctuation to break sentences can cause mistakes.

Example:
Text: "Dr. Smith is here. He arrived at 5 p.m. He is a good speaker."

Here:

- "Dr." and "p.m." have **dots**, but they are **not ends of sentences**.
- So, a naive system might wrongly split "Dr. Smith is here" into two parts: "Dr." and "Smith is here."

That's why **smart sentence segmentation tools** consider:

- Capitalization (does the next word start with a capital?)
- Abbreviation lists (like "Dr.", "Mr.")
- Grammar rules or trained models

Summary

Concept	**Meaning**	**Example**
Tokenization	Breaking text into words or sentences	"Hello world" → ["Hello", "world"]
Space-delimited language	Language where words are separated by spaces	English: "I love dogs"

Concept	Meaning	Example
Unsegmented language	No spaces between words	Chinese: "我喜欢狗"
Sentence Segmentation	Breaking text into full sentences	"He left. She cried."
Sentence Boundary Punctuation	Using ., !, ? to find sentence ends (with care!)	"He works at 5 p.m." (Don’t break at "p.m.")