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.

Con el análisis posterior se ha comprobado que la eficiencia de la extracción de datos no ha sido óptima, ya que una keyword concurrente ha sido 'presidente' o 'presidenta'. Esto podría darse que a la introducción de la intervención agradecen a la presidencia del congreso por concederles la palabra, además que no aporta mucho contexto o significado de la temática de la intervención.

A falta de tiempo no se ha podido implementar una mejora de este proceso de extracción pero se han contemplado otros sistemas como la utilización de modelos de clasificación temática supervisada, o la incorporación de técnicas de ponderación de términos como TF-IDF para reducir la influencia de palabras frecuentes pero poco informativas. También se consideró filtrar las palabras clave mediante un análisis de contexto más profundo, por ejemplo, utilizando embeddings semánticos para agrupar keywords similares y eliminar aquellas con menor relevancia temática. Estas posibles mejoras quedan como trabajo futuro para perfeccionar la calidad del análisis y lograr una representación más fiel de las temáticas tratadas en las intervenciones parlamentarias.