geturls.py

概要

第一引数に指定したURLのWebページにアクセスし、ページ内に含まれているリンクをたどって行く。
アクセスしたURLを標準出力に出力する。

第二引数は最大で何階層までのページリンクをたどるかを指定する。スクリプトに指定したURLは0階層目。そのページ内に含まれるリンクが1階層目。1階層目のページに含まれていたリンクは2階層目、と数えて何階層までのページリンクをたどるかを指定する。デフォルトは5が指定されている（5階層目までのページリンクをたどる、6階層目はたどらない）。

スクリプトの引数で指定できるパラメーターは上記の2つだけ。あと、いくつかスクリプト内でユーザーが調整できるパラメーターが存在する。詳細はソース参照。

このスクリプトはrobots.txtに対応している。デフォルトで10000個のrobots.txt設定をキャッシュし、禁止されているURLはGETせずにスキップする。

拡張子が正規表現 gif|png|bmp|jpg|jpeg|img|mov|mpg|mpeg|avi|exe|gz|zip|css|pdf|rdf にマッチするコンテンツにはアクセスしない。これはスクリプト内のカスタムパラメーター（EXC_EXTENSIONS_RESTR）で指定しているので、ユーザーが追加できる。