Práctica 5: Implementación de la aplicación wget

<table>
<tr><td colspan="2">UNIDAD DE APRENDIZAJE : Aplicaciones para comunicaciones en red
UNIDAD TEMÁTICA IV: Hilos</td></tr>
<tr><td>No. Y Título de la práctica:
Práctica no. 5 Aplicación Wget</td><td rowspan="2">Tiempo de realización: 4.5 horas</td></tr>
<tr><td>Objetivo de la práctica: El estudiante implementará una aplicación que implemente la herramienta de descarga WGET para uno o más recursos a través de sus respectivos URL’s.</td></tr>
<tr><td colspan="2">Situación problemática: Se desea implementar una herramienta de descarga que permita al usuario descargar múltiples recursos de un servidor HTTP mediante peticiones de tipo GET. Además, se desea que dicho servicio de descarga sea concurrente, ya que dentro del contenido de una página web pueden existir uno o más objetos incrustados (imágenes, scripts, hipervínculos, etc.) y éstos también deberán ser descargados de manera concurrente.</td></tr>
<tr><td colspan="2">Competencia específica: Desarrolla aplicaciones en red, con base en el modelo cliente-servidor y utilizando de sockets de flujo e hilos.</td></tr>
<tr><td>Competencias genéricas:
Aplica los conocimientos en la práctica
Demuestra habilidad para trabajar en equipo
Demuestra capacidad de investigación
Desarrolla aplicaciones en red con base en la tecnología más adecuada</td><td>Elementos de competencia:
Programa aplicaciones en red con base en el modelo Cliente-Servidor y la interfaz de aplicaciones de sockets de flujo, así como hilos.
Analiza los servicios definidos en la capa de transporte
Emplea el modelo Cliente-Servidor para construir aplicaciones en red
Programa aplicaciones Cliente-Servidor utilizando sockets de flujo
Programa aplicaciones utilizando hilos de ejecución para distribuir tareas de descarga</td></tr>
<tr><td colspan="2">Criterios de evaluación: La práctica 4 aportará el 25% de la unidad temática IV</td></tr>
</table>

Introducción

GNU Wget es una herramienta libre que permite la descarga de contenidos desde servidores web de una forma simple. Su nombre deriva de World Wide Web (w), y de «obtener» (en inglés get), esto quiere decir: obtener desde la WWW. Fue escrito originalmente por Hrvoje Nikšić y por ser un proyecto de software libre tiene una gran cantidad de colaboradores directos e indirectos.

Actualmente admite descargas mediante los protocolos HTTP, HTTPS y FTP. Entre las características más destacadas que ofrece Wget está la posibilidad de fácil descarga de mirrors (repositorios) complejos de forma recursiva, conversión de enlaces para la visualización de contenidos HTML localmente, soporte para proxies, etc.

Su primera versión se lanzó en 1996, coincidiendo con el boom de popularidad de la web. Es un programa utilizado a través de línea de comandos, principalmente en sistemas tipo UNIX, especialmente en GNU/Linux. Escrito en el lenguaje de programación C.

Recursos y/o materiales

- Manual de prácticas de laboratorio de Aplicaciones para Comunicaciones en Red
- Plumones
- Bibliografía
- Internet
- Computadora
- IDE de desarrollo
- Apuntes

Instrucciones

En esta práctica debes implementar una versión básica de la herramienta Wget, la cual permitirá al usuario mediante línea de comandos indicar el URL a ser descargado, así como el uso del parámetro -r indicando que además de descargar el URL solicitado, la herramienta también descargará de manera recursiva todos los enlaces encontrados en dicho recurso.

Desarrollo de la práctica

- En esta práctica solo se implementará el lado del cliente, ya que el servidor será cualquier servidor HTTP definido en el URL de petición desde línea de comandos.
- En cuanto se inicie la aplicación cliente, se mostrará un mensaje mostrando al usuario la sintaxis de la línea de comandos
  ej. Sintaxis: wget -r -t 10 - - tries http://unapagina.io/
  donde: -r indica uso en modo recursivo
  -t define el tamaño del pool de descarga
  --tries indica el número de intentos para descargar el recurso

- Una vez establecida la conexión mediante un socket de flujo con la dirección establecida en el URL (primero extraer la dirección IP asociada al URL mediante una resolución de nombre de dominio). El programa generará una petición HTTP de tipo GET para solicitar el recurso del URL (estructurar la petición HTTP a partir del URL y enviarla al servidor).
- Después deberá validarse la respuesta devuelta por el servidor consultado y en caso de tener una respuesta afirmativa (código 200 ok) se procederá a descargar el contenido del recurso solicitado (extraer del cuerpo del encabezado de la respuesta el tamaño del archivo a ser descargado).
- Mientras se lleva a cabo la recepción del archivo solicitado y antes de que éste sea escrito en el sistema de archivos local (disco duro) se analizará el código html del archivo en búsqueda de objetos incrustados o enlaces a otros recursos, los cuales también deberán ser descargados. Adicionalmente la ruta de los enlaces deberá ser modificada de modo que refleje una ruta hacia el sistema de archivos local para encontrar el recurso referenciado, permitiendo así la navegación local del recurso HTTP sin necesidad de hacer peticiones fuera de la máquina.
- En caso de encontrarse un enlace a un nuevo recurso, este deberá ser copiado y validado para su descarga en una lista de URLS visitados de modo que URLS previamente descargados no sean nuevamente planificados para su descarga. En caso de que el URL sea nuevo, éste se agregará a la lista de URLs visitados y también se encolará para su descarga en una alberca de hilos que realizará dicha descarga. En caso de que el URL ya exista en la lista de URLS visitados, éste simplemente será ignorado.

- Para evitar que el programa pudiera ciclarse en una secuencia de descargas sin fin, el usuario podrá definir el nivel de profundidad en la búsqueda de enlaces a ser descargados.
- Cada archivo generado a partir de una descarga deberá conservar el nombre y estructura del archivo original que existe en el servidor.

Cierre de la práctica

Preguntas:

1. ¿Qué usos adicionales se le pueden dar a esta herramienta?
2. ¿Se obtiene alguna ventaja al usar una alberca de hilos para implementar la descarga de los archivos?, ¿cuál?