21/11/2023

Pre-procesamiento de texto

No sabemos cómo viene el texto, es por ello que debemos:

* Limpiar y prepara daos de texto
* Motivación: los datos pueden tener diferentes origines y formatos
* Existen diferentes ténicas que se pueden utilizar

Eliminación de ruido

Tokenizacion:

Dividir el texto en componentes más pequeños, cada uno de estos componentes en un token

Normalizar

Derivación

Quitar prefijos y sufijos

Lematización

Pasar de verbos conjuados al infinitivo

Etiquetado morfológico (“Part-of-Speech Tagging”)

* Mejora el rendimiento de la lematización