



BOOM
AiGeneration

COSA SI NASCONDE DIETRO LE PAROLE?

LA TOKENIZATION SPIEGATA BENE

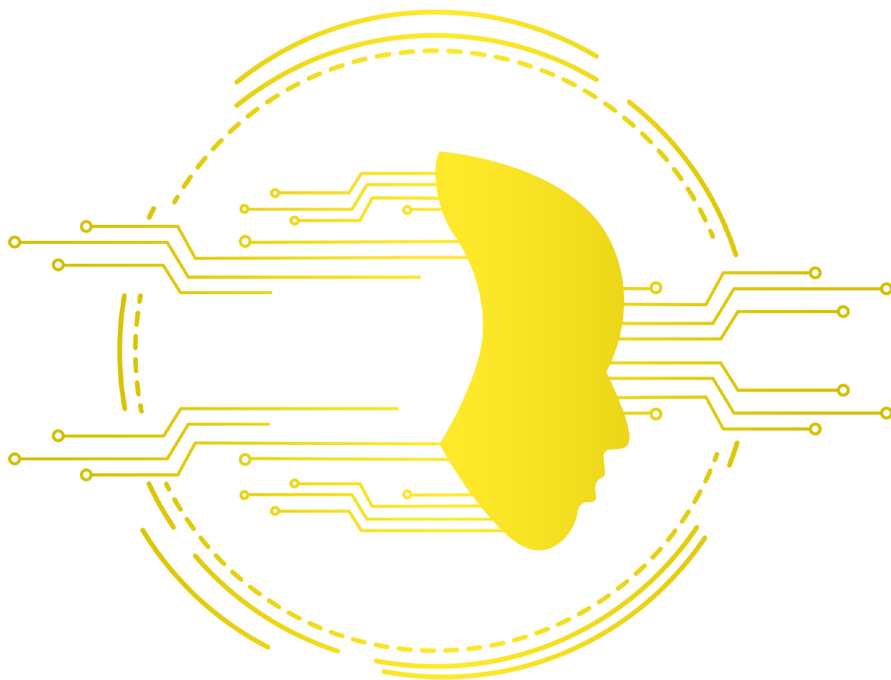
SCOPRI DI PIÙ →

Quando scriviamo un messaggio, leggiamo un articolo o parliamo a un assistente vocale, stiamo usando il linguaggio in modo naturale, fluido, continuo.

Ma le macchine, per capire il linguaggio umano, devono **spezzarlo, segmentarlo, strutturarlo**.

Ecco dove entra in gioco un concetto fondamentale per l'elaborazione automatica del linguaggio: la **tokenization**.





In poche parole, la tokenization è **il processo con cui il testo viene suddiviso in unità fondamentali dette "token"**. È il primo passo in quasi tutti i modelli di **Natural Language Processing (NLP)** e **Intelligenza Artificiale linguistica**.



lin l g guag
g linguaggio u
n gio a i

Cos'è un token?

Un **token** è una singola unità di testo su cui lavora un modello linguistico. Può essere:

- Una **parola intera** (es. "linguaggio"),
- Un **segmento di parola** (es. "lin "guag" + "gio"),
- Un **carattere singolo** (es. "l", "i", "n", ...),
- O persino **simboli speciali** come punteggiatura, emoji, o spazi.

La scelta di cosa sia un "token" dipende dal **tipo di modello** e dal **compito linguistico**.

Perché è così importante?

La tokenization non è solo una questione tecnica: è ciò che **determina come una macchina percepisce il linguaggio umano**. Se la segmentazione è sbagliata, anche l'interpretazione e la generazione del testo saranno errate.

È come dare a un traduttore umano un testo con le parole tagliate male: il risultato finale sarà confuso o addirittura incomprensibile.





Per i modelli linguistici moderni (come BERT, GPT, RoBERTa), la tokenization è una parte essenziale del **preprocessing del testo**, ed è direttamente collegata all'efficienza, alla comprensione semantica e alla capacità di generalizzazione.

Tipi di tokenizzazione

Nel corso dell'evoluzione dell'NLP, sono stati sviluppati vari paradigmi di tokenizzazione, ciascuno con implicazioni teoriche e applicative:

1. Whitespace Tokenization

Segmenta il testo in parole separate da spazi o punteggiatura. È semplice ma inefficace in lingue senza spazi (es. cinese) o in testi non strutturati.

linguaggio



2. Regex-based Tokenization

Utilizza espressioni regolari o dizionari linguistici per distinguere parole, numeri, segni diacritici e simboli. È più sofisticata ma linguaggio-dipendente.

3. Subword Tokenization

(es. Byte Pair Encoding – BPE)

Basata su algoritmi statistici di compressione, segmenta le parole in sottosequenze frequenti. È robusta rispetto a parole rare, morfologia complessa e multilinguismo. Ampiamente usata in modelli neurali (es. GPT, BERT).

4. Character-level Tokenization

Ogni singolo carattere è trattato come un token. Utile in compiti di spelling, linguaggi artificiali o quando il vocabolario è estremamente variabile.

l g
g a i u
n i o g



Token ≠ parola

È importante ricordare: **un token non corrisponde sempre a una parola.** Questo ha implicazioni dirette anche sui modelli IA che usano i token come “unità di costo”.

Ad esempio: GPT-3.5 considera un token come una media di $\frac{3}{4}$ di **parola**. Quindi, quando si parla di “**limite di 4096 token**”, non si intendono 4096 parole, ma circa **3000**.



Tokenization e multilinguismo

Uno dei grandi vantaggi dei tokenizzatori moderni è che riescono a gestire testi in **lingue diverse**, anche con alfabeti differenti (cirillico, cinese, arabo...). Questo avviene grazie all'uso di token universali o caratteri Unicode, combinati in modo flessibile.



C O N C L U S I O N E

Un'operazione invisibile ma fondamentale

La tokenization è spesso nascosta all'utente finale, ma è uno dei passaggi più critici in ogni sistema NLP. Senza di essa, i modelli non saprebbero nemmeno da dove iniziare.

Capire **come vengono suddivise le parole, i concetti e i simboli** ci aiuta a capire anche **come le macchine “leggono” il linguaggio**, e perché a volte possono fraintenderlo o usarlo in modo bizzarro.



E tu?

E tu? Avevi mai pensato che le macchine non leggono parole... ma token?
Raccontaci nei **commenti**
in che contesto ti sei
scontrato con i limiti
(o i vantaggi) della
tokenizzazione.

Seguici e unisciti a noi
per diventare parte
della Community

BOOM
AiGeneration

Collaborer.Ai