requests-html Documentation

Release v0.3.4

Kenneth Reitz

Feb 28, 2018

Contents:

This library intends to make parsing HTML (e.g. scraping the web) as simple and intuitive as possible.

When using this library you automatically get:

- **Full JavaScript support**!
- *CSS Selectors* (a.k.a jQuery-style, thanks to PyQuery).
- *XPath Selectors*, for the faint at heart.
- Mocked user-agent (like a real web browser).
- Automatic following of redirects.
- Connection–pooling and cookie persistence.
- The Requests experience you know and love, with magical parsing abilities.

CHAPTER 1

Installation

```
$ pipenv install requests-html
```

Only later versions of **Python 3** are supported.

CHAPTER 2

Tutorial & Usage

Make a GET request to 'python.org', using Requests:

```
>>> from requests_html import HTMLSession
>>> session = HTMLSession()

>>> r = session.get('https://python.org/')
```

Grab a list of all links on the page, as–is (anchors excluded):

```
>>> r.html.links
{'//docs.python.org/3/tutorial/', '/about/apps/', 'https://github.com/python/
↪pythondotorg/issues', '/accounts/login/', '/dev/peps/', '/about/legal/', '//docs.
↪python.org/3/tutorial/introduction.html#lists', '/download/alternatives', 'http://
↪feedproxy.google.com/~r/PythonInsider/~3/kihd2DW98YY/python-370a4-is-available-for-
↪testing.html', '/download/other/', '/downloads/windows/', 'https://mail.python.org/
↪mailman/listinfo/python-dev', '/doc/av', 'https://devguide.python.org/', '/about/
↪success/#engineering', 'https://wiki.python.org/moin/PythonEventsCalendar
↪#Submitting_an_Event', 'https://www.openstack.org', '/about/gettingstarted/',
↪'http://feedproxy.google.com/~r/PythonInsider/~3/AMoBel8b8Mc/python-3.html', '/
↪success-stories/industrial-light-magic-runs-python/', 'http://docs.python.org/3/
↪tutorial/introduction.html#using-python-as-a-calculator', '/', 'http://pyfound.
↪blogspot.com/', '/events/python-events/past/', '/downloads/release/python-2714/',
↪'https://wiki.python.org/moin/PythonBooks', 'http://plus.google.com/+Python',
↪'https://wiki.python.org/moin/', 'https://status.python.org/', '/community/
↪workshops/', '/community/lists/', 'http://buildbot.net/', '/community/awards',
↪'http://twitter.com/ThePSF', 'https://docs.python.org/3/license.html', '/psf/
↪donations/', 'http://wiki.python.org/moin/Languages', '/dev/', '/events/python-user-
↪group/', 'https://wiki.qt.io/PySide', '/community/sigs/', 'https://wiki.gnome.org/
↪Projects/PyGObject', 'http://www.ansible.com', 'http://www.saltstack.com', 'http://
↪planetpython.org/', '/events/python-events', '/about/help/', '/events/python-user-
↪group/past/', '/about/success/', '/psf-landing/', '/about/apps', '/about/', 'http://
↪www.wxpython.org/', '/events/python-user-group/665/', 'https://www.python.org/psf/
↪codeofconduct/', '/dev/peps/peps.rss', '/downloads/source/', '/psf/sponsorship/
↪sponsors/', 'http://bottlepy.org', 'http://roundup.sourceforge.net/', 'http://
↪pandas.pydata.org/', 'http://brochure.getpython.info/', 'https://bugs.python.org/',
↪'/community/merchandise/', 'http://tornadoweb.org', '/events/python-user-group/650/
↪', 'http://flask.pocoo.org/', '/downloads/release/python-364/', '/events/python-
↪user-group/660/', '/events/python-user-group/638/', '/psf/', '/doc/', 'http://blog.
↪python.org', '/events/python-events/604/', '/about/success/#government', 'http://
↪python.org/dev/peps/', 'https://docs.python.org', 'http://feedproxy.google.com/~r/
↪PythonInsider/~3/zVC80sq9s00/python-364-is-now-available.html', '/users/membership/
↪', '/about/success/#arts', 'https://wiki.python.org/moin/Python2orPython3', '/
```

Grab a list of all links on the page, in absolute form (anchors excluded):

```
>>> r.html.absolute_links
{'https://github.com/python/pythondotorg/issues', 'https://docs.python.org/3/tutorial/
↪', 'https://www.python.org/about/success/', 'http://feedproxy.google.com/~r/
↪PythonInsider/~3/kihd2DW98YY/python-370a4-is-available-for-testing.html', 'https://
↪www.python.org/dev/peps/', 'https://mail.python.org/mailman/listinfo/python-dev',
↪'https://www.python.org/doc/', 'https://www.python.org/', 'https://www.python.org/
↪about/', 'https://www.python.org/events/python-events/past/', 'https://devguide.
↪python.org/', 'https://wiki.python.org/moin/PythonEventsCalendar#Submitting_an_Event
↪', 'https://www.openstack.org', 'http://feedproxy.google.com/~r/PythonInsider/~3/
↪AMoBel8b8Mc/python-3.html', 'https://docs.python.org/3/tutorial/introduction.html
↪#lists', 'http://docs.python.org/3/tutorial/introduction.html#using-python-as-a-
↪calculator', 'http://pyfound.blogspot.com/', 'https://wiki.python.org/moin/
↪PythonBooks', 'http://plus.google.com/+Python', 'https://wiki.python.org/moin/',
↪'https://www.python.org/events/python-events', 'https://status.python.org/',
↪'https://www.python.org/about/apps', 'https://www.python.org/downloads/release/
↪python-2714/', 'https://www.python.org/psf/donations/', 'http://buildbot.net/',
↪'http://twitter.com/ThePSF', 'https://docs.python.org/3/license.html', 'http://wiki.
↪python.org/moin/Languages', 'https://docs.python.org/faq/', 'https://jobs.python.org
↪', 'https://www.python.org/about/success/#software-development', 'https://www.
↪python.org/about/success/#education', 'https://www.python.org/community/logos/',
↪'https://www.python.org/doc/av', 'https://wiki.qt.io/PySide', 'https://www.python.
↪org/events/python-user-group/660/', 'https://wiki.gnome.org/Projects/PyGObject',
↪'http://www.ansible.com', 'http://www.saltstack.com', 'https://www.python.org/dev/
↪peps/peps.rss', 'http://planetpython.org/', 'https://www.python.org/events/python-
↪user-group/past/', 'https://docs.python.org/3/tutorial/controlflow.html#defining-
↪functions', 'https://www.python.org/community/diversity/', 'https://docs.python.org/
↪3/tutorial/controlflow.html', 'https://www.python.org/community/awards', 'https://
↪www.python.org/events/python-user-group/638/', 'https://www.python.org/about/legal/
↪', 'https://www.python.org/dev/', 'https://www.python.org/download/alternatives',
↪'https://www.python.org/downloads/', 'https://www.python.org/community/lists/',
↪'http://www.wxpython.org/', 'https://www.python.org/about/success/#government',
↪'https://www.python.org/psf/', 'https://www.python.org/psf/codeofconduct/', 'http://
↪bottlepy.org', 'http://roundup.sourceforge.net/', 'http://pandas.pydata.org/',
↪'http://brochure.getpython.info/', 'https://www.python.org/downloads/source/',
↪'https://bugs.python.org/', 'https://www.python.org/downloads/mac-osx/', 'https://
↪www.python.org/about/help/', 'http://tornadoweb.org', 'http://flask.pocoo.org/',
↪'https://www.python.org/users/membership/', 'http://blog.python.org', 'https://www.
↪python.org/privacy/', 'https://www.python.org/about/gettingstarted/', 'http://
↪python.org/dev/peps/', 'https://www.python.org/about/apps/', 'https://docs.python.
↪org', 'https://www.python.org/success-stories/', 'https://www.python.org/community/
↪forums/', 'http://feedproxy.google.com/~r/PythonInsider/~3/zVC80sq9s00/python-364-
↪is-now-available.html', 'https://www.python.org/community/merchandise/', 'https://
↪www.python.org/about/success/#arts', 'https://wiki.python.org/moin/Python2orPython3
↪', 'http://trac.edgewall.org/', 'http://feedproxy.google.com/~r/PythonInsider/~3/
↪wh73_1A-N7Q/python-355rc1-and-python-348rc1-are-now.html', 'https://pypi.python.org/
↪', 'https://www.python.org/events/python-user-group/650/', 'http://www.
↪riverbankcomputing.co.uk/software/pyqt/intro', 'https://www.python.org/about/quotes/
↪', 'https://www.python.org/downloads/windows/', 'https://www.python.org/events/
↪calendars/', 'http://www.scipy.org', 'https://www.python.org/community/workshops/',
↪'https://www.python.org/blogs/', 'https://www.python.org/accounts/signup/', 'https:/
↪/www.python.org/events/', 'https://kivy.org/', 'http://www.facebook.com/pythonlang?
↪fref=ts', 'http://www.web2py.com/', 'https://www.python.org/psf/sponsorship/
↪sponsors/', 'https://www.python.org/community/', 'https://www.python.org/download/
↪other/', 'https://www.python.org/psf-landing/', 'https://www.python.org/events/
↪python-user-group/665/', 'https://wiki.python.org/moin/BeginnersGuide', 'https://
↪www.python.org/accounts/login/', 'https://www.python.org/downloads/release/python-
↪364/', 'https://www.python.org/dev/core-mentorship/', 'https://www.python.org/about/
↪success/#business', 'https://www.python.org/community/sigs/', 'https://www.python.
↪org/events/python-user-group/', 'http://ipython.org', 'https://www.python.org/shell/
↪', 'https://www.python.org/community/irc/', 'https://www.python.org/about/success/
↪#engineering', 'http://www.pylonsproject.org/', 'http://pycon.blogspot.com/',
↪'https://www.python.org/about/success/#scientific', 'https://www.python.org/doc/
```

Select an `Element` with a CSS Selector (learn more):

```
>>> about = r.html.find('#about', first=True)
```

Grab an `Element`'s text contents:

```
>>> print(about.text)
About
Applications
Quotes
Getting Started
Help
Python Brochure
```

Introspect an `Element`'s attributes (learn more):

```
>>> about.attrs
{'id': 'about', 'class': ('tier-1', 'element-1'), 'aria-haspopup': 'true'}
```

Render out an `Element`'s HTML:

```
>>> about.html
'<li aria-haspopup="true" class="tier-1 element-1 " id="about">\n<a class="" href="/
↪about/" title="">About</a>\n<ul aria-hidden="true" class="subnav menu" role="menu">
↪\n<li class="tier-2 element-1" role="treeitem"><a href="/about/apps/" title="">
↪Applications</a></li>\n<li class="tier-2 element-2" role="treeitem"><a href="/about/
↪quotes/" title="">Quotes</a></li>\n<li class="tier-2 element-3" role="treeitem"><a␣
↪href="/about/gettingstarted/" title="">Getting Started</a></li>\n<li class="tier-2␣
↪element-4" role="treeitem"><a href="/about/help/" title="">Help</a></li>\n<li class=
↪"tier-2 element-5" role="treeitem"><a href="http://brochure.getpython.info/" title="
↪">Python Brochure</a></li>\n</ul>\n</li>'
```

Select an `Element` list within an `Element`:

```
>>> about.find('a')
[<Element 'a' href='/about/' title='' class=''>, <Element 'a' href='/about/apps/'␣
↪title=''>, <Element 'a' href='/about/quotes/' title=''>, <Element 'a' href='/about/
↪gettingstarted/' title=''>, <Element 'a' href='/about/help/' title=''>, <Element 'a
↪' href='http://brochure.getpython.info/' title=''>]
```

Search for links within an element:

```
>>> about.absolute_links
{'http://brochure.getpython.info/', 'https://www.python.org/about/gettingstarted/',
↪'https://www.python.org/about/', 'https://www.python.org/about/quotes/', 'https://
↪www.python.org/about/help/', 'https://www.python.org/about/apps/'}
```

Search for text on the page:

```
>>> r.html.search('Python is a {} language')[0]
programming
```

More complex CSS Selector example (copied from Chrome dev tools):

```
>>> r = session.get('https://github.com/')
>>> sel = 'body > div.application-main > div.jumbotron.jumbotron-codelines > div >␣
↪div > div.col-md-7.text-center.text-md-left > p'
```

```
>>> print(r.html.find(sel, first=True).text)
GitHub is a development platform inspired by the way you work. From open source to␣
↪business, you can host and review code, manage projects, and build software␣
↪alongside millions of other developers.
```

XPath is also supported (learn more):

```
>>> r.html.xpath('a')
[<Element 'a' class='btn' href='https://help.github.com/articles/supported-browsers'>]
```

CHAPTER 3

JavaScript Support

Let's grab some text that's rendered by JavaScript:

```
>>> r = session.get('http://python-requests.org/')

>>> r.html.render()

>>> r.html.search('Python 2 will retire in only {months} months!')['months']
'<time>25</time>'
```

Note, the first time you ever run the `render()` method, it will download Chromium into your home directory (e.g. `~/.pyppeteer/`). This only happens once.

CHAPTER 4

Using without Requests

You can also use this library without Requests:

```
>>> from requests_html import HTML
>>> doc = """<a href='https://httpbin.org'>"""

>>> html = HTML(html=doc)
>>> html.links
{'https://httpbin.org'}
```

CHAPTER 5

API Documentation

5.1 Main Classes

These classes are the main interface to `requests-html`:

class `requests_html.`**HTML**(*, *url='https://example.org/'*, *html*, *default_encoding='utf-8'*) → None
An HTML document, ready for parsing.

absolute_links
All found links on page, in absolute form (learn more).

base_url
The base URL for the page. Supports the `<base>` tag (learn more).

encoding
The encoding string to be used, extracted from the HTML and `HTMLResponse` headers.

find(*selector: str*, *first: bool = False*, *_encoding: str = None*)
Given a CSS Selector, returns a list of *`Element`* objects.

Example CSS Selectors:

- `a`
- `a.someClass`
- `a#someID`
- `a[target=_blank]`

See W3School's CSS Selectors Reference for more details.

If `first` is `True`, only returns the first *`Element`* found.

full_text
The full text content (including links) of the *`Element`* or *`HTML`*..

html
Unicode representation of the HTML content (learn more).

links
All found links on page, in as–is form.

lxml
lxml representation of the *Element* or *HTML*.

pq
PyQuery representation of the *Element* or *HTML*.

raw_html
Bytes representation of the HTML content (learn more).

render (*retries: int = 8*, *script: str = None*, *scrolldown=False*, *sleep: int = 0*)
Reloads the response in Chromium, and replaces HTML content with an updated version, with JavaScript executed.

If `scrolldown` is specified, the page will scrolldown the specified number of times, after sleeping the specified amount of time (e.g. `scrolldown=10, sleep=1`).

If just `sleep` is provided, the rendering will wait *n* seconds, before returning.

If `script` is specified, it will execute the provided JavaScript at runtime. Example:

```
script = """
    () => {
        return {
            width: document.documentElement.clientWidth,
            height: document.documentElement.clientHeight,
            deviceScaleFactor: window.devicePixelRatio,
        }
    }
"""
```

Returns the return value of the executed `script`, if any is provided:

```
>>> r.html.render(script=script)
{'width': 800, 'height': 600, 'deviceScaleFactor': 1}
```

Warning: the first time you run this method, it will download Chromium into your home directory (`~/.pyppeteer`).

search (*template: str*) → parse.Result
Searches the *Element* for the given parse template.

search_all (*template: str*) → parse.Result
Searches the *Element* (multiple times) for the given parse template.

set_html
Unicode representation of the HTML content (learn more).

text
The text content of the *Element* or *HTML*.

xpath (*selector: str*, *first: bool = False*, *_encoding: str = None*)
Given an XPath selector, returns a list of *Element* objects.

If a sub-selector is specified (e.g. `//a/@href`), a simple list of results is returned.

See W3School's XPath Examples for more details.

If `first` is `True`, only returns the first *Element* found.

class requests_html.**Element**(*, *element*, *url*, *default_encoding*) → None
An element of HTML.

absolute_links
All found links on page, in absolute form (learn more).

attrs
Returns a dictionary of the attributes of the *Element* (learn more).

base_url
The base URL for the page. Supports the <base> tag (learn more).

encoding
The encoding string to be used, extracted from the HTML and HTMLResponse headers.

find(*selector: str*, *first: bool = False*, *_encoding: str = None*)
Given a CSS Selector, returns a list of *Element* objects.

Example CSS Selectors:

- a
- a.someClass
- a#someID
- a[target=_blank]

See W3School's CSS Selectors Reference for more details.

If first is True, only returns the first *Element* found.

full_text
The full text content (including links) of the *Element* or *HTML*..

html
Unicode representation of the HTML content (learn more).

links
All found links on page, in as–is form.

lxml
lxml representation of the *Element* or *HTML*.

pq
PyQuery representation of the *Element* or *HTML*.

raw_html
Bytes representation of the HTML content (learn more).

search(*template: str*) → parse.Result
Searches the *Element* for the given parse template.

search_all(*template: str*) → parse.Result
Searches the *Element* (multiple times) for the given parse template.

set_html
Unicode representation of the HTML content (learn more).

text
The text content of the *Element* or *HTML*.

xpath(*selector: str*, *first: bool = False*, *_encoding: str = None*)
Given an XPath selector, returns a list of *Element* objects.

If a sub-selector is specified (e.g. //a/@href), a simple list of results is returned.

See W3School's XPath Examples for more details.

If `first` is `True`, only returns the first *`Element`* found.

5.2 Utility Functions

`requests_html.`**`user_agent`**(*style='chrome'*) → str
Returns a random user-agent, if not requested one of a specific style. Defaults to a Chrome-style User-Agent.

5.3 HTML Sessions

These sessions are for making HTTP requests:

class `requests_html.`**`HTMLSession`**(*mock_browser=True*, **args*, ***kwargs*)
A consumable session, for cookie persistence and connection pooling, amongst other things.

`close`()
Closes all adapters and as such the session

`delete`(*url*, ***kwargs*)
Sends a DELETE request. Returns `Response` object.

Parameters

- **`url`** – URL for the new `Request` object.
- **`**kwargs`** – Optional arguments that `request` takes.

Return type requests.Response

`get`(*url*, ***kwargs*)
Sends a GET request. Returns `Response` object.

Parameters

- **`url`** – URL for the new `Request` object.
- **`**kwargs`** – Optional arguments that `request` takes.

Return type requests.Response

`get_adapter`(*url*)
Returns the appropriate connection adapter for the given URL.

Return type requests.adapters.BaseAdapter

`get_redirect_target`(*resp*)
Receives a Response. Returns a redirect URI or `None`

`head`(*url*, ***kwargs*)
Sends a HEAD request. Returns `Response` object.

Parameters

- **`url`** – URL for the new `Request` object.
- **`**kwargs`** – Optional arguments that `request` takes.

Return type requests.Response

`merge_environment_settings`(*url*, *proxies*, *stream*, *verify*, *cert*)
Check the environment and merge it with some settings.

Return type dict

mount(*prefix*, *adapter*)
Registers a connection adapter to a prefix.

Adapters are sorted in descending order by prefix length.

options(*url*, ***kwargs*)
Sends a OPTIONS request. Returns `Response` object.

Parameters

- **url** – URL for the new `Request` object.
- ****kwargs** – Optional arguments that `request` takes.

Return type requests.Response

patch(*url*, *data=None*, ***kwargs*)
Sends a PATCH request. Returns `Response` object.

Parameters

- **url** – URL for the new `Request` object.
- **data** – (optional) Dictionary, bytes, or file-like object to send in the body of the `Request`.
- ****kwargs** – Optional arguments that `request` takes.

Return type requests.Response

post(*url*, *data=None*, *json=None*, ***kwargs*)
Sends a POST request. Returns `Response` object.

Parameters

- **url** – URL for the new `Request` object.
- **data** – (optional) Dictionary, bytes, or file-like object to send in the body of the `Request`.
- **json** – (optional) json to send in the body of the `Request`.
- ****kwargs** – Optional arguments that `request` takes.

Return type requests.Response

prepare_request(*request*)
Constructs a `PreparedRequest` for transmission and returns it. The `PreparedRequest` has settings merged from the `Request` instance and those of the `Session`.

Parameters **request** – `Request` instance to prepare with this session's settings.

Return type requests.PreparedRequest

put(*url*, *data=None*, ***kwargs*)
Sends a PUT request. Returns `Response` object.

Parameters

- **url** – URL for the new `Request` object.
- **data** – (optional) Dictionary, bytes, or file-like object to send in the body of the `Request`.
- ****kwargs** – Optional arguments that `request` takes.

Return type requests.Response

rebuild_auth(*prepared_request*, *response*)
When being redirected we may want to strip authentication from the request to avoid leaking credentials. This method intelligently removes and reapplies authentication where possible to avoid credential loss.

rebuild_method(*prepared_request*, *response*)
When being redirected we may want to change the method of the request based on certain specs or browser behavior.

rebuild_proxies(*prepared_request*, *proxies*)
This method re-evaluates the proxy configuration by considering the environment variables. If we are redirected to a URL covered by NO_PROXY, we strip the proxy configuration. Otherwise, we set missing proxy keys for this URL (in case they were stripped by a previous redirect).

This method also replaces the Proxy-Authorization header where necessary.

Return type dict

resolve_redirects(*resp*, *req*, *stream=False*, *timeout=None*, *verify=True*, *cert=None*, *proxies=None*, *yield_requests=False*, ***adapter_kwargs*)
Receives a Response. Returns a generator of Responses or Requests.

send(*request*, ***kwargs*)
Send a given PreparedRequest.

Return type requests.Response

CHAPTER 6

Indices and tables

- genindex
- modindex
- search

Python Module Index

r

Index

T

U

X