ACTIVIDAD SESIÓN ELEMENTOS BÁSICOS DE SPARK

Una plataforma de streaming desea analizar cuáles son las películas más vistas en Chile durante el último mes. Para esto, se dispone de un dataset con información sobre las visualizaciones de películas, incluyendo el usuario, el nombre de la película, la cantidad de minutos vistos y la calificación otorgada por el usuario.

Tu tarea es procesar estos datos utilizando **Apache Spark y RDDs**, aplicando transformaciones y acciones para obtener información relevante.

Dataset: peliculas_mas_vistas.csv

Cada fila del archivo representa una visualización de una película por un usuario e incluye la siguiente información:

usuario	película	minutos_vistos	rating	género
Juan	La Gran Aventura	120	4.5	Animación
Ana	Acción Extrema	90	3.8	Acción
Pedro	La Gran Aventura	110	4.2	Animación
Carla	Drama Profundo	150	4.9	Drama
Juan	Acción Extrema	85	4.0	Acción
Ana	Romance Inesperado	130	4.7	Romance
Luis	Documental de la Naturaleza	95	4.1	Documental
Pedro	Romance Inesperado	120	4.4	Romance
Carla	Acción Extrema	100	4.2	Acción
Luis	Drama Profundo	140	4.8	Drama

INSTRUCCIONES

1.- Carga y Preprocesamiento de Datos (1 punto)

- Cargar el dataset en un RDD y asegurarse de eliminar la primera fila (encabezado).
- Convertir los datos a una estructura de tuplas adecuadas: (usuario, película, minutos_vistos, rating, género).

2.- Cantidad de Visualizaciones por Película (1 punto)

- Contar cuántas veces ha sido vista cada película.

3.- Tiempo Total de Visualización por Película (1 punto)

- Sumar el total de minutos vistos por cada película y mostrar el top 3 de las más vistas.

4.- Películas con un Rating Promedio Mayor a 4.5 (1 punto)

- Calcular el rating promedio de cada película y filtrar las que tengan más de 4.5.

5.- Promedio de Minutos Vistos por Género (1 punto)

- Obtener el tiempo promedio de visualización por cada género.

6.- Usuarios con Mayor Tiempo de Visualización Acumulado (1 punto)

- Calcular el tiempo total visto por usuario y mostrar los 3 usuarios con más tiempo de visualización.

7.- Género Más Popular (1 punto)

- Determinar cuál es el género con más visualizaciones en total.

8.- Película con Mayor Rating en Cada Género (1 punto)

- Para cada género, obtener la película con mayor rating promedio.

9.- Distribución de Ratings (1 punto)

- Contar cuántas películas tienen un rating entre 1-2, 2-3, 3-4 y 4-5.

10.- Optimización y Explicación del Código (1 punto)

- Explicar brevemente el uso de lazy evaluation en Spark y cómo optimizar el código usando persist() o cache().

INSTRUCCIONES ADICIONALES:

- Puntos totales = 10 puntos.
- Comprimir el archivo en formato .zip o .rar.
- Subir el archivo a la plataforma.