HTTPget by Matthai, v. 1.0

About the program

HTTPget is a Perl based program for analysing (parsing) websites.
You should put your URL's in an input text file and program tries to open the links and identifies what type of a document is on the other side. If it is HTML dokument it makes an analysis of the document.
Program also extracts all links from the document. As the result are get two tab-delimited text files - in the first there is an analysis of URL's in the second are extracted links.
Program is also able to identify if URL uses WebTracker.

This package includes

HTTPget package is available in ZIP format. ZIP file includes:

- get.pl - a main script;
- import_get.sps - script to import data to SPSS;
- httpget.pdf - program manual in PDF format;
- install.txt - program manual in TXT format;
- sample.txt - sample input file with URL's.

License

Support

Since this program is a free software I do not offer any technical support.

Latest version

The latest version of the software could be downloaded from my website:
http://www.ljudmila.org/matej/httpget

About the author

See my website at http://www.ljudmila.org/matej if you want to know more about me.

Requirements

This program is intended to be installed to computer connected to the internet. HTTPget has been designed to use Perl with the following libraries:

- LWP
- HTML::HeadParser (included in LWP)
- HTTP::Headers (included in LWP)
- HTML::LinkExtor (included in LWP)
- Time::HiRes

You can get these libraries at CPAN. If you use *ActiveState ActivePerl* under Windows you can use *PPM* to install Time::HiRes library. Simply go to commandline and type PPM. Then type install Time::HiRes and then quit.

Perl and libraries must be installed before running this program. It is always recommended to use the latest stable version Perl and libraries.
Some previous versions of LWP library has a bug. To remove a bug, open a file *\Perl\site\lib\LWP\Simple.pl* in text editor like TextPad, go to line *298* and repair this:

```
my $sock = IO::Socket::INET->new(PeerAddr => $host,
                                 PeerPort => $port,
                                 Proto    => 'tcp',
                                 Timeout  => 60) || return;
```

like that:

```
my $sock = IO::Socket::INET->new(PeerAddr => $host,
                                 PeerPort => $port,
                                 Proto    => 'tcp',
                                 Timeout  => 60) || return undef;
```

Special thanks

I want to send out thanks to Gregor Petrič who motivated me for writing this software and users of Slo-Tech forum (http://www.slo-tech.com), who helped me with many useful suggestions.

Installation

There is no special installation needed after Perl and libraries are properly installed, but you can change name of "browser" (that name will be mediated to the web server where URL is located as a browser name), written in variable *$ua->agent*.
Currently it is set to "*MatejKovacicGregaPetric/InternetResearchProject*".

How to use a program

First, you have to prepare input file. It should be in the following format:

1. http://www.first_url.org *tabulator* code_number
2. http://www.second_url.org *tabulator* code_number
3. etc...

Code number could be anything, but it should be a number. The reason why this code number exists is that I am using some special data where this code indicates something important for me. But you can set anything, for instance zero:

1. http://www.ljudmila.org/matej/httpget 0
2. http://www.ljudmila.org/matej/webtracker 0

Save this file in text format.

In command line run get.pl program with text file with input data as command line parameter. If program breaks, find out which is the last line number in the input file which has been parsed before program stopped. Then type this line number in get.pl in variable *$breaked*.

Data

Data are automatically exported in two files. First is a **.dat** file, which contains variables describing your URL.
The second is **.net** file, which contains network of links in the form:
http://your_url http://link_from_this_URL.

Importing data to SPSS

Data from .dat file you can import to SPSS using *import-get.sps* script. SPSS is a statistical package mostly used in social sciences. You can also use other statistical programs, even Excel, but importing your data to those programs is up to you.
Import-get.sps script imports data and equip them with labels. Analysis is, again, up to you. It depends of a level of your statistical knowledge.

Future features

Exporting data from .net file to Pajek format, used for network analysis is not implemented yet.
There would also be fine that program automatically builds a tree of links under top link provided.