# Procesamiento de Lenguaje Natural

1. **Preprocesamiento/Normalización:**

* Crear uniformidad en los datos.
* Se eliminan stopwords (Palabras que no dan valor como artículos o conectores), por ejemplo “El vaso es rojo” -> “vaso rojo”.
* Se quitan acentos y es mejor apegarse al alfabeto americano.
* Debe quedar un vector de palabras.

1. **Vectorización:**

* Convierto un vector de palabras a uno de números: [palabras] -> [números]

1. **Modelado:**

* Aplica algoritmo de clasificación, sentimiento y/o predicción a los números.

**Conceptos:**

**Token:**  Unidad más chica que puedo interpretar, puede ser una palabra, letra, etc.

**Vocabulario:** Obtener los tokens únicos, para evitar que se repitan.

**Corpus:** Un libro compuesto de documentos o archivos, libro dicho de forma metafórica.

**Documento:** Texto, como una hoja de texto, etc., una hoja del corpus.