Trabajo Práctico NLP
Detección de Tópicos y clasificación

Descripción del problema

El problema a resolver es el siguiente:

- Una vez por día se reciben los textos de portales de noticias y con estos, se pide calcular los tópicos del día.
- Lo esperado es que al calcular los tópicos de un día, muchos coincidan con los del anterior y aparezcan otros tópicos nuevos. Para esto se requiere diseñar un algoritmo de merging (agrupar tópicos iguales)
- Durante el día se van recibiendo de a batches o de a un textos a clasificar en alguno de los tópicos de los días anteriores. El modelo debe retornar los ids de los tópicos a los que el texto pertenece
- Además de los tópicos, se debe encontrar las entidades, keywords y análisis sentimiento

Entrenamiento

Se tiene un dataset de entrenamiento que tiene los siguientes campos: Título, Texto, Fecha, Entidades, Keywords

Tareas:

- Dividir el dataset por día y realizar el entrenamiento de un modelo de detección de tópicos diario. Se debe usar un modelo que utilice **embeddings** para realizar el clustering
- **Merging**: Definir un criterio de agrupación de tópicos aplicado al mismo día y entre distintos días.
- Mergear tópicos similares
- Almacenar los embeddings de tópicos en una base de datos que soporte vectores (OpenSearch por ejemplo)
- El modelo de datos debe tener los siguientes campos para cada tópico: Keywords, embeddings, fecha, umbral de detección y alternativamente un nombre generado por chatGPT u otra técnica

Inferencia

Entrada:

- Título
- Texto

Salida:

- Id de los tópicos al que pertenece el texto
- Entidades y keywords matcheadas con el tópico