{desafío} latam_ Academia de talentos digitales

Actividad 6 - Inferencia de Tópicos con EM

- Para poder realizar esta actividad debes haber revisado la lectura correspondiente a la semana.
- Crea una carpeta de trabajo y guarda todos los archivos correspondientes (notebook y csv).
- Una vez terminada la actibidad, comprime la carpeta y sube el `.zip`

Descripción de Actividades

- En esta sesión trabajaremos con una serie de base de datos sobre letras musicales de distintos artistas. Cada uno de los `csv` se encuentra en la carpeta `dump`.
- Cada `csv` tiene el nombre del artista a analizar. Los archivos contienen el nombre del artista, el género musical del artista, el nombre de la canción y las letras.
- En base a esta información, el objetivo del ejercicio es generar un modelo probabilístico que pueda identificar el género musical más probable dado la letra de una canción.
- Para ello implementaremos un modelo conocido como Latent Dirichlet Allocation que hace uso de una variante del algoritmo EM para inferir clases latentes a partir de una matriz de documentos.

Desafío 1: Preparar el ambiente de trabajo

- Importe los módulos `numpy`, `pandas`, `matplotlib`, `seaborn`, `glob` y `os` siguiendo las buenas prácticas. Los últimos dos módulos permitirán realizar la importación de múltiples archivos dentro de la carpeta `dump`.
- Para ello genere un objeto que guarde en una lista todos los archivos alojados en `dump` utilizando `glob.glob` y `os.getcwd()` para extraer las rutas absolutas. Posteriormente genere un objeto `pd.DataFrame` que contenga todos los csv.
- Asegúrese de eliminar la columna `Unnamed: 0` que se genera por defecto.

Desafío 2: Matriz de ocurrencias

- Importe la clase `CountVectorizer` dentro de los módulos `feature_extraction.text` de la librería `sklearn`.
- Aplique la clase para extraer las 5000 palabras más repetidas en toda la base de datos.
- Con la clase inicializada, incorpore las letras con el método `fit_transform` y guarde los resultados en un nuevo objeto

Desafío 3: Entrenamiento del Modelo

- Importe `sklearn.decomposition.LatentDirichletAllocation` y `sklearn.model_selection.GridSearchCV`.
- Genere una búsqueda de grilla con los siguientes hiperparámetros:
  - `n_components: [5, 10, 15]`.
  - `learning_decay: [0.7, 0.5]`.
- Entrene la búsqueda de grilla con las letras en un formato vectorizado con `CountVectorizer`.
- Reporte brevemente cuál es la mejor combinación de hiperparámetros.

Desafío 4 : Inferencia e Identificación de Tópicos

- En base a la mejor combinación de hiperparámetros, entrene el modelo con la matriz de atributos de las letras.
- Para identificar de qué se trata cada tópico, necesitamos identificar las principales 15 palabras asociadas con éste. Puede implementar la siguiente línea de código para identificar las principales palabras en un tópico:
- Comente a qué tópicos está asociada cada clase inferida.

Desafío 5: Identificación de probabilidades

- En base a la información generada, es posible identificar cuales van a ser los géneros más probables de ocurrir para un artista.
- Para ello necesitamos guardar la probabilidad de cada canción en nuestra base de datos original. Podemos implementar esto de la siguiente manera:
- Genere una matriz de correlaciones entre la probabilidad de tópicos inferidos. Comente brevemente cuales son las principales asociaciones existentes.

- Con esta nueva base de datos, identifique las probabilidades de pertenencia para un artista específico.
- Grafique la distribución de las probabilidades para algún artista en específico.