html5charref Documentation

Release 0.1.0

Brendan Abel

February 29, 2016

Contents

Python library for escaping/unescaping HTML5 Named Character References.

The standard python library includes the HTMLParser package for unescaping HTML named entities and HTML unicode escapes. Unfortunately, it doesn't include any of the named character entity references defined in HTML5. This library intends to provide a solution for escaping/unescaping HTML character references defined in HTML5.

CHAPTER 1

Installation

This project is still under development, so you should install it via GitHub instead of PyPI:

```
pip install git+https://github.com/bpabel/html5charref.git
```

CHAPTER 2

Usage

The main purpose of html5charref is to unescape HTML named entities. It will also handle HTML unicode character escapes.

```
html = u'This has &copy; and &lt; and &#x000a9; symbols'
print html5charref.unescape(html)
# u'This has \uxa9 and < and \uxa9 symbols'
```

You can also use html5charref to find the HTML5 named entity for a given unicode character.

```
import html5charref
# The copyright character
print html5charref.escape_char(u'\u00a9')
# u'&copy;'
```

CHAPTER 3

Updating Named Entity References

It is possible that additional named entity references will be added to the HTLM5 spec. You can update the list maintained by html5charref using the *update_charrefs()* function. This queries the latest named entity definitions from the w3 HTML5 site.

```
import html5charref
html5charref.update_charrefs()
```

CHAPTER 4

Licensing

This project is licensed under the MIT license.

CHAPTER 5

API Reference

html5charref.**escape_char**(*c*, *named_only=False*)

Return an HTML5 named character reference for the given unicode character. If no character entity reference is available, return a an html unicode escape, or the original unicode char if that cannot be done. Characters that are part of ASCII are not escaped.

Parameters **named_only** (*bool*) – If set to True, will only try to use named entities. If a named entity can't be found, the original character will be returned instead of an html unicode escape.

Note: Because several character references may refer to the same unicode point, the returned character reference may not be the one you expect. Use the *escape_char_advanced()* function to get a list of all named character references for a given unicode point and choose the specific one you want.

html5charref.**escape_char_advanced**(*c*)

Return a list of all HTML5 named character references for the given unicode character.

html5charref.**unescape**(*html*)

Return a unicode string with html character entity references and html unicode escapes converted to their unicode equivalent.

This closely matches HTMLParser.unescape(), but supports the HTML5 named entities.

html5charref.**unescape_charref**(*charref*)

Return the matching unicode character for the given HTML5 named character reference.

html5charref.**update_charrefs**()

Update the named entity dictionary from the w3 html5 specification site.

h

E

H

U