Bajarse una página web entera: wget

Por Paco Aldarias Raya

Impreso: 3 de diciembre de 2005

Email: pacolinux arroba inicia punto es
Web: `http://pagina.de/pacodebian`
Con Linux Debian. En Valencia (España)

El documento tiene version .html, y .pdf, cambiando en el navegador la parte final podrás acceder a ambos.

Realizado con: LaTeX

Índice

Índice **1**

1. Introducción **1**

2. Instalación **1**

3. Uso **2**

4. Ejemplo **2**

5. Descarga controlada por un fichero. **3**

6. Descargar una pagina con nc **3**

7. Comentarios **3**

8. Bibliografía **4**

1. Introducción

Podemos guardar una página web con el navegador, pero sólo se guarda el texto que hay dentro.

Existe la posibilidad de traernos todo el contenido de una página web usando wget.

2. Instalación

Desde consola como root:
apt-get install wget lynx

Siendo:

1. wget. Permite bajarse webs o ficheros.
2. lynx. Navegador web en modo texto.

3. Uso

1. **wget http://loquesea.com**

   Baja una página entera tal y como esta.

2. **wget -r -l x -A jpg,jpeg,gif,png,mpg,mpeg http://loquesea.com**

   Para bajar sólo las imagenes jpg,jpeg,gif,png,mpg,mpeg:

   Siendo donde x=nivel de recursión

3. **wget –limit-rate=1k http://loquesea.com**

   Para bajar a una velocidad. Siendo 1k=limite de velocidad.

4. **wget -rL -T 150 -np -k http://loquesea.com**

   -k para que transforme los links absolutos a relativos -np no parent. No coge los subdirectorio superiores.

5. **wget -rL -k -T 150 http://www.lapagina.com**

   Para q baje todos los archivos .jpg o .mpg de un link se puede poner q sea recursivo.

6. **wget -c -nd -r -l 5 -T 150 -k http://loquesea.com**

   Baja una página entera y metiendolo todo en la misma carpeta donde estamos:

   Siendo:

   -c indica que continue por donde se quedo la última vez.
   -nd no crea la estructura jerarquica de directorios, lo mete todo junto.
   -r recursivo. Indica que coga tb directorios.
   -l nivel de prófundidad máxima.
   -T segundo que se espera en caso de retrasos.
   -k Una vez descargada la página convierte los enlaces para verse localmente.

4. Ejemplo

Para bajar la web de IES 25 abril:
`http://intercentres.cult.gva.es/intercentres/46016713/` seria desde consola:

1. Crearmos la carpeta web:
   mkdir web

2. Nos cambiamos a esa carpeta:
   cd web

3. Nos bajamos la web principal.
   wget -c -nd -np -r -l 5 -T 150 -k `http://intercentres.cult.gva.es/intercentres/46016713/index.htm`

5. Descarga controlada por un fichero.

1. Vamos a la página web:
   `http://intercentres.cult.gva.es/intercentres/46016713/index.htm`

2. Guardamos la pagina en el fichero lista.txt

```
lynx --dump \
http://intercentres.cult.gva.es/intercentres/46016713/index.htm \
> lista.txt
```

3. Editamos el fichero para dejar los sólo los enlaces que nos interesan:
   nano lista.txt

4. Bajamos esos enlaces:
   wget -i lista.txt

6. Descargar una pagina con nc

Utilidad tcp/ip que lee y escribe.

```
nc -l -p 80 < fichero.html
```

Siendo la p es puerto. y la l listen mode para conexiones entrantes.

7. Comentarios

1. Hay servidores webs q limitan el número de páginas q se pueden bajar.

2. Wget sólo baja las páginas que tienen enlaces a otra. El resto de ficheros no los baja.

3. A partir de la versión 1.8 podemos limitar la velocidad de descarga. La versión de debian woody es la 1.8.1.

8. Bibliografía

1. Esta página:
   `http://pagina.de/pacodebian`

2. Com baixar varis arxius dúna pagina web gastant wget
   `http://bulma.net/body.phtml?nIdNoticia=716`

3. Nova versió del potent wget (Descarreges web)
   `http://bulma.net/body.phtml?nIdNoticia=1054`