NARROWTEQ

DARCY RIPPER

May 01, 2013

Document Version 1.0.0.0
Darcy Ripper Version 1.0.0.0

www.darcyripper.com

Table of Contents

1 Overview

1.1 About

Darcy Ripper is a powerful pure Java multi-platform web crawler (web spider) with great work load and speed capabilities. This is a standalone multi-platform Graphical User Interface application that can be used by simple users as well as programmers to download web resources on the fly.

Based on proven Java technology, the intuitive Darcy GUI is easy-to-use and provides robust functionality for creating and running simple or complex download jobs.

1.2 Features and Benefits

Darcy Ripper offers a large list of features that will enhance the efficiency of the download process as far as the processing time, network time, memory used and accuracy go.

Graphical User Interface

- Multi-platform;
- Real-time view of the download job progress;
- Pause/Resume/Stop download job anytime;
- Save and Load download job template files;
- Regular Expression Editor;
- Check for Updates support;
- Online Help and support.

General Download Features

- Multithreaded – configurable number of parallel download jobs to run at a certain period of time;
- Memory control options – user can control what happens to download jobs after they finish;
- Multiple starting points (URLs) for download job – user can specify multiple hosts on which a download job can run.

HTTP Connection Features

- HTTP/HTTPS support;
- GZip compression support;
- HTTP Proxy support;
- WWW Authentication support;
- Cookies support;
- Request customization support: referral behavior, configurable agent name;
- HTTP response code analysis and configurable behavior;
- Connection limits support: number of maximum connections per server, retries number control, bandwidth limitation, limitation depending on the HTTP response code.

Download Control Features

- Maximum search depth support;
- Maximum number of followed links support;
- Maximum time limit support;
- Downloaded file size support;

- Followed URL prefix support;
- Host Name limitation support;
- Save to Disk limitation support;
- Response behavior limitation matching response header with regular expressions;
- Response behavior limitation matching response content with regular expressions;
- Downloaded file content limitation support.

2 GUI Overview

Darcy Ripper offers an intuitive and robust interface that makes it easier to create, load and run download jobs in a transparent and secure manner.

2.1 Menu/Tool bar

File Menu

The File menu includes the following commands:

New
Creates a new Job Package and launches the Job Package Configuration dialog;
Open...
Opens an existing Job Package file and loads its configuration;
Edit
Opens the Job Package Configuration dialog for the current selected Job Package;
Save
Saves the current selected Job Package to a file;
Save As...
Saves the current select Job Package to a different file;
Save All
Saves all the opened Job Packages to files;
Close
Closes the current selected Job Package;
Close All
Closes all the opened Job Packages;
Exit
Exits the application.

Job Menu

The Job menu includes the following commands:

Start
Starts processing the download of the current selected Job Package;
Pause
Pauses the current running download process;
Stop
Stops the current running download process;
Clear
Clears the data associated with the current selected Job Package.

Utilities Menu

The Utilities menu includes the following commands:

Regular Expressions Editor
Starts the regular expressions editor dialog.

Help Menu

The Help menu includes the following commands:

Help
Launches the Darcy Ripper Help dialog;
Send Feedback

Opens the default system browser and launches the Darcy Ripper Feedback URL;
Check for Updates...
Checks if there are any Darcy Ripper updates available for download;
About
Provides a few details regarding the current Darcy Ripper application.

Tool bar

The application's tool bar contains the following commands:

New
Creates a new Job Package and launches the Job Package Configuration dialog;
Open...
Opens an existing Job Package file and loads its configuration;
Edit
Opens the Job Package Configuration dialog for the current selected Job Package;
Save
Saves the current selected Job Package to a file;
Save As...
Saves the current select Job Package to a different file;
Save All
Saves all the opened Job Packages to files;
Start
Starts processing the download of the current selected Job Package;
Pause
Pauses the current running download process;
Stop
Stops the current running download process;
Clear
Clears the data associated with the current selected Job Package.

2.2 Job Package Overview

This main window section gives the user an overview of the Job Package configuration as well as the entire Job Package download process. Thus, the following sections are available:

Configuration
Defines a summary of the Job Package Settings;
In Progress...
Displays all the Job Package download connections that are being processed at a certain moment of time. A connection that is processed may be a page that is being downloaded or a page that was already downloaded and it is being processed in order for the links to be extracted;
Opened...
Displays all the Job Package download connections that are opened at a certain moment of time. A connection that is opened refers to a page that is downloaded at that particular moment of time;
Finished
Displays all the Job Package processed pages (downloads);
All
Displays all the Job Package download connections.

2.3 Utilities

2.3.1 Reg Exp Editor

This internal tool makes it easier for the user to control the regular expression that he uses in the Job Package configuration process.

Regular expressions syntax

Examples

`.*sometext.*`
Fully matches every line containing the text “sometext”.

`com`
Partially matches every line containing the text “text”.
Matches: “http://www.darcyripper.com”
Does not match: “http://www.darcyripper.org/download.html”

`.*\.com$`
Fully matches every line ending with “.com”.
Matches: “http://www.darcyripper.com”
Does not match: “http://www.darcyripper.com/download.html”

`.*\.com$|.*\.org$`
Fully matches every line ending with “.com” or “.org”.
Matches: “http://www.darcyripper.com”
Does not match: “http://www.darcyripper.com/download.html”

Special characters:

`\`	Indicates that the next character is not special and should be interpreted literally;
`.`	Any character except newline;
`\.`	A period (and so on for *, \(, \\, etc.);
`?`	Zero or one of the preceding element;
`*`	Zero or more of the preceding element;
`+`	One or more of the preceding element;
`^`	The start of the string;
`$`	The end of the string;
`\d,\w,\s`	A digit, word character [A-Za-z0-9_], or whitespace;
`\D,\W,\S`	Anything except a digit, word character, or whitespace.

Character classes:

`[abc]`	Character a, b or c;
`[a-z]`	Characters a through z;
`[a-zA-Z]`	Characters a through z or A through Z;

[^abc]	Any character except a, b, or c;
aa|bb	Either aa or bb;

2.4 Check For Updates

Darcy Ripper offers the possibility of checking if any newer versions are available for download.

We encourage you to check for Darcy Ripper versions from time to time. This will ensure that your application version has the latest features and capabilities.

If a newer version is available, you can choose to download (install) the update.

Note: On some operating systems special write permission may be needed in order for your update to work as expected. Before downloading an update, please ensure that you have sufficient access rights to install it. A password, usually the administrator's or root password, may be required.

3 Job Package Configuration

3.1 Basic Settings

This section refers to the most basic Job Package configuration parameters. Most of these parameters are mandatory and pretty important.

The available basic settings are:

Name
Defines the Job Package name. Multiple Job Packages can have the same name but we do not recommend this approach because it will lead to confusion in organizing these Job Packages. *This is a mandatory field*;
URL(s)
One (or more) URL from which the Job Package will start its processing. The URL(s) specified here must be valid (according to the RFC #3986) otherwise Darcy will signal the invalidity with an error. Multiple URLs can be added by pressing the "Add..." button. *This is a mandatory field*;
Save Path
The absolute path of the directory where downloaded resources must be saved. In order for a Job Package download process to work as expected, enough rights must be given to Darcy in order for it to be able to write files at this location. *This is a mandatory field*.

Concurrency Settings

Parallel Downloads
The maximum number of parallel downloads that can run at a certain moment of time. This is mandatory field.

Memory Settings
These settings will help save a lot of memory as download information will not be kept in the application's memory.

Drop Ignored Links
If checked, all the links that have been ignored (rules not satisfied, limits impose etc.) will be removed from memory and they will not be present in the overall results.
Drop Finished Links
If checked, all the links that have been downloaded and processed completely will be removed from memory and they will not be present in the overall results;

3.1.1 Edit URLs List

This section offers the possibility of setting up multiple URLs from which the Job Package will start its processing.

For each of the URLs defined here all the other Job Package settings are valid, meaning that there cannot be set different rules for each of the URLs defined here. In order to achieve this multiple Job Packages must be defined.

Processing multiple Job Packages at a single moment of time is not supported at this moment, but we are working at it.

Each URL must be defined on a single line.

Note: The URLs list must not be empty.

Note: Each URL must be valid (according to the RFC #3986) otherwise Darcy will signal the invalidity with an error.

3.2 Basic Connection Settings

This section offers the possibility of customizing Darcy Ripper behavior with regard to the connections handled during a Job Package download process.

The available basic connection settings are:

Max Connections Number
Defines the maximum number or connections per server that this Job Package is allowed to create. Once this limit is reached the Job Package download process will stop. This limitation applies to a single server, thus for a single host. For an unlimited number of connection set this limit to "*-1*". The default value of this field is "*-1*";
Retries Number
Defines the number of times Darcy must try to retrieve a certain web resource, when the first try resulted in an error. If this value is reached and the server did not provide yet the resource, that certain resource will be considered unreachable. Even if this is helpful in some cases (e.g. the server is too slow) setting this value to a greater number will increase also the time and resources Darcy needs in order to process the entire Job Package. Note that this value refers to only a URL. The default value of this field is "*3*".

Proxy Settings

Address
The address of the proxy server;
Port
The port on which the proxy server is listening on;
User Name
The user name authentication detail to be used for proxy server connection;
User Password
The user password authentication detail to be used for proxy server connection.

Request Settings

Send Referral
Signals that the "*Referral*" request header must be added to the sent requests;
User Agent
Defines the "*User-Agent*" request header that must be added to the sent requests. By default, Darcy Ripper uses it's own value for this request header, but there are Web Server which require specific user agent value to work as expected. Some examples are:
- Mozilla Firefox 20

```
 Mozilla/5.0 (Windows NT 6.1; WOW64; rv:20.0) Gecko/20100101 Firefox/20.0
```

- IE 9

```
 Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.1; WOW64; Trident/6.0; SLCC2; .NET CLR 2.0.50727; .NET CLR 3.5.30729; .NET CLR 3.0.30729; Media Center PC 6.0)
```

Bandwidth Limit

Bandwidth Limit
The bandwidth that Darcy Ripper must not exceed during the Job Package download process. This limitation applies to a single download thread.

3.3 WWW Authentication Settings

This section refers to the WWW Authentication details that must be used in the Job Package download process for authentication processes to certain websites.

Note: These credentials will only be used for WWW Authentication mechanisms. The login through HTML forms will not use such details.

Note: The order in which the WWW Authentication details are defined is important and must not be ignored in order for the Job Package to be correctly processed. This means, that the defined WWW Authentication details are processed, for a particular server, in the same order they are defined in this section.

The available WWW Authentication settings are:

Host Name
The host name for which these WWW Authentication details are valid;
User Name
The user name property of the WWW Authentication details;
User Password
The user password property of the WWW Authentication details.

3.4 Cookie Settings

These settings refer to the cookie information that will be used in the Job Package download process. Cookies defined here will be used for authentication to certain web sites. In order to add cookies you must Login to that particular web site using your favorite browser, then analyze the obtained cookies and add them here.

Note: Even if most HTTP Servers do not take into account the order of cookies, there have been reports which state that some HTTP Server do validate authentications based on the cookie order.

The available cookie details are:

Name
The name of the cookie to be used;
Value
The value of the cookie to be used;
Domain
The domain on which the cookie is valid;
Path
The path on which the cookie is valid;

The cookie defined here are used for creating the value of the "*Cookie*" request header for each request. This "*Cookie*" header follows the following syntax:

```
Cookie: cookieName1=cookieValue1; cookieName2=cookieValue2
```

Even if the domain and path are not added to the header value, these fields are mandatory in order for Darcy to decide what cookies must be used for which host.

3.5 HTTP Response Settings

These settings refer to defining special behavior depending in the HTTP Status Code received as part of the HTTP Response. For example, there are HTTP Server which may respond with a “**403 Forbidden**” or “**404 Not Found**” replies when they receive too many requests in a small period of time. Thus, in case of such a case, specifying here that if such a HTTP Response code is received a retry is needed after a delay period may fix the issue.

The available behavior properties are:

Host Name
Defines the name of the Host for which the behavior is valid;
Status Code (start)
Defines the start of the codes interval for which the behavior is valid;
Status Code (end)
Defines the end of the codes interval for which the behavior is valid;
Action
Defines the action that must be took in case all the conditions are met. At this moment there are three possible such actions:
- **Ok**: The request is considered finished with success;
- **Retry**: A retry request must be issued;
- **Failed**: The request is considered finished with error.

Delay
The delay period (in milliseconds) that must pass before the retry request must be issued.

The HTTP response status codes referred above represent the basic settings filtering criteria for these settings. HTTP standard defines some rules that must be respected by the Web Server with regard to these codes. We will depict next some of these specifications and status codes:

1XX Informational
The codes of this class refer to information messages (header settings or process progress) and must not be associated with HTTP/1.0 servers;
2XX Success
The codes of this class signal to the client that the server received the request, understood it and successfully processed it. Some of the important codes of this class are “**200 OK**”, “**204 No Content**” and “**206 Partial Content**”;
3XX Redirection
The codes of this class inform the client that at least another request must be performed in order to it to receive the request. The most encountered status code of this class is “**301 Moved Permanently**”;
4XX Client Error
The codes of this class signal to the client that an error was encountered, error originated from the client side. Some of the most encountered codes of this class are “**400 Bad Request**”, “**401 Unauthorized**”, “**403 Forbidden**” and “**405 Request Timeout**”;
5XX Server Error
The codes of this class signal to the client that an error was encountered, error originated from the server side. Some of the most encountered codes of this class are “**500 Internal Server Error**”, “**502 Bad Gateway**” and “**503 Service Unavailable**”.

For more details with regard to the HTTP specifications with regard to the Response Status Codes you can further read:

http://support.google.com/webmasters/bin/answer.py?hl=en&answer=40132

3.6 Basic Rules

Specify here special rules to customize the Job Package download process. The available basic rules are:

Depth
Defines the maximum recursion depth that must be reached during a Job Package download job.
A website structure, with regard to the web pages depth, can be viewed like this:

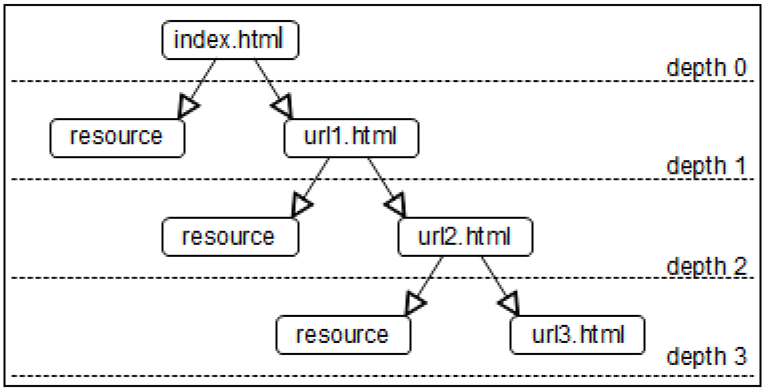

Fig. 1. Depth Sample

Links Limit
Defines the maximum number of links that must be followed during the Job Package download process. When this limit is reached then the download process will stop;
Time Limit
Defiles the maximum time (in milliseconds) that a Job Package download process must not exceed. When this time limit is reached then the download process will stop.

File Size Filter
By means of these settings, the decisions can be made depending on the web resources file size. For example, in order to avoid downloading large files, these settings may be used.

The available file size filter properties are:

File Size (from)
Defines the start value of the file size interval, from which files will be considered by this filter;
File Size (to)
Defines the end value of the file size interval, from which files will be considered by this filter;
Reply 'Content-Length' not available action
Defines the actual action that must be took for the file whose size if between **File Size (from)** and **File Size (to)**. At this moment the two possible values are:

- **Save To Disk**: saves the file to disk;
- **Reject File**: rejects that particular file and will not download it.

URL Prefix Filter

URL Prefix Filter
Defines, yet another, method of limiting the followed links. Thus, only URLs which begin with this value will be followed.

Host Name Filter

This filter enables the limitation of followed links by the Host Name, using regular expressions. Thus adding regular expressions in the Host Name Filter list will enable only the links whose Host Name match that regular expression to be processed.

Note: All the defined host name rules will be verified. Thus, having a regular expression “.*” will make all the others obsolete and useless.

Note: Removing all entries will lead to no link being processed.

Save To Disk Filter

This filter enables the control of the files that will be downloaded and saved to disk. Thus, a Job Package can process links without saving their content to disk.

For instance, this filter can be used when only CSS files are wanted to be saved to disk. By setting a regular expression “*.css*” will make this possible.

Note: All the defined regular expressions will be processed. Thus, having a regular expression “.*” will enable the save to disk of all files, ignoring all the other rules.

Note: Removing all entries will lead to no resource being saved to disk.

3.7 Request Filters

These settings offer the possibility of HTTP requests to be filtered by regular expressions and assigning them special actions.

The available properties for such a filter are:

Name
Defines the name of this filter. We strongly recommend to use significant names in order to have a good vision of all the filters;
Description
Defines a small description of this filter;
Match
The regular expression that will be matched against the request;
Match Success
The action that must be taken in case the regular expression is matched;
Match Failure
The action that must be taken in case the regular expression is matched.

There can be defined three possible actions for a filter:

No Change
No action will be taken. This represents the default action;
Accept
The request is considered valid and will be processed;
Reject
The request is considered invalid and will not be processed.

3.8 Reply Content Filters

These settings offer the possibility of HTTP responses content to be filtered by regular expressions.

The available properties for such a filter are:

Name
Defines the name of this filter. We strongly recommend to use significant names in order to have a good vision of all the filters;
Description
Defines a small description of this filter;
Match
The regular expression that will be matched against the response content;
Match Success
The action that must be taken in case the regular expression is matched;
Match Failure
The action that must be taken in case the regular expression is matched.

There can be defined two possible actions for a filter:

No Action
The request is considered valid and will be processed;
Abort Download
The request is considered invalid and will not be processed.