# Índice de contenidos

## 1.- Advertencias y recomendaciones

Se recomienda la lectura de los siguientes puntos a tener en cuenta:

- El presente código se encuentra en fase beta. Eso significa que pueden producirse errores debido a situaciones no previstas. En este caso, contacte a través del correo electrónico que consta al final del presente documento.
- Se ha limitado a 5000 las ejecuciones del programa. El código tarda un tiempo en ejecutarse que se relaciona directamente con el número de webs, las ejecuciones a realizar, la velocidad de internet de que se dispone y si se desea descargar imágenes y/o documentos. Esto significa que, a mayor número de ejecuciones, mayor tiempo de ejecución.
- No es recomendable detener la ejecución del scraping una vez ha comenzado. Si se hace, con toda probabilidad se habrá escrito de forma incorrecta en el archivo resources.php. Si fuera el caso, existe una opción (Reiniciar todos los registros) en el menú de configuración para arreglar el problema.
- Se recomienda usar el código cuando se disponga de una conexión a internet de calidad y estable para que el código funcione correctamente y con mayor rapidez.
- El código se ha testeado usando Mozilla Firefox 53.0.3 y Windows 8 y en unas webs determinadas. Se usan atributos que únicamente funcionan en HTML5.
- El código usa REGEX (expresiones regulares) para analizar las url introducidas. Dado esto, si la url a analizar contiene expresiones que no se hayan contemplado en la expresión regular utilizada, no se obtendrán urls de dicho archivo.
- El código se encuentra protegido con una licencia MIT.

## 2.- Instalación

Para la instalación, debe usarse un servidor con PHP. Durante el desarrollo del código se usó XAMPP.

Para usar el código con XAMPP, se deben colocar los archivos y carpetas en la carpeta htdocs. Para una correcta visualización del menú principal (index.html) es necesario que los iconos (contenidos en la carpeta imgs) estén en la carpeta “htdocs/imgs”.

**Importante:** dado que el script tiene un tiempo relativamente alto de ejecución, es necesario modificar la variable “max_execution_time”, en el archivo php/php.ini, a un número lo suficientemente alto como para que el script no se interrumpa.

## 3.- Uso

Al entrar al menú nos encontramos con 4 enlaces: al script de scraping, a un script para explorar los archivos descargados, a un script para configurar diferentes variables y registros usados en el código y a la presente documentación.

### 3.1.- Scraping

Al usar el script de scraping, es necesario tener en cuenta las opciones de guardar imágenes y documentos, ya que estas opciones influyen en el tiempo de ejecución del script.

- Se descargarán imágenes con extensiones .jpg, .jpeg y .png.
- Se descargarán documentos con extensiones .doc, .docx, .xls, .odt, .ods, .xml, .rtf, .pdf, .txt, .dot, .dotx, .dotm, .docm, .dic, .zip, .rar, .tar, .r.gz y .bz2.

Existe la opción de que se despliegue un informe con información sobre la ejecución del script. Esto no influye en el tiempo de ejecución del mismo.

**Importante:** puesto que no se controla el peso de los archivos a descargar, ni la naturaleza de los mismos más allá de su extensión, la responsabilidad sobre lo que se descarga y lo que ocupa en el disco duro es del usuario del script.

**Importante:** al determinar el número de ejecuciones del script, el usuario debe tener en cuenta que se trata del número de ejecuciones que se realizarán **para cada** url introducida en el apartado de configuración.

### 3.2.- Explorador

Es posible que se creen automáticamente archivos en las carpetas que el script del explorador lee, tales como thumbs.db. En este caso el script los leerá normalmente.

### 3.3.- Configuración

La opción “Reiniciar registros” borra los datos sobre qué urls, imágenes y documentos se han analizado. No se descargarán urls, imágenes y documentos que consten en estos registros.

**Importante:** es fundamental que, al introducir nuevas webs, se introduzcan correctamente nombre y url en sus respectivos campos.

**Importante:** la opción de “Reiniciar todos los registros” borra también todas las url, nombres y palabras clave introducidas.

## 4.- Licencia y otra información

El código consta de una licencia MIT, que se encuentra en el archivo correspondiente(LICENSE.txt) en la carpeta raíz.

Para cualquier error, sugerencia o comentario puede dirigirse a memerto@hotmail.com.