🔹 What is NLP?

- **Definition:** AI branch that helps computers understand & process human language.
- **Purpose:** Bridge between **human language** and **machine code**.
- **Applications:**
    - Chatbots → Customer support
    - Email filters → Spam detection
    - Voice assistants → Speech understanding
    - Search engines → Query interpretation
    - Translators → Language conversion
    - Sentiment analysis → Mood/opinion detection

Tokenization

- Break text into **smaller units (tokens)** → words, sentences, or characters.

🔹 Why Tokenize?

- Makes text machine-readable
- Essential for **sentiment analysis, translation, text generation**

🔹 Types of Tokenization

1. **Word-level:** Splits by words → "I love AI" → [I, love, AI]
2. **Character-level:** Splits by characters → "AI" → [A, I]
3. **Sentence-level:** Splits by sentences → "Hello. How are you?" → [Hello., How are you?]

👉 Most common = **Word-level tokenization**

🔹 Handling Punctuation & Spaces

- **split():** Joins punctuation, ignores spaces
- **NLTK:** Separates punctuation, handles spaces
- **spaCy:** Best handling of punctuation, contractions, multilingual text

---

🔹 Corner Cases

- **Punctuation joins with words** → fix: NLTK/spaCy
- **Contractions split weirdly (don’t → do, n't)** → fix: spaCy
- **Emojis ignored** → fix: emoji-aware tokenizers
- **Multilingual text** → fix: spaCy or Hugging Face multilingual models