Ruth Parajó Ferrer

Constantin Rey-Coquais

Procesamiento de Datos Masivos

Tarea 1

1. Procesamiento de datos

Para esta parte de la tarea se decidió aplicar el modelo de lenguaje KeyBert, siguiendo el ejemplo de la ayudantía.

Primero elimina las menciones de usuarios (por ejemplo, aquellas precedidas por `@`) mediante expresiones regulares. Luego elimina caracteres de salto de línea (`\n`) y tabulaciones (`\t`). Después convierte todo el texto a minúsculas, para evitar diferencias por capitalización. Posteriormente, elimina acentos y caracteres diacríticos utilizando sus códigos Unicode. Luego aplica la función `eliminar_stopwords`, que remueve las palabras vacías (como 'el', 'y', 'pero'), dejando solo los términos relevantes. Finalmente, elimina los espacios en blanco al inicio y al final del texto. La línea comentada `lematizar_texto` sugiere que podría incluirse un paso de lematización, aunque actualmente no está activo.