

L U M B R E



# Tokenización y normalización

**Grado:** 7

**Materia:** Procesamiento de Lenguaje Natural · Inteligencia Artificial

11/10/2026

## Antes de empezar

Antes de cualquier modelo, el texto bruto hay que trocearlo y homogeneizarlo. La tokenización elige las piezas con las que trabajará la máquina, y la normalización reduce el ruido para que variantes equivalentes no dispersen los datos.

### QUÉ APRENDERÁS

- Dividir el texto en unidades, normalizarlo y limpiarlo para poder modelarlo computacionalmente.

## Tokenización y normalización

### Trocear bien es medio camino

Un modelo no entiende frases, entiende tokens. Cómo se parte el texto decide qué verá la máquina: demasiado fino dispersa, demasiado grueso no generaliza. Es la primera decisión de todo sistema de lenguaje.

#### VIDEO

### Cómo funcionan los tokens

Vídeo en español que muestra cómo el texto se parte en tokens antes de entrar a un modelo.



Escanea para verlo en Lumbre

<https://www.youtube.com/watch?v=6bKdKTI09Y>

## Qué es un token

Un token es la unidad mínima que el modelo trata como símbolo: una palabra, un subfijo o un carácter. Tokenizar es convertir texto en una secuencia de esos símbolos identificables.

## Por palabras

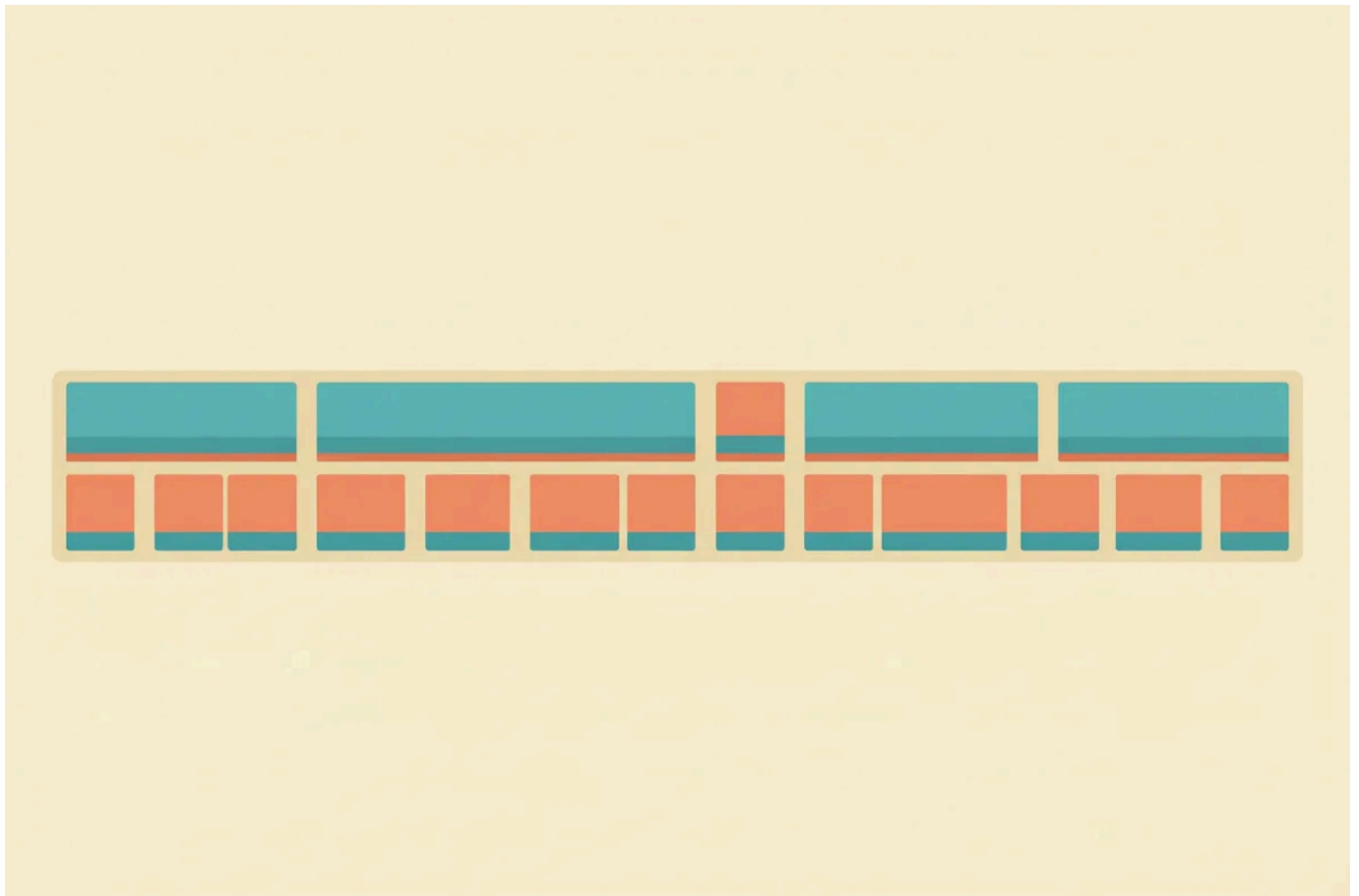
- Separar por espacios y puntuación es intuitivo y legible.
- Genera vocabularios enormes y sufre con palabras nuevas.
- La riqueza morfológica —tildes, mayúsculas, sufijos— fragmenta aún más.

## Normalizar

Pasar a minúsculas, acentuar o no, unificar espacios y eliminar ruido reduce variantes que serían tokens distintos. La normalización conviene decidirla según la tarea: en algunos idiomas perder matices cuesta exactitud.

## Lematización y stemming

Reducir palabras a su raíz agrupa flexiones: «corriendo» y «corre» comparten base. El stemming corta de forma mecánica y barata; la lematización usa diccionario y contexto y da formas válidas.



*Tokenizar convierte el texto en unidades; las subpalabras parten los términos raros en piezas frecuentes.*

## Por subpalabras

Métodos como BPE o WordPiece dividen palabras raras en piezas frecuentes vistas en el entrenamiento. Así el vocabulario se mantiene acotado y ninguna palabra es completamente desconocida.

## Tokens especiales

Se añaden símbolos para inicio, fin, relleno o incógnito. Gestionarlos bien es crucial: un modelo confunde relleno con contenido si no se enmascara.

## El coste del vocabulario

Un vocabulario grande aumenta los parámetros de la capa de embedding. Subpalabras equilibran cobertura y tamaño, sobre todo en idiomas morfológicamente ricos.

## Tokenizadores modernos

Los grandes modelos de lenguaje usan tokenizadores de subpalabras entrenados con datos masivos. Una misma frase puede partirse de forma distinta según el tokenizador, cambiando coste y comportamiento.

## Errores frecuentes

- Normalizar de más y borrar información útil —mayúsculas o negaciones— que después cambia el sentido.
- Ignorar los tokens de relleno y dejar que el modelo los trate como contenido real.
- Elegir una granularidad inadecuada: por palabra dispersa el vocabulario, por carácter alarga las secuencias.

## Ejemplo resuelto: tokenizar una frase

Quieres pasar una frase a un modelo y decides cómo trocearla:

1. Limpias la frase de ruido y espacios sobrantes.
2. Normalizas mayúsculas según convenga a la tarea.
3. Segmentas con un tokenizador de subpalabras tipo BPE.
4. Añades los tokens especiales de inicio y fin.
5. Verificas que ninguna palabra caiga como desconocida.

## ¿Para qué sirve en la realidad?

Todo sistema de lenguaje —buscadores, traductores, chatbots— empieza tokenizando; cómo partes el texto afecta al coste, al vocabulario y a si el modelo entiende o no una

palabra nueva.

**1. La tokenización por subpalabras (BPE, WordPiece) se prefiere a la por palabras sobre todo porque:**

- ☐ produce siempre menos tokens que el número de palabras
- ☐ limita el vocabulario y maneja palabras nuevas dividiéndolas en piezas
- ☐ elimina la necesidad de normalizar
- ☐ preserva mejor las mayúsculas

**2. El stemming y la lematización son exactamente equivalentes y dan siempre los mismos resultados.**

- ☐ Verdadero
- ☐ Falso

ACTIVIDAD INTERACTIVA DISPONIBLE EN LUMBRE

**3. Tokenización**



**4. Ordena las etapas típicas de preprocesar un texto:**

**Elementos:**

- quitar ruido, etiquetas y caracteres inútiles
- unificar mayúsculas, acentos y espacios
- dividir en tokens por palabra o subpalabra
- agrupar flexiones reduciendo a la raíz

|                   |       |
|-------------------|-------|
| <b>Limpiar</b>    | . . . |
| <b>Normalizar</b> | . . . |
| <b>Segmentar</b>  | . . . |
| <b>Reducir</b>    | . . . |

Análisis léxico [https://en.wikipedia.org/wiki/Lexical\\_analysis](https://en.wikipedia.org/wiki/Lexical_analysis)Tokenización y procesamiento de texto.

## 5. Une cada término de tokenización con su definición.

|    |                                                |       |
|----|------------------------------------------------|-------|
| 5. | unidad mínima que el modelo trata como símbolo | ----- |
| 5. | dividir palabras raras en piezas frecuentes    | ----- |
| 5. | reducir una palabra a su raíz con diccionario  | ----- |
| 5. | pasar todo a minúsculas para reducir variantes | ----- |

**Opciones:** token · subpalabras · lematizacion · normalizacion

## Trocear bien es medio camino

Tokenización y normalización

Normalizar

minúsculas

unificar espacios

Reducir

stemming

lematizacion

Segmentar

por palabra

por subpalabra

Especiales

inicio y fin

relleno y enmascarado

El mapa mental interactivo es interactivo en la versión online.



Tokenizer <https://en.wikipedia.org/wiki/Tokenizer> Componente que convierte el texto en tokens para el modelo.

**6. Vas a clasificar tweets en español. ¿Normalizarías las tildes y las mayúsculas? Razona qué ganas y qué podrías perder para esa tarea concreta.**

---



---



---



---

## Referencia curricular

|                 |                                                             |
|-----------------|-------------------------------------------------------------|
| <b>Grado:</b>   | 7                                                           |
| <b>Materia:</b> | Procesamiento de Lenguaje Natural · Inteligencia Artificial |

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://univ.lumbre.lat/object/cdia-nlp-001-tokenizacion-y-normalizacion>

Material educativo generado desde Lumbre · [lumbre.lat](https://lumbre.lat)