Expresiones regulares - grep

Arquitectura y Sistemas Operativos
Tecnicatura Superior en Programación.
UTN-FRA

Autores: *Prof. Martín Isusi Seff*

Revisores: *Prof. Marcos Pablo Russo*

Versión: 1

1. A continuación, utilizará el comando ***wget*** para descargar el contenido del sitio web de google.

   Como se puede apreciar en la imagen, se utilizó el comando ***wget www.google.com*** y en a salida del mismo se ve que se realiza una conexión y se "salva" el archivo **index.html**. Podemos confirmar esto ejecutando el comando ***ls***. Allí vemos que efectivamente el archivo está en el directorio.

2. Revisando el contenido del archivo, vemos que el mismo no es más ni menos que código HTML[1]. En HTML, cada elemento de una página web (imágenes, links, párrafos, etc.) se representa con etiquetas de la siguiente manera:

   <etiqueta>

   Contenido del elemento.

   </etiqueta>

   Nótese que el contenido del elemento está "encerrado" por etiquetas que se abren y se cierran. Existen distintos tipos de etiquetas. Por ejemplo, los links se representan de la siguiente manera:

   <a>Título

   Dentro de las etiquetas se puede agregar más información. Esto da, por ejemplo, que tengamos etiquetas de la siguiente manera:

   Link

   Analizando la estructura de HTML, podemos decir que los links empiezan con **<a** y terminan con **/a>**, pudiendo haber cualquier carácter en el medio.

3. Habiendo completado el punto 1 y entendido el punto 2, en este punto se pide desarrollar una expresión regular que sirva para detectar links dentro del archivo descargado previamente. Una vez encontrados los links, redirigirlos hacia un nuevo archivo.

[1] Lenguaje de marcado utilizado para el maquetado de sitios web.