Paralelización con hebras: pthread y mutex en Catálogo de Títulos de Netflix

Sistemas Operativos

27 de Septiembre de 2025

1. Descripción del Problema

En esta tarea de programación, se requiere implementar un programa en C que procese un archivo de datos real del catálogo de Netflix, descargado desde Kaggle: *Netflix Movies and TV Shows* (kaggle.com/datasets/shivamb/netflix-shows). El dataset original es un archivo CSV (`netflix_titles.csv`) que contiene, entre otras, las columnas `type`, `title`, `director`, `cast`, `country`, `date_added`, `release_year`, `rating`, `duration`, `listed_in`, `description`.

El objetivo principal es contar cuántos títulos pertenecen a cada país **principal**, utilizando múltiples hebras (`pthreads`) para paralelizar el procesamiento del archivo y un `mutex` para garantizar la correcta sincronización al actualizar los contadores compartidos. Para esta tarea, se define **país principal** como el primer país listado en la columna `country` (si hay múltiples países separados por comas, se toma el que aparece antes de la primera coma; si `country` está vacío o `NULL`, se ignora esa fila).

2. Requisitos

1. El programa debe aceptar dos argumentos al ejecutarse:
   a) El nombre del archivo de datos (ej. `netflix_titles.csv`).
   b) El número de hebras (N) que se utilizarán para procesar el archivo.
2. El programa debe procesar **todas las filas de datos** exceptuando la cabecera que indica las categorías. Cada línea relevante incluye la columna `country` de la cual se extrae el **país principal** según la regla descrita.

Figura 1: Ejemplo referencial de columnas en `netflix_titles.csv`. Por ejemplo, para el primer show del catálogo (segunda línea) en la columna country aparece **United States** como país principal.

3. El programa principal leerá el archivo para determinar el número total de líneas de datos (L) a procesar excluyendo la cabecera.
4. Se deben crear N hebras trabajadoras utilizando `pthread_create`.
5. El trabajo (las L líneas) debe distribuirse equitativamente entre las N hebras. Cada hebra será responsable de procesar un subconjunto de las líneas del archivo. Se debe manejar correctamente la división si L no es divisible exactamente por N.
6. Cada hebra procesará las líneas asignadas, extrayendo el **país principal** desde la columna `country` de cada registro. Si el campo `country` está vacío, esa fila no se considera.
7. Se debe utilizar una estructura de datos compartida accesible por todas las hebras para almacenar el conteo de títulos por país principal (por ejemplo, una tabla hash o un array)
8. Se debe implementar un `pthread_mutex_t` para proteger el acceso a la estructura de datos compartida. Cada vez que una hebra necesite actualizar el contador de un país, debe adquirir el `mutex`, realizar la actualización y luego liberarlo.
9. El proceso principal (la hebra inicial) debe esperar a que todas las N hebras trabajadoras terminen su ejecución utilizando `pthread_join`.
10. Una vez que todas las hebras hayan finalizado, el proceso principal debe crear un archivo de salida llamado `reporte_paises_netflix.txt` y escribir en él un resumen de los resultados: una lista de cada país principal que aparece al menos una vez y el número total de títulos asociados, **ordenado de mayor a menor**. El formato puede ser similar al siguiente:

```
País principal       | Títulos
--------------------------------
United States        | 2818
India                | 972
United Kingdom       | 419
Japan                | 245
... (resto de países) ...
```

3. Pautas Adicionales

- Asegurar una distribución eficiente y correcta de las líneas del archivo entre las hebras.
- El código debe ser claro, legible y estar bien comentado, explicando las partes clave de la lógica de paralelización y sincronización. Por ejemplo, indicar en dónde se usa el mutex y por qué motivo.
- Se sugiere normalizar espacios en blanco y `trim` del país principal (por ejemplo, eliminar espacios antes/-después de la coma).

4. Evaluación

La tarea será evaluada en base a los siguientes criterios:

- **Documentación**: Variables y funciones declaradas acorde al problema, código limpio y correctamente comentado.
- **Hebras y Sincronización**: Creación y gestión correcta de hebras (`pthread_create`, `pthread_join`). Implementación y uso adecuado del `mutex` (`pthread_mutex_t`) para proteger el recurso compartido (contador por país), evitando condiciones de carrera. Distribución correcta de la carga de trabajo entre hebras.
- **Solución**: El programa compila y se ejecuta correctamente. Divide correctamente el archivo de entrada. Agrega los conteos por país de forma precisa y maneja la concurrencia correctamente, produciendo un archivo (`reporte_paises_netflix.txt`) con los resultados correctos y ordenados.

Datos

Dataset: *Netflix Movies and TV Shows* — Kaggle (autor: Shivam Bansal).
URL de descarga: https://www.kaggle.com/datasets/shivamb/netflix-shows?resource=download

Plazo de Entrega

Fecha Límite de Entrega: 18 de Octubre de 2025 hasta las 23:59 hrs.