Help.WebExtract

NAME

`Web Extract` is used to scrap(extract) the web contents.

SYNOPSYS

Prerequisite

First of all, save the web contents as an HTML file.

How to save as an HTML file

1. Oepn your browser and go to the target URL
2. Copy the source of the target page to clipboard: CTRL+U (View page source), CTRL+A (Select All) CTRL+C (Copy)
3. Open your favorite editor and paste the clipboard: CTRL+V (Paste)
4. Save as an HTML (`UTF8` file encoding is recommended)

Second, you have to build your own specification as `YAML`. You can refer to this document details.

Parameter

HTML File

- A full path & filename of the source HTML file to be extraced.

Specification File

- A specification file as YAML format, which is used to defined a thing what a bot will extract from `HTML File`.
- See 'How to use Web Extract plugin' details.

Options

Set Num of Res...

- Set number of results.
- If set, results will be returned by the specified number.
- Defaults: 0 (No limit)

File encoding

- File encoding type of 'HTML File'.
- Default: UTF8
- When you saving a web source to an HTML, UTF8 is preferred.

- If failed, please, check the file encoding of 'HTML File' first.

HTML Parser

- The type of 'HTML parser' is to be used to parse the HTML file.
- Default: LXML
- If using LXML fails, please try again using 'html.parser'.

PLATFORM

Here is the supported platform for this plugin.

- This plugin support Windows 10 and above.
- This plugin support Linux (Ubuntu).
- This plugin support Mac.

Version

1.522.1458

Limitation

SEE ALSO

LICENSE