**Unidad 3 – Procesamiento del lenguaje natural**

1. **Preprocesamiento de texto**
   1. **Introducción**

El preprocesamiento de texto consiste en limpiar y prepara los datos de texto para su posterior uso en un contexto especifico. Es utilizado en todas la tareas del PLN, ya que estos datos incluyen mucho ruido y pueden varias. Su objetivo es reducir un texto a solo las palabras que realmente se necesitan.

Las técnicas más comunes son: regex, nltk, eliminación de caracteres y formato innecesario, tokenizacion y normalización

* 1. **Eliminación de ruido**

El limpiado de texto se usa en una gran cantidad de ámbitos. Dependiendo del problema es posible que se quiera eliminar información no deseada como: puntuación y acentos, caracteres especiales, formato html, etc…

El tipo de ruido que se desea eliminar normalmente de la fuente (Web-scrapping, APIs, etc…)

El método ‘.sub()’ del modulo ‘re’ permite eliminar gran parte del ruido

* 1. **Tokenización**

La tokenización consiste en dividir el texto en componentes más pequeños (token). Los tokens pueden ser palabras, frases, silabas, etc…

La librería nltk dispone de las funciones word\_tokenize(text) y sent\_tokenize(text)

* 1. **Normalización**

La normalización es un termino genérico que engloba varias tareas de preprocesamiento de texto:

* Pasar de mayúsculas a minuscular
* Eliminar palabras vacías (stopwords)
* Stemming: eliminar prefijos y sufijos
* Lematización: remplazar un token de una palabra con su raíz.

Eliminar palabras irrelevantes (o más comunes) cuando no importa la estructura de la oración. Con nltk puedes importarlas con stop\_words = set(stopwords.words(‘english’))

La derivación elimina los prefijos y sufijos de las palabras, tuilizado normalmente por los motores de búsqueda para mejorar la coincidencia con la entrada del usuario. NLTK tiene un derivador llamado PorterStemmer

La lematización es un proceso de conversión a su forma raíz. Es importante etiquetar las palabras con su parte del discurso (sustenativo, verbo, etc…)

Etiquetado de localización en el discurso (Part-of-Speech Tagging) permite mejorar el rendimiento de la lematización para saber que palabra ocupa cierto lugar en la cadena de texto. La librería spacy tiene funciones que identifican los términos (sustantivos, adjetivos, verbos, adverbios…)