Clustering-sando a José Madero

Fecha de entrega: 12 de mayo

Problema a resolver: Agrupar 31 canciones de José Madero usando el método K-means. Analizaremos las letras de dos álbumes y un EP del artista, cada una almacenada en un archivo .txt:

- **PSalmos** Con 12 canciones, este álbum aborda temas de depresión y vacío emocional.
- **Giallo**: Este álbum tiene 13 canciones y trata sobre la superación y el progreso personal
- **Aurora**: Compuesto por 6 temas, explora las dificultades ligadas a la fama.

Instrucciones:

1. **(+1 pt)** Carga y exploración datos
   - **(+0.8 pt)** Implementa una función para cargar los archivos .txt desde un directorio en específico. Asegúrate de que la función lea solo archivos de textos y capture tanto el contenido del texto como el nombre del archivo.
   - **(+0.2 pts)** Carga los datos de los álbumes en listas separadas y luego combínalos en una solta lista para su análisis posterior.
2. **(+0.5 pts)** Vectorización de datos
   - **(+0.5 pts)** Utiliza `TfidfVectorizer` para convertir los textos cargados en una matriz de características TF-IDF. Explica brevemente en tu reporte qué es TF-IDF y por qué es util para el procesamiento de textos.
3. **(+3.2 pts)** Aplicación de K-Means
   - **(+2.2 pts)** Usa una instancia de `KMeans()` de `sklearn.cluster` para clasificar textos en k clusters (tú eliges el valor de k). Justifica la elección del valor que le asignes a k y menciona detalladamente qué representa cada uno en términos de los álbumes de José Madero.

     (+1 pt) Asocia cada archivo con su cluster e imprime los resultados. ¿Cómo se agruparon los datos y qué diferencia hay con su agrupación original?
4. **(+3 pts)** Visualización de Clusters
   - **(+3 pts)** Realiza una reducción de dimensionalidad de la matriz TF-IDF usando `TruncatedSVD` y representa gráficamente los clusters resultantes. Describe detalladamente qué representan estos gráficos y cómo se pueden interpretar los resultados. Cada uno de los integrantes del equipo debe dar su análisis.

5. **(+2.3 pts)** Nuevos datos

   - **(+0.0 pts)** Carga los archivos `que_pretendes.txt` y `nuestra_afliccion.txt`. que contienen las letras de las canción *Qué pretendes* de Bad Bunny y J Balvin, y *Nuestra Aflicción* de PXNDX, respectivamente.
   - **(+0.8 pts)** Vuelve a realizar el clustering, esta vez incluyendo estos archivos. ¿En cuál álbum se agruparon? ¿A qué crees que se debe esto? Justifica ampliamente tu respuesta. Cada uno de los integrantes del equipo debe dar su análisis.
   - **(+1.5 pts)** Reflexiona sobre el inciso anterior. ¿Qué ajustes podrías realizar para que estas nuevas canciones se clasifiquen correctamente en un clúster diferente? Justifica detalladamente tu respuesta e implementa la solución propuesta. ¿Lograste los resultados esperados? ¿Por qué? Cada uno de los integrantes del equipo debe dar su análisis.

Entregables:

- Código fuente del proyecto en un archivo `.ipynb`
- Reporte en `pdf` donde se explique la implementación, la experimentación y el análisis de los resultados. Esto también puede ir en el notebook.

Estos dos archivos deben de estar dentro un de `.zip` con el nombre `equipo`-practica09.

Sobre la entrega:

- La práctica se realizará en equipos de exactamente **5 personas**.
- Ante cualquier duda con la práctica, por favor envía un correo a **taniarubi@ciencias.unam.mx** o manda un mensajito por **telegram** :D
- **En caso de no seguir los lineamientos de entrega, no se calificará la práctica en cuestión.**
- En caso de detectarse copias entre equipos, se evaluará a ambas partes con cero sin realizar indagaciones sobre el asunto.
- En caso de detectarse el uso de IA generativa en la mayor parte de la realización de la práctica, se realizará una entrevista a todo el equipo para determinar la calificación final del trabajo.