EGENIX.COM

mxTidy

HTML Tidy for Python

Version 3.0

Printed in Germany.

Contents

1. Introduction

mxTidy provides a Python interface to a thread-safe, library version of the HTML Tidy command line tool.

HTML Tidy helps you to cleanup coding errors in HTML and XML files and produce well-formed HTML, XHTML or XML as output. This allows you to preprocess web-page for inclusion in XML repositories, prepare broken XML files for validation and also makes it possible to write converters from well-known word processing applications such as MS Word to other structured data representations by using XML as intermediate format.

1.1 Thread-safe Tidy Library

During the development of this interface, the original HTML Tidy command line version was significantly modified to turn it from a single run, command line tool into a thread-safe C library which not only interfaces to files, but also to memory buffers.

This approach was later picked up by a team of volunteers to create the HTML Tidy Lib Project. mxTidy (currently) does not use the new Tidy library, but continues to include and work with the modified and enhanced original code base written by Dave Raggett.

Most of mxTidy's operations are automatic or can be manipulated by large number of configuration options. It also provides you with access to the error and warning information generated by HTML Tidy.

1.2 Speed and Memory

HTML Tidy is very good at trying to restructure the HTML or XML input, but unfortunately not too fast at it. The main reason for this is the single character input/output strategy used in the code which causes quite a few C function calls.

Changing the code to use a buffer and pointer strategy would enhance the performance, but requires a lot of work.

The memory requirements in string to string mode amount to about twice the size of the input string in addition to the parser tree overhead. In file to file mode, only the tree overhead is introduced.

Note that the current releases reconfigure HTML Tidy for every run which causes additional overhead.

2. HTML Tidy Options

The following is a list of HTML Tidy options that you can pass to the underlying HTML Tidy engine.

Most of the original HTML Tidy options are also available in the mxTidy interface. Some options have been removed, since they don't map well to an embedded module, e.g. there is no configuration file support and the slide bursting options have also been removed.

The default values used in mxTidy are given for each option.

> Note that some options have different defaults than in the command line version of HTML Tidy.

For more information about the background and workings of HTML Tidy, please see the *HTML Tidy Overview* page which is also included in the package.

`add_xml_decl = 0`

If set to 1, Tidy will add the XML declatation when outputting XML or XHTML. The default is 0.

Note that if the input document includes an <?xml?> declaration then it will appear in the output independent of the value of this option.

`add_xml_space = 0`

If set to 1, this causes Tidy to add xml:space="preserve" to elements such as pre, style and script when generating XML.

This is needed if the whitespace in such elements is to be parsed appropriately without having access to the DTD. The default is 0.

`assume_xml_procins = 0`

If set to 1, this changes the parsing of processing instructions to require ?> as the terminator rather than >.

The default is 0. This option is automatically set if the input is in XML.

`break_before_br = 0`

If set to 1, Tidy will output a line break before each <br> element. The default is 0.

clean = 0

If set to 1, causes Tidy to strip out surplus presentational tags and attributes replacing them by style rules and structural markup as appropriate. It works well on the html saved from Microsoft Office'97. The default is 0.

`drop_empty_paras = 1`

If set to 1, empty paragraphs will be discarded. If set to no, empty paragraphs are replaced by a pair of br elements as HTML4 precludes empty paragraphs. The default is 1.

`drop_font_tags = 0`

If set to 1 together with the clean option (see above), Tidy will discard font and center tags rather than creating the corresponding style rules. The default is 0.

`enclose_block_text = 0`

If set to 1, this causes Tidy to insert a p element to enclose any text it finds in any element that allows mixed content for HTML transitional but not HTML strict. The default is 0.

`fix_backslash = 1`

If set to 1, this causes backslash characters "\" in URLs to be replaced by forward slashes "/". The default is 1.

`fix_bad_comments = 1`

If set to 1, this causes Tidy to replace unexpected hyphens with '=' characters when it comes across adjacent hyphens. The default is 1. This option is provided for users of Cold Fusion which uses the comment syntax: <!--- --->

`gnu_emacs = 0`

If set to 1, Tidy changes the format for reporting errors and warnings to a format that is more easily parsed by GNU Emacs. The default is 0.

`hide_endtags = 0`

If set to 1, optional end-tags will be omitted when generating the pretty printed markup. This option is ignored if you are outputting to XML. The default is 0.

`indent_attributes = 0`

If set to 1, each attribute will begin on a new line. The default is 0.

`input_xml = 0`

If set to 1, Tidy will use the XML parser rather than the error correcting HTML parser. The default is 0.

`literal_attributes = 0`

If set to 1, this ensures that whitespace characters within attribute values are passed through unchanged. The default is 0.

`logical_emphasis = 0`

If set to 1, causes Tidy to replace any occurrence of i by em and any occurrence of b by strong. In both cases, the attributes are preserved unchanged. The default is 0. This option can now be set independently of the clean and drop-font-tags options.

`numeric_entities = 0`

Causes entities other than the basic XML 1.0 named entities to be written in the numeric rather than the named entity form. The default is 0.

`output_error = 1`

Generate error information. Default if 1.

`output_markup = 1`

Generate markup output. Turning this off is useful for checking for errors only. Default if 1.

`output_xhtml = 0`

If set to 1, Tidy will generate the pretty printed output writing it as extensible HTML. The default is 0. This option causes Tidy to set the doctype and default namespace as appropriate to XHTML. If a doctype or namespace is given they will checked for consistency with the content of the document. In the case of an inconsistency, the corrected values will appear in the output. For XHTML, entities can be written as named or numeric entities according to the value of the "numeric-entities" property. The tags and attributes will be output in the case used in the input document, regardless of other options.

`output_xml = 0`

If set to 1, Tidy will use generate the pretty printed output writing it as well-formed XML. Any entities not defined in XML 1.0 will be written as numeric entities to allow them to be parsed by an XML parser. The tags and attributes will be in the case used in the input document, regardless of other options. The default is 0.

`quiet = 0`

If set to 1, Tidy won't output the welcome message or the summary of the numbers of errors and warnings to the error stream. The default is 0.

`quote_ampersand = 1`

If set to 1, this causes unadorned & characters to be written out as &amp;. The default is 1.

`quote_marks = 0`

If set to 1, this causes " characters to be written out as &quot; as is preferred by some editing environments. The apostrophe character ' is written out as &#39; since many web browsers don't yet support &apos;. The default is 0.

`quote_nbsp = 1`

If set to 1, this causes non-breaking space characters to be written out as entities, rather than as the Unicode character value 160 (decimal). The default is 1.

`raw = 0`

Avoid mapping values > 127 to entities. Default is 0.

`show_warnings = 0`

If set to 0, warnings are suppressed. This can be useful when a few errors are hidden in a flurry of warnings. The default is 1.

`tidy_mark = 0`

f set to 1 (the default) Tidy will add a meta element to the document head to indicate that the document has been tidied. To suppress this, set tidy-mark to 0. Tidy won't add a meta element if one is already present.

`uppercase_attributes = 0`

If set to 1 attribute names are output in upper case. The default is 0 resulting in lowercase, except for XML where the original case is preserved.

`uppercase_tags = 0`

Causes tag names to be output in upper case. The default is 0 resulting in lowercase, except for XML input where the original case is preserved.

`word_2000 = 0`

If set to 1, Tidy will go to great pains to strip out all the surplus stuff Microsoft Word 2000 inserts when you save Word documents as "Web pages". The default is 0.

Microsoft has developed its own optional filter for exporting to HTML, and the 2.0 version is much improved. You can download the filter free from the Microsoft Office Update site.

`wrap_asp = 1`

If set to 0, this prevents lines from being wrapped within ASP pseudo elements, which look like: <% ... %>. The default is 1.

`wrap_attributes = 0`

If set to 1, attribute values may be wrapped across lines for easier editing. The default is 0. This option can be set independently of wrap-scriptlets.

`wrap_jste = 1`

If set to 0, this prevents lines from being wrapped within JSTE pseudo elements, which look like: <# ... #>. The default is 1.

`wrap_php = 1`

If set to 0, this prevents lines from being wrapped within PHP pseudo elements. The default is 1.

`wrap_script_literals = 0`

If set to 1, this allows lines to be wrapped within string literals that appear in script attributes. The default is 0.

`wrap_sections = 1`

Wrap within <![...]> section tags. Default is 1.

`indent_spaces = 2`

Sets the number of spaces to indent content when indentation is enabled. The default is 2 spaces.

`tab_size = 8`

Sets the number of columns between successive tab stops. The default is 8. It is used to map tabs to spaces when reading files. Tidy never outputs files with tabs.

`wrap = 72`

Sets the right margin for line wrapping. Tidy tries to wrap lines so that they do not exceed this length. The default is 72. Set wrap to 0 if you want to disable line wrapping.

`alt_text = None`

This allows you to set the default alt text for img attributes. This feature is dangerous as it suppresses further accessibility warnings.

`indent = "no"`

If set to "yes", Tidy will indent block-level tags. The default is "no". If set to "auto" Tidy will decide whether or not to indent the content of tags such as title, h1-h6, li, td, th, or p depending on whether or not the content includes a block-level element. You are advised to avoid setting indent to yes as this can expose layout bugs in some browsers.

```
char_encoding = "ascii"
```

Determines how Tidy interprets character streams.

For "ascii", Tidy will accept Latin-1 character values, but will use entities for all characters whose value > 127.

For "raw", Tidy will output values above 127 without translating them into entities. For "latin1" characters above 255 will be written as entities.

For "utf8", Tidy assumes that both input and output is encoded as UTF-8.

You can use "iso2022" for files encoded using the ISO2022 family of encodings e.g. ISO 2022-JP.

The default is "ascii".

These descriptions were extracted from the *HTML Tidy documentation* and fall under the HTML Tidy copyright.

3. mx.Tidy Functions

The package defines these functions:

`tidy(input, output=None, errors=None, **options)`

Beautify the HTML/XML input and return a tuple `(nerrors, nwarnings, outputdata, errordata)`.

`input` may be a string or a file open for reading data.

If `output` is given as file open for writing, the generated markup is written to this file and `outputdata` is set to None. Otherwise, output is written to a string which is returned by the function in `outputdata`.

The same is true for error information which Tidy generates. This is either written to `errors` or returned via errordata.

`nerrors` and `nwarnings` are integers which are set to the number of errors/warnings which TIDY generated.

Tidy options can be passed to the function using *keyword parameters*, e.g. `output_xhtml=1`.

Configuration files are not supported by the interface.

4. mx.Tidy Constants

The package defines these constants:

`Error`

This exception will be raised for problems related to the Tidy interface.

5. Examples of Use

This script demonstrates how to use mxTidy to clean up MS Word HTML exports:

```
import sys
from mx import Tidy

### Globals

_debug = 1

tidy_options = {
    'break_before_br': 1,
    'clean': 1,
    'output_xhtml': 1,
    'word_2000': 1,
    'indent': 'yes',
    'wrap': 0,
    }

###

def run_tidy(input, output=None, errors=None,
             tidy_options=tidy_options):

    (nerrors, nwarnings, outputdata, errordata) = \
              Tidy.tidy(input, output, errors, **tidy_options)
    if _debug and errordata:
        print 'Tidy messages:'
        print
        print errordata
        print
    return outputdata

def process(inputfile, outputfile):

    input = open(inputfile, 'rb')
    output = open(outputfile, 'wb')
    outputdata = run_tidy(input)
    input.close()
    output.write(outputdata)
    output.close()

###

if __name__ == '__main__':
    process(sys.argv[1], sys.argv[2])
```

More examples will appear in the `Examples` subdirectory of the package.

6. Package Structure

```
[Tidy]
       Doc/
       [Examples]
       [mxTidy]
              libtidy/
              test.py
       Tidy.py
```

Names with trailing / are plain directories, ones with []-brackets are Python packages, ones with ".py" extension are Python submodules.

The package imports all symbols from the extension module and also registers the types so that they become compatible to the pickle and copy mechanisms in Python.

7. Support

eGenix.com is providing commercial support for this package. If you are interested in receiving information about this service please see the *eGenix.com Support Conditions*.

8. Copyright & License

The mxTidy software and the modifications to the HTML Tidy source code are covered by the ***eGenix.com Public License Agreement***, which is included in the following section. The text of the license is also included as file "LICENSE" in the package's main directory.

The included HTML Tidy software is covered by the following W3C license:

By downloading, copying, installing or otherwise using the software, you agree to be bound by the terms and conditions of the following *eGenix.com Public License Agreement* and the above HTML Tidy license.

EGENIX.COM PUBLIC LICENSE AGREEMENT

Version 1.1.0

This license agreement is based on the Python CNRI License Agreement, a widely accepted open-source license.

1. Introduction

This "License Agreement" is between eGenix.com Software, Skills and Services GmbH ("eGenix.com"), having an office at Pastor-Loeh-Str. 48, D-40764 Langenfeld, Germany, and the Individual or Organization ("Licensee") accessing and otherwise using this software in source or binary form and its associated documentation ("the Software").

2. License

Subject to the terms and conditions of this eGenix.com Public License Agreement, eGenix.com hereby grants Licensee a non-exclusive, royalty-free, world-wide license to reproduce, analyze, test, perform and/or display publicly, prepare derivative works, distribute, and otherwise use the Software alone or in any derivative version, provided, however, that the eGenix.com Public License Agreement is retained in the Software, or in any derivative version of the Software prepared by Licensee.

3. NO WARRANTY

eGenix.com is making the Software available to Licensee on an "AS IS" basis. SUBJECT TO ANY STATUTORY WARRANTIES WHICH CAN NOT BE EXCLUDED, EGENIX.COM MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED. BY WAY OF EXAMPLE, BUT NOT LIMITATION, EGENIX.COM MAKES NO AND DISCLAIMS ANY REPRESENTATION OR WARRANTY OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE SOFTWARE WILL NOT INFRINGE ANY THIRD PARTY RIGHTS.

4. LIMITATION OF LIABILITY

EGENIX.COM SHALL NOT BE LIABLE TO LICENSEE OR ANY OTHER USERS OF THE SOFTWARE FOR ANY INCIDENTAL, SPECIAL, OR CONSEQUENTIAL DAMAGES OR LOSS (INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF BUSINESS PROFITS, BUSINESS INTERRUPTION, LOSS OF BUSINESS INFORMATION, OR OTHER PECUNIARY LOSS) AS A RESULT OF USING, MODIFYING OR

DISTRIBUTING THE SOFTWARE, OR ANY DERIVATIVE THEREOF, EVEN IF ADVISED OF THE POSSIBILITY THEREOF.

SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THE ABOVE EXCLUSION OR LIMITATION MAY NOT APPLY TO LICENSEE.

5. Termination

This License Agreement will automatically terminate upon a material breach of its terms and conditions.

6. Third Party Rights

Any software or documentation in source or binary form provided along with the Software that is associated with a separate license agreement is licensed to Licensee under the terms of that license agreement. This License Agreement does not apply to those portions of the Software. Copies of the third party licenses are included in the Software Distribution.

7. General

Nothing in this License Agreement affects any statutory rights of consumers that cannot be waived or limited by contract.

Nothing in this License Agreement shall be deemed to create any relationship of agency, partnership, or joint venture between eGenix.com and Licensee.

If any provision of this License Agreement shall be unlawful, void, or for any reason unenforceable, such provision shall be modified to the extent necessary to render it enforceable without losing its intent, or, if no such modification is possible, be severed from this License Agreement and shall not affect the validity and enforceability of the remaining provisions of this License Agreement.

This License Agreement shall be governed by and interpreted in all respects by the law of Germany, excluding conflict of law provisions. It shall not be governed by the United Nations Convention on Contracts for International Sale of Goods.

This License Agreement does not grant permission to use eGenix.com trademarks or trade names in a trademark sense to endorse or promote products or services of Licensee, or any third party.

The controlling language of this License Agreement is English. If Licensee has received a translation into another language, it has been provided for Licensee's convenience only.

8. Agreement

By downloading, copying, installing or otherwise using the Software, Licensee agrees to be bound by the terms and conditions of this License Agreement.

For question regarding this License Agreement, please write to:

eGenix.com Software, Skills and Services GmbH

Pastor-Loeh-Str. 48

D-40764 Langenfeld

Germany