La tokenización es el proceso de dividir texto en unidades más pequeñas llamadas tokens, que pueden ser palabras, subpalabras o caracteres. Es un paso clave en NLP, ya que permite transformar texto en un formato que los modelos pueden procesar. Existen tokenizadores como WordPiece, Byte-Pair Encoding (BPE) o SentencePiece. El tipo de tokenización influye en el rendimiento del modelo, especialmente en idiomas con alta variabilidad morfológica o escritura no separada por espacios...