Procesamiento de Lenguaje Natural

Maestría en Inteligencia Artificial Aplicada
Tecnológico de Monterrey

Tarea: Audio-a-Texto con LDA y LLM

Número de Equipo: _____

Nombres: _____

Matrículas: _____

La presente actividad se realiza en equipos y utiliza como datos de entrada archivos de audio con algunas de las fábulas de Esopo en idioma español.

Deberán considerar los siguientes puntos:

Ejercicio 1:

- Los audios son del dominio público tomados de la página del proyecto Gutenberg y los encuentran en la siguiente liga:
  https://www.gutenberg.org/ebooks/21144
- Del total de fábulas que hay en dicha página (30 diferentes), solamente deberán trabajar con los siguientes 10 audios:
  - 1, 4, 5, 6, 14, 22, 24, 25, 26, 27.
- Todos los audios son en español, aunque con diferentes acentos de personas de habla hispana y uno de ellos de una persona no nativa de habla hispana (el audio 25). De preferencia usar los archivos de audio en formato MP3; pero en dado caso pueden usar el formato Apple iTunes si usas Mac y se les facilita su manejo. Cada audio tiene una duración aproximada de 1 minuto. Indiquen el tipo de formato que utilizaron.
- De cada archivo de audio deberán extraer el texto en español.
  - Todos los audios comienzan y terminan con el mismo contenido. Inician con “las fábulas de Esopo … fábula número ##”. Y terminan con “fin de la fábula, esta grabación …”. Este inicio y final de texto deberás eliminarlos para que no sean parte del análisis que realizarán.

Ejercicio 2:

- Para la extracción del audio a texto pueden usar cualquier modelo audio-to-text que crean adecuado. En el archivo **MNA_NLP_modelos_audio2txt.ipynb** que está en Canvas se encuentran algunos ejemplos, pero pueden usar algún otro que consideren extrae mejor la información. De preferencia prueben varios de ellos y seleccionen el más adecuado. Indiquen qué modelo seleccionaron y por qué tomaron dicha decisión.

Ejercicio 3:

- Pueden considerar aplicar algún proceso de limpieza de texto si lo creen adecuado. Lo estándar es eliminar caracteres especiales, stopwords, palabras con una longitud muy corta de caracteres o con muy poca frecuencia; pero en dado caso consideren que cada texto en sí es muy corto y podrían estar eliminando información valiosa. Comenten la decisión que hayan tomado, inclusive si deciden no aplicar procesamiento alguno.

Ejercicio 4:

- De cada fábula deberán extraer las palabras clave mediante el algoritmo LDA (Asignación Latente de Dirichlet). Consideren que cada fábula tiene solo un tópico y que el total de palabras por tópico serían en principio 20 (pero pueden ajustar este valor si consideran que obtienen un mejor resultado). Deberán desplegar las palabras clave obtenidas en cada fábula.

Ejercicio 5:

- Mediante el uso de un modelo LLM y con base a cada lista de palabras clave de cada fábula:
    - Generar un único enunciado que describa o resuma cada fábula.
    - Además, para cada una de dichas fábulas generar tres posibles subtemas (enunciados) diferentes.
    - Pueden seleccionar el modelo LLM de su preferencia. Nuevamente, se les recomienda utilizar varios de ellos para que comparen resultados.

Ejercicio 6:

- Incluyan sus conclusiones de dicha actividad, comentando en particular los mayores problemas que hayan enfrentado.