Trabajo Práctico Nº 3

Tema: Introducción – era parte

Fecha Inicio: 16/04/2024 **Fecha de Entrega:** 30/04/2024

Actividades:

1. Empleando el corpus Brown de NLTK, detokenize el archivo ***cg73***.

1) Tokenize en oraciones.
2) Muestre las primeras 10.

2. Realice paso a paso el preprocesamiento del texto obtenido en el punto anterior, ello incluye:

1) Eliminación de ruido (etiquetas HTML, XML, emoticones)
2) Tokenización
3) Normalización (eliminación de signos de puntuación, pasar a minúsculas)
4) Eliminación de palabras vacías
5) Obtener un listado de las 50 palabras más frecuentes
6) Stemming. Obtener un listado de las 50 palabras más frecuentes
7) Lematizacion. Obtener un listado de las 50 palabras más frecuentes
8) Lematizacion indicando el PoS para los verbos.
9) Realizar una representación tabular de los primeros 30 tokens indicando la palabra normal, realizado el stemming, lematización y lematización con PoS (verbos)

3. Armar un dataframe con lo obtenido en los subitems e), f) y g). Hacer un único gráfico comparativo de las 20 palabras más frecuentes de cada subitem.