Práctica de Apache Spark

Fecha de entrega: 25 de abril

Como parte de la práctica se han entregado 3 ficheros de datos en formato TSV que contienen información relacionada con 10.000 canciones extraídas del set de datos "Million Song Dataset" (https://labrosa.ee.columbia.edu/millionsong/).

El contenido de cada uno de ellos se especifica a continuación:

Fichero albums.tsv:
- **id:** Identificador único del disco.
- **title:** Título del disco.

Fichero artists.tsv:
- **id:** Identificador único del artista.
- **name:** Nombre del artista.
- **hotness:** Nivel de popularidad del artista.
- **familiarity:** Reconocimiento del artista.
- **location:** Ubicación del artista.

Fichero songs.tsv:
- **id:** Identificador único de la canción.
- **title:** Título de la canción.
- **year:** Año de publicación de la canción.
- **hotness:** Popularidad de la canción.
- **id_artist:** Identificador único del artista de la canción.
- **id_album:** Identificador único del álbum de la canción.
- **duration:** Duración en segundos de la canción.
- **end_of_fade_in:** Segundo de la canción en el que termina el fade in.
- **start_of_fade_out:** Segundo de la canción en el que empieza el fade out.
- **tempo:** Tempo de la canción.
- **time_signature:** Número de tiempos por compás de la canción.
- **key:** Escala de la canción (de 0 a 11).
- **loudness:** Volumen de la canción.
- **mode:** Tipo de escala de la canción (mayor = 0 o menor = 1)
- **style:** Estilo de la canción.

Utilizando la información suministrada, se deberán resolver los siguientes 4 ejercicios. Se deberá entregar **un Jupyter Notebook por cada uno de ellos.**

Ejercicio 1 – Spark Core (3 puntos)

Haciendo uso de Spark Core, resuelve las 5 consultas siguientes (0,6 puntos por consulta). En principio, las consultas no siguen ningún orden específico de dificultad.

IMPORTANTE: No se podrá utilizar ninguna funcionalidad asociada a Spark SQL en este ejercicio.

1. ¿Cuál es el estilo más rápido (*tempo*) en media?
2. ¿Cuales son los 5 artistas, ubicados en UK (cualquier territorio de UK), con mayor número de canciones en escala menor (*mode = 1*)?
3. Desde 1970 hasta hoy, ¿las canciones son más rápidas (*tempo*), altas (*loudness*) y cortas (*duration*) en media? Ordena los resultados por año ascendente.
4. ¿Cuál es el estilo que más abusa de los efectos de fade in y fade out (mayor número de segundos desde inicio al final del fade in más desde el inicio del fade out al final de la canción)?
5. ¿Cual es la canción más popular (*hotness*) de los 5 artistas más populares (*hotness*)?

Ejercicio 2 – Spark SQL: funciones (1,5 puntos)

Haciendo uso de Spark SQL basado en funciones, resuelve las 5 consultas del primer ejercicio (0,3 puntos por consulta). En principio, las consultas no siguen ningún orden específico de dificultad.

IMPORTANTE: No se podrán utilizar en este ejercicio consultas en sintaxis SQL y la lectura de los ficheros tendrá que hacerse como se ha visto en clase (lectura básica más creación explícita de DataFrames).

Ejercicio 3 – Spark SQL: consultas (1,5 puntos)

Haciendo uso de Spark SQL basado en consultas, resuelve las 5 consultas del primer ejercicio (0,3 puntos por consulta). En principio, las consultas no siguen ningún orden específico de dificultad.

IMPORTANTE: La lectura de los ficheros tendrá que hacerse como se ha visto en clase (lectura básica más creación explícita de DataFrames).

Ejercicio 4 – Spark ML: regresión (4 puntos)

Haciendo uso de Spark ML, se requiere desarrollar un **modelo de regresión** que permita predecir la **popularidad (*hotness*)** que tendrá una canción en base al resto de información suministrada. Se deberán tener en cuenta los siguientes puntos:

- Se debe entregar todo el código necesario para:
    - Carga de datos.
    - Análisis exploratorio.
    - Preprocesado.
    - Entrenamiento.
    - Validación del modelo (se deja al alumno la elección de las métricas de evaluación más apropiadas para el modelo).

- No se evaluará, de forma directa, el rendimiento de los modelos en términos de precisión, pero sí se valorará el esfuerzo dedicado a desarrollar el mejor modelo posible (alternativas probadas, decisiones tomadas, creatividad, calidad del análisis previo, etc.).

La entrega deberá consistir en un fichero zip con tu nombre y apellidos (p.e. ***miguel_angel_corella_spark.zip***) que contenga cada uno de los scripts suministrados con las soluciones desarrolladas.