INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Prof. Moreno Cervantes Axel Ernesto

Práctica 4
Scraping Página Web

I N T E G R A N T E S

Salinas Monroy America Joana

Sanchez Barragan Rodrigo

Introducción

La práctica consiste en un script de Java diseñado para descargar archivos desde una página web dada, siguiendo enlaces y considerando la estructura de directorios de la página. Además, se implementa un sistema de bloqueo para la creación de directorios, evitando posibles conflictos en la creación de carpetas.

El objetivo principal es proporcionar una herramienta que permita a los usuarios descargar archivos de una página web de manera recursiva, manteniendo la estructura de directorios original de la página. El script explora los enlaces de la página proporcionada, descargando archivos y creando carpetas según sea necesario.

Desarrollo

Clase Link

Maneja una lista de enlaces (list) con mecanismos de bloqueo para la concurrencia usando ReadWriteLock. La clase tiene métodos para añadir (add) y obtener (get) enlaces de la lista de forma segura en un entorno concurrente, así como para verificar la existencia de un enlace (exists) y obtener el tamaño de la lista (size). Los bloqueos de lectura y escritura se utilizan para asegurar que las operaciones de lectura y escritura no ocurran simultáneamente, previniendo posibles inconsistencias en los datos.

Clase Archivo

Permite la descarga de archivos desde una URL en un hilo separado. La clase utiliza un objeto Link para gestionar una lista de archivos descargados y evitar duplicados. En el método run, se invoca el método download que maneja la descarga de archivos y directorios, y procesa el contenido HTML para encontrar y descargar enlaces y recursos adicionales como imágenes. Se utiliza HttpURLConnection para manejar las conexiones HTTP y BufferedReader, BufferedWriter, DataInputStream, y DataOutputStream para leer y escribir los datos de los archivos. La ruta de descarga predeterminada se define en una variable estática.

Main

Utiliza un objeto Scanner para leer la URL desde la entrada del usuario. La clase incluye una lista estática downloadFilesList de tipo Link para gestionar los enlaces de descarga. El programa crea un objeto URL a partir de la URL introducida y utiliza un grupo de hilos (ExecutorService) con un máximo de 100 hilos (MAX_T) para gestionar las descargas concurrentemente. Se inicia un nuevo hilo que ejecuta la clase Archivo para realizar la descarga de los archivos o carpetas desde la URL especificada. Si ocurre una excepción, como una URL mal formada, se imprime el rastro de la excepción.