
nvhtml Documentation

Release 0.5

dli

Apr 16, 2020

Contents

1	ReadMe	3
1.1	Installation	3
1.2	Documentation	3
1.3	License	3
1.4	Quickstart	4
1.5	Features	4
1.6	References	4
2	Installation	5
2.1	From pip	5
2.2	From sources	5
3	Usage	7
3.1	Command Line	7
3.2	Examples	20
3.3	lvsrch	26
3.4	engine	29
4	Package API	31
4.1	engine	31
4.2	lvsrch	31
5	Contributing	33
5.1	Types of Contributions	33
5.2	Pull Request Guidelines	33
6	Credits	35
6.1	Developers	35
6.2	Contributors	35
7	History	37
7.1	0.0.1 (2018-12-18)	37
8	TODO	39
9	Indices and tables	41



CHAPTER 1

ReadMe



1.1 Installation

```
$ pip3 install nvhtml
```

- refer to [Installation](#)

1.2 Documentation

- refer to <https://nvhtml.readthedocs.io>

1.3 License

- MIT

1.4 Quickstart

- refer to [Usage](#)

1.5 Features

- search html via tags-pathlist
- search html via breadth,depth
- search html via tag
- width first traverse html
- description matrix of html
- depth first traverse html (sax)
- beautify html
- **get html node via relationship:**
 - ancestor,descendants,parent,grand-parent,children,
 - siblings,lsib,rcib,preceding-sib,following-sib,
 - lcin,rcin

1.6 References

- [elist](#)
- <https://github.com/ihgazni2/elist>
- [mlconjug](#)(for docs template format)
- <https://github.com/SekouD/mlconjug/tree/master/docs>

2.1 From pip

To install nvhtml, run this command in your terminal:

```
$ pip3 install nvhtml
```

2.2 From sources

The sources for nvhtml can be downloaded from the github repo:

```
$ git clone git@github.com:ihgazni2/nvhtml.git
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```


3.1 Command Line

3.1.1 rshtml write rule

pound *comment*
- dash *attrib*
. dot *text and tail*
| vbar *text and tail content can multiline*

3.1.2 rshtml

```
NVHTML-BENCH# cat tstrs.rshtml
html
  head
    meta
      -http-equiv X-UA-Compatible
      -content IE=edge,chrome=1
    meta
      -name viewport
      -content user-scalable=yes, initial-scale=1.0, minimum-scale=1.0,
↪ maximum-scale=3.0
    link
  body
    div
      -id menu-item-27961
      -class qtranxs-lang-menu-item menu-item-object-custom
      .text
        |hello
        |hihihi
```

(continues on next page)

(continued from previous page)

```

        .tail
        |this is a tail
    li
div
    li
div
#comment
    .text
    |this is acomment
NVHTML-BENCH#

NVHTML-BENCH# nvhtml_rsh2html -input tstrs.rshtml

NVHTML-BENCH# ls -l | egrep tstrs.rshtml
-rw-r--r-- 1 root root    597 Aug 19 11:36 tstrs.rshtml
-rw-r--r-- 1 root root    658 Aug 19 11:56 tstrs.rshtml.html

NVHTML-BENCH# cat tstrs.rshtml.html
<html>
  <head>
    <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
    </meta>
    <meta name="viewport" content="user-scalable=yes, initial-scale=1.0,
↳minimum-scale=1.0, maximum-scale=3.0">
    </meta>
    <link>
    </link>
  </head>
  <body>
    <div id="menu-item-27961" class="qtranxs-lang-menu-item menu-item-
↳object-custom">
      hello
      hihihi
      <li>
      </li>
    </div>
    this is a tail
    <div>
      <li>
      </li>
    </div>
    <div>
    </div>
  </body>
  <!--
    this is acomment
  -->
</html>
NVHTML-BENCH#

vice versa
NVHTML-BENCH# nvhtml_html2rsh -input tstrs.html

```

3.1.3 vertical show

with color

```
NVHTML-BENCH# nvhtml_struct_show -input disp.html
```

```
color:  True
```

```
-----
|          html          |
|-----|
| head |          body   | |
|---|---|---|
|      | div |      div  | |
|---|---|---|---|
|      |      | div | div |
|-----|
```

```
NVHTML-BENCH#
NVHTML-BENCH# nvhtml_struct_show -input disp.html
```

```
-----
|          html          |
|-----|
| head |          body   | |
|---|---|---|
|      | div |      div  | |
|---|---|---|---|
|      |      | div | div |
|-----|
```

```
NVHTML-BENCH#
```

```
NVHTML-BENCH#
```

no color

```
#if too big, disable color , and open it in editor such as notepad ++
NVHTML-BENCH# nvhtml_struct_show -input opis.html -color "no" > html_txt_tb
```

```

-----
.....footer.....
-----
v.....
-----
|.....div.....|
-----
|.....ul.....|
-----
|.....li.....| style |.....li.....|
-----
| h3 | .....form..... | ..... | h3 | .....ul..... | div |
-----
| ..... | .....label..... | .....button..... | ..... | .....li..... | .....li..... | .....li..... |
-----
| ..... | .....span..... | .....input..... | .....span..... | i | ..... | ..... | .....a..... | .....a..... | .....a..... |
-----
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | .....span..... | .....span..... | .....span..... |
-----
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
-----
| ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... |
-----

```

nvssh_struct_show

```
TEST# nvssh_struct_show -input jobj2.rshtml
```

```
color: True
```

```

-----
|          html          |
-----
|          body          |
-----
|      im-dict-root      |
-----
|   l   |   t   |   s   |
-----
| o0 | o1 | o0 | o1 | u0 | u1 |
-----
| a0 | a1 | t0 | t1 | s0 | s1 |
-----

```

```
NVHTML# cat TEST/jobj2.rshtml
```

```
im-dict-root
```

```
  l
```

```
    o0
```

```
      a0
```

```
    o1
```

```
      a1
```

```
  t
```

```
    o0
```

```
      t0
```

(continues on next page)

(continued from previous page)

```

        o1
          t1
s
        u0
          s0
        u1
          s1
NVHTML#

```

3.1.4 beautify

```

nvhtml_beauty -input opis.html
vim opis.html.out.html

# nvhtml_beauty -h
usage: nvhtml_beauty [-h] [-input INPUT_HTML_FILE] [-output OUTPUT_HTML_
↪FILE] [-codec INPUT_CODEC]

optional arguments:
  -h,            --help                show this help message and exit
  -input         --input_html_file     input html file name
  -output        --output_html_file    output html file name
  -codec         --input_codec         input html file codec

```

3.1.5 search with loc

```

NVHTML-BENCH# nvhtml_loc -h
usage: nvhtml_loc [-h] [-input INPUT_HTML_FILE] [-codec INPUT_CODEC]
                [-loc MAT_LOC]

optional arguments:
  -h,            --help                show this help message and exit
  -input         --input_html_file     input html file name
  -codec         --input_codec         input html file codec
  -loc           --mat_loc             depth,breadth

NVHTML-BENCH# nvhtml_loc -input opis.html -loc 11,2
<-----json-----
{
  'pl':
    [
      'html',
      'body',
      'div',
      'div',
      'header',
      'div',
      'nav',
      'div',
      'ul',
      'li',

```

(continues on next page)

(continued from previous page)

```

        'ul',
        'li'
    ],
    'breadth': 2,
    'depth': 11,
    'pbreadth': 2,
    'samepl_sibseq': 0,
    'samepl_breadth': 0,
    'tag': 'li',
    'sibseq': 0,
    'attrib':
        {
            'id': 'menu-item-22951',
            'class': 'menu-item menu-item-type-post_type menu-item-object-
↪page menu-item-22951'
        },
    'text': None,
    'tail': '\n\t',
    'text_intag': ''
}
-----json---->

<-----html-----
<li id="menu-item-22951" class="menu-item menu-item-type-post_type menu-item-
↪object-page menu-item-22951">
    <a href="https://opistobranquis.info/en/home/presentacio/">
        <span>
            Presentation
        </span>
    </a>
</li>

<-----html-----

```

3.1.6 tag search with depth

```

NVHTML-BENCH# nvhtml_tag -h
usage: nvhtml_tag [-h] [-input INPUT_HTML_FILE] [-codec INPUT_CODEC]
                  [-tag HTML_TAG] [-which WHICH_TAG]
                  [-sdepth START_LEVEL_DEPTH] [-edepth END_LEVEL_DEPTH]

optional arguments:
  -h, --help                show this help message and_
↪exit
  -input, --input_html_file  input html file name
  -codec, --input_codec      input html file codec
  -tag, --html_tag           html tag for search
  -which, --which_tag        sequence of tag-array
  -sdepth, --start_level_depth  start level depth
  -edepth, --end_level_depth   end level depth
NVHTML-BENCH#

NVHTML-BENCH# nvhtml_tag -input opis.html.out.html -tag img -sdepth 3 -
↪edepth 6


```

(continues on next page)

(continued from previous page)

```

</img>


</img>


</img>


</img>


</img>

NVHTML-BENCH#

```

3.1.7 search with tags-path

```

NVHTML-BENCH# nvhtml_tgpth -input opis.html -tgpth html.body.di
[
  'div',
  'div',
  'div',
  'div'
]

NVHTML-BENCH# nvhtml_tgpth -input opis.html -tgpth html.body.div
[
  0,
  1,
  2,
  3
]

NVHTML-BENCH#
NVHTML-BENCH# nvhtml_tgpth -input opis.html -tgpth html.body.div.3
<div id="cookie-banner">

    <div id="cookie-banner-container">

        <div class="left">

            Our website uses cookies. By
↳accessing our website and
            <br>
            </br>
            agreeing to this policy, you consent to our use of cookies.
        </div>

        <div class="right">

```

(continues on next page)

(continued from previous page)

```

        <a class="accept" href="#">
            ACCEPT
        </a>

        <a class="more-info" href="https://opistobranquis.info/1HWEw">
            MORE INFO
        </a>

    </div>

</div>
</div>
NVHTML-BENCH#

usage: nvhtml_tgpth [-h] [-input INPUT_HTML_FILE] [-codec INPUT_CODEC]
                  [-tgpth TAG_PATH]

optional arguments:
  -h, --help            show this help message_
  -input INPUT_HTML_FILE  input html file name
  -codec INPUT_CODEC      input html file codec
  -tgpth TAG_PATH         html tag dot path

```

3.1.8 html to db

```

NVHTML-BENCH# nvhtml_sqlite -input opis.html
db: ./opis.html.sqlite.db
table: tb_html
NVHTML-BENCH#

NVHTML-BENCH# sqlite3 opis.html.sqlite.db
SQLite version 3.22.0 2018-01-22 18:45:57
Enter ".help" for usage hints.
sqlite>
sqlite> .table
tb_html
sqlite>
sqlite> .schema tb_html
CREATE TABLE IF NOT EXISTS "tb_html" (
  "index" INTEGER,
  "_pl" TEXT,
  "_breadth" TEXT,
  "_depth" TEXT,
  "_pbreadth" TEXT,
  "_samepl_sibseq" TEXT,
  "_samepl_breadth" TEXT,
  "_tag" TEXT,
  "_sibseq" TEXT,
  "_text" TEXT,
  "_tail" TEXT,

```

(continues on next page)

(continued from previous page)

```

"class" TEXT,
"href" TEXT,
"id" TEXT,
"style" TEXT,
"type" TEXT,
"src" TEXT,
"rel" TEXT,
"target" TEXT,
"title" TEXT,
"content" TEXT,
"alt" TEXT,
"media" TEXT,
"name" TEXT,
"align" TEXT,
"property" TEXT,
"role" TEXT,
"value" TEXT,
"data-shared" TEXT,
"hreflang" TEXT,
"for" TEXT,
"aria-current" TEXT,
"colspan" TEXT,
"method" TEXT,
"action" TEXT,
"sizes" TEXT,
"placeholder" TEXT,
"height" TEXT,
"width" TEXT,
"http-equiv" TEXT,
"autocomplete" TEXT,
"data-layout" TEXT,
"data-orig-file" TEXT,
"data-href" TEXT,
"lang" TEXT,
"data-image-title" TEXT,
"data-recalc-dims" TEXT,
"data-attachment-id" TEXT,
"data-text" TEXT,
"data-flxmap" TEXT,
"size" TEXT,
"data-image-description" TEXT,
"data-medium-file" TEXT,
"async" TEXT,
"language" TEXT,
"srcset" TEXT,
"data-comments-opened" TEXT,
"data-large-file" TEXT,
"data-via" TEXT,
"defer" TEXT,
"data-permalink" TEXT,
"data-image-meta" TEXT,
"data-noptimize" TEXT,
"data-url" TEXT,
"data-orig-size" TEXT
);
CREATE INDEX "ix_tb_html_index" ON "tb_html" ("index");
sqlite>

```

(continues on next page)

(continued from previous page)

```

sqlite>
sqlite> SELECT src FROM tb_html WHERE _tag=="img" AND src like "%facebook.png";
↪https://opistobranquis.info/wp-content/themes/tempera/images/socials/
↪Facebook.png
↪https://opistobranquis.info/wp-content/themes/tempera/images/socials/
↪Facebook.png
↪https://opistobranquis.info/wp-content/themes/tempera/images/socials/
↪Facebook.png
↪https://opistobranquis.info/wp-content/themes/tempera/images/socials/
↪Facebook.png
sqlite>
sqlite>
sqlite>
sqlite>
sqlite> SELECT href FROM tb_html WHERE _tag=="link" AND href like "%.com";
↪//s0.wp.com
↪//c0.wp.com
↪//i0.wp.com
↪//i1.wp.com
↪//i2.wp.com
sqlite>

```

3.1.9 html to dir

```

NVHTML-BENCH# mkdir TMP
NVHTML-BENCH# nvhtml_dir -input opis.html -wkdir TMP

NVHTML-BENCH# tree -fdL 4 TMP | head
TMP
├── TMP/html.0
│   ├── TMP/html.0/body.1
│   │   ├── TMP/html.0/body.1/<comment>.91
│   │   ├── TMP/html.0/body.1/div.90
│   │   │   ├── TMP/html.0/body.1/div.90/<comment>.4
│   │   │   ├── TMP/html.0/body.1/div.90/<comment>.7
│   │   │   ├── TMP/html.0/body.1/div.90/div.0
│   │   │   ├── TMP/html.0/body.1/div.90/div.1
│   │   │   └── TMP/html.0/body.1/div.90/div.2
NVHTML-BENCH#
NVHTML-BENCH# tree -fdL 4 TMP | tail
├── TMP/html.0/head.0/style.45
├── TMP/html.0/head.0/style.55
├── TMP/html.0/head.0/style.56
├── TMP/html.0/head.0/style.57
├── TMP/html.0/head.0/style.58
├── TMP/html.0/head.0/style.78
├── TMP/html.0/head.0/style.79
└── TMP/html.0/head.0/title.7

138 directories

NVHTML-BENCH# ls -l TMP/html.0/body.1/div.90/div.2

```

(continues on next page)

(continued from previous page)

```

total 36
drwxr-xr-x 3 root root 4096 Aug 11 02:49 a.3
drwxr-xr-x 3 root root 4096 Aug 11 02:49 a.4
-rw-r--r-- 1 root root    7 Aug 11 02:49 attrib.class
-rw-r--r-- 1 root root    7 Aug 11 02:49 attrib.id
-rw-r--r-- 1 root root  538 Aug 11 02:49 outter_html
-rw-r--r-- 1 root root    3 Aug 11 02:49 tag
-rw-r--r-- 1 root root    1 Aug 11 02:49 tail
-rw-r--r-- 1 root root    4 Aug 11 02:49 text
-rw-r--r-- 1 root root    8 Aug 11 02:49 text_intag
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/attrib.id
srights
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/attrib.class
socials
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/outter_html
<div class="socials" id="srights">
    <a target="_blank" href="https://twitter.com/InfoOpk
↪" class="socialicons social
-Twitter external" title="Twitter">
        
    </a>
    <a target="_blank" href="https://www.facebook.com/
↪OPK.Opistobranquis/" class="s
ocialicons social-Facebook external" title="Facebook">
        
    </a></div>
NVHTML-BENCH#

NVHTML-BENCH# ls -al TMP/html.0/body.1/div.90/div.2 | egrep " \.[a-z]"
-rw-r--r-- 1 root root    1 Aug 11 02:49 .breadth
-rw-r--r-- 1 root root    1 Aug 11 02:49 .depth
-rw-r--r-- 1 root root  27 Aug 11 02:49 .mkdir_pth
-rw-r--r-- 1 root root    2 Aug 11 02:49 .pbreadth
-rw-r--r-- 1 root root  18 Aug 11 02:49 .pl
-rw-r--r-- 1 root root    1 Aug 11 02:49 .samepl_breadth
-rw-r--r-- 1 root root    1 Aug 11 02:49 .samepl_sibseq
-rw-r--r-- 1 root root    1 Aug 11 02:49 .sibseq
NVHTML-BENCH#
NVHTML-BENCH#
NVHTML-BENCH#
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.breadth
2
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.depth
3
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.pbreadth
90
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.pl
/html/body/div/div
NVHTML-BENCH#
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.samepl_breadth
2
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.samepl_sibseq
2

```

(continues on next page)

(continued from previous page)

```
NVHTML-BENCH# more TMP/html.0/body.1/div.90/div.2/.sibseq
2
NVHTML-BENCH#
```

3.1.10 find all

```
NVHTML-BENCH# nvhtml_find_all -input opis.html -attrib "http-equiv"
[
  'X-UA-Compatible',
  'Content-Type'
]
NVHTML-BENCH#
NVHTML-BENCH# nvhtml_find_all -input opis.html -attrib "href" | egrep
↪ "jorunna-e"
  'https://opistobranquis.info/en/guia/nudibranchia/doridina/doridoidei/
↪ doridoidea/jorunna-efe/',
  'https://opistobranquis.info/en/guia/nudibranchia/doridina/doridoidei/
↪ doridoidea/jorunna-evansi/',
NVHTML-BENCH#

NVHTML-BENCH# nvhtml_find_all -input opis.html
common attribs:
[
  '_pl',
  '_breadth',
  '_depth',
  '_pbreadth',
  '_samepl_sibseq',
  '_samepl_breadth',
  '_tag',
  '_sibseq',
  '_text',
  '_tail'
]
attrib_names:frequency
{
  'class': 947,
  'href': 810,
  'id': 181,
  'style': 80,
  'type': 78,
  'src': 55,
  'rel': 49,
  'target': 41,
  'title': 36,
  'content': 23,
  'alt': 19,
  'media': 17,
  'name': 15,
  'align': 13,
  'property': 12,
  'role': 9,
  'value': 7,
  'hreflang': 4,
  'data-shared': 4,
```

(continues on next page)

(continued from previous page)

```

'colspan': 3,
'for': 3,
'aria-current': 3,
'sizes': 3,
'action': 3,
'method': 3,
'placeholder': 2,
'width': 2,
'http-equiv': 2,
'height': 2,
'data-permalink': 1,
'data-recalc-dims': 1,
'srcset': 1,
'size': 1,
'data-layout': 1,
'data-orig-size': 1,
'language': 1,
'data-medium-file': 1,
'data-href': 1,
'data-image-description': 1,
'data-image-title': 1,
'data-orig-file': 1,
'defer': 1,
'data-flxmap': 1,
'data-nooptimize': 1,
'data-image-meta': 1,
'lang': 1,
'data-url': 1,
'data-large-file': 1,
'autocomplete': 1,
'data-via': 1,
'async': 1,
'data-comments-opened': 1,
'data-attachment-id': 1,
'data-text': 1
}
NVHTML-BENCH#

```

3.1.11 nvhtml_wfs_udlrpls

```
NVHTML-BENCH#nvhtml_wfs_udlrpls -input xxx.html
```

3.1.12 nvhtml_wfs_dulrpls

```
NVHTML-BENCH#nvhtml_wfs_dulrpls -input xxx.html
```

3.2 Examples

3.2.1 tagsrch

```
from lxml.etree import HTML as LXHTML
from lxml.etree import XML as LXML
from xdict.jprint import pdir, pobj
from nvhtml import txt
from nvhtml import lvsrch
from nvhtml import fs
from nvhtml import engine
from nvhtml import utils
import lxml.sax
```

```
html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
eles = lvsrch.a(root, 7, 8, show=False)
print(eles[0])
print(eles[5])
eles = lvsrch.a(root, 7, 8, which=0)
eles = lvsrch.a(root, 7, 8, which=0, source=False)
```

```
>>>
>>> from lxml.etree import HTML as LXHTML
>>> from lxml.etree import XML as LXML
>>> from xdict.jprint import pdir, pobj
>>> from nvhtml import txt
>>> from nvhtml import lvsrch
>>> from nvhtml import fs
>>> from nvhtml import engine
>>> from nvhtml import utils
>>> import lxml.sax
>>>
>>>
>>> html_str = fs.rfile("./test.html")
>>> root = LXHTML(html_str)
>>> eles = lvsrch.a(root, 7, 8, show=False)
>>> print(eles[0])
<a href="/tags/index.asp" title="HTML 4.01 / XHTML 1.0 参考手册">
  标签列表（字母排序）
</a>
>>> print(eles[5])
<a href="/tags/html_ref_canvas.asp" title="HTML Canvas 参考手册">
  HTML 画布
</a>
>>> eles = lvsrch.a(root, 7, 8, which=0)
<a href="/tags/index.asp" title="HTML 4.01 / XHTML 1.0 参考手册">
  标签列表（字母排序）
</a>
>>> eles = lvsrch.a(root, 7, 8, which=0, source=False)
<Element a at 0x155fef4a648>
>>>
```


3.2.2 relation get

```
html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
ele = engine.xpath(root, "//div", 5)

engine.parent(ele)
engine.grand_parent(ele)
engine.ancestors(ele)
engine.parent(ele)
engine.grand_parent(ele)
engine.ancestors(ele)
engine.lsib(ele)
engine.rsib(ele)
engine.lcin(ele)
engine.rcin(ele)
engine.siblings(ele)
engine.descendants(ele, 5, 6)

engine.layer(ele)
engine.breadth(ele)
engine.depth(ele)
engine.pathlist(ele)
```

```

>>>
>>>
>>> html_str = fs.rfile("./test.html")
>>> root = LXHTML(html_str)
>>> ele = engine.xpath(root, "//div", 5)
>>>
>>> engine.parent(ele)
<Element div at 0x155fef7dac8>
>>> engine.grand_parent(ele)
<Element div at 0x155fef92d88>
>>> engine.ancestors(ele)
[<Element div at 0x155fef83c88>, <Element div at 0x155fef92d88>, <
Element body at 0x155fef838c8>, <Element html at 0x155fef5c588>]
>>> engine.parent(ele)
<Element div at 0x155fef83c88>
>>> engine.grand_parent(ele)
<Element div at 0x155fef98548>
>>> engine.ancestors(ele)
[<Element div at 0x155fef9be48>, <Element div at 0x155fef98548>, <
Element body at 0x155fef9ba88>, <Element html at 0x155fef5c588>]
>>> engine.lsisib(ele)
>>> engine.rsisib(ele)
<Element div at 0x155fef9a78c8>
>>> engine.lcin(ele)
<Element ul at 0x155fef9a8248>
>>> engine.rcin(ele)
>>> engine.siblings(ele)
[<Element div at 0x155fef5ad88>, <Element div at 0x155fef9e1c8>]
>>> engine.descendants(ele, 5, 6)
[<Element h2 at 0x155fefb0bc8>, <Element ul at 0x155fefb05c8>, <El
ement h2 at 0x155fefb0608>, <Element ul at 0x155fefb0f48>]
>>>
>>> engine.layer(ele)
[<Element a at 0x155fefc1888>, <Element div at 0x155fefc1d48>, <El
ement ul at 0x155fefc1808>, <Element div at 0x155fef5ad88>, <Eleme
nt div at 0x155fefc1588>, <Element h1 at 0x155fefc1fc8>, <Element
div at 0x155fefc1bc8>, <Element div at 0x155fefc1f88>, <Element di
v at 0x155fefc2208>, <Element div at 0x155fefc2188>, <Element div
at 0x155fefc2f88>, <Element div at 0x155fefc2ec8>, <Element p at 0
x155fefc2108>, <Element p at 0x155fefc2d48>]
>>> engine.breadth(ele)
3
>>> engine.depth(ele)
4
>>> engine.pathlist(ele)
['html', 'body', 'div', 'div', 'div']
>>>
>>>

```

3.2.3 description matrix

```
html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
wfs = engine.WFS(root)
pobj(wfs.mat[3][1])
```

```
>>>
>>> html_str = fs.rfile("./test.html")
>>> root = LXHTML(html_str)
R>>> wfs = engine.WFS(root)
>>> pobj(wfs.mat[3][1])
{
  'pl':
    [
      'html',
      'body',
      'div',
      'div'
    ],
  'samepl_sibseq': 1,
  'samepl_breadth': 1,
  'tag': 'div',
  'sibseq': 1,
  'attrib':
    {
      'id': 'navfirst'
    },
  'text': '\r\n',
  'tail': '\r\n\r\n',
  'text_intag': '\r\n\r\n',
  'node': <Elementdivat0x155fef50388>
}
>>>
>>>
>>>
```

3.2.4 width-first-traverse

```
html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
pls = engine.wfspls(root)
utils.parr(pls[:10])
```

```
>>>
>>> html_str = fs.rfile("./test.html")
>>> root = LXHTML(html_str)
>>> pls = engine.wfspls(root)
>>> utils.parr(pls[:10])
['html']
['html', 'head']
['html', 'body']
['html', 'head', '<cyfunction Comment at 0x00000155FE772550>']
['html', 'head', 'script']
['html', 'head', 'script']
['html', 'head', 'meta']
['html', 'head', 'meta']
['html', 'head', 'meta']
['html', 'head', 'link']
>>>
>>>
>>>
```

3.2.5 depth-first-traverse

```
import lxml.sax
html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
dfs = engine.DFS()
lxml.sax.saxify(root, dfs)
utils.parr(dfs.pls[:5])
utils.parr(dfs.pls[-10:])
```

```

>>> utils.parr(dfs.pls[:5])
['html', 'head', 'script']
['html', 'head', 'script']
['html', 'head', 'meta']
['html', 'head', 'meta']
['html', 'head', 'meta']
>>> utils.parr(dfs.pls[-10:])
['html', 'body', 'div', 'div', 'div', 'script']
['html', 'body', 'div', 'div', 'div']
['html', 'body', 'div', 'div']
['html', 'body', 'div', 'div', 'p']
['html', 'body', 'div', 'div', 'p', 'a']
['html', 'body', 'div', 'div', 'p']
['html', 'body', 'div', 'div']
['html', 'body', 'div']
['html', 'body']
['html']
>>>
>>>

```

3.2.6 beautify

```

html_str = fs.rfile("./test.html")
root = LXHTML(html_str)
html_str = engine.beautify(root)
print(html_str[:480])

```

```

>>> import lxml.sax
>>> html_str = fs.rfile("./test.html")
>>> root = LXHTML(html_str)
>>> html_str = engine.beautify(root)
>>> print(html_str[:480])
<html lang="zh-cn">
  <head>
    <script async="" src="https://www.googletagmanager.com/gtag/js?id=UA-878633-1">
    </script>
    <script>
      window.dataLayer = window.dataLayer || [];
    </script>
    <meta charset="gbk">
    </meta>
    <meta name="robots" content="all">
    </meta>
    <meta name="author" content="w3school.com.cn">
    </meta>
    <link rel="stylesheet" type="text/css" href="/c5_20171220.
css">
  </head>
  <body>
  </body>
</html>

```

3.3 lvsrch

```

[
'a',
'abbr',
'acronym',
'address',
'applet',
'area',
'arguments',
'article',
'aside',
'audio',
'b',
'base',
'basefont',
'bdi',
'bdo',
'big',
'blockquote',
'body',
'br',
'button',
'canvas',
'caption',

```

(continues on next page)

(continued from previous page)

```
'center',
'cite',
'code',
'col',
'colgroup',
'command',
'datalist',
'dd',
'del_',
'details',
'dfn',
'dialog',
'dir',
'div',
'dl',
'dt',
'el1',
'em',
'embed',
'engine',
'fieldset',
'figcaption',
'figure',
'font',
'footer',
'form',
'frame',
'frameset',
'h1',
'h2',
'h3',
'h4',
'h5',
'h6',
'head',
'header',
'hr',
'html',
'i',
'iframe',
'img',

'ins',
'isindex',
'kbd',
'keygen',
'label',
'legend',
- 'link',
'map',
'mark',
'menu',
'menuitem',
'meta',
'meter',
'nav',

```

(continues on next page)

(continued from previous page)

```
'noframes',
'noscript',
'object',
'ol',
'optgroup',
'option',
'output',
'p',
'param',
'pre',
'progress',
'q',
'rp',
'rt',
'ruby',
's',
'samp',
'script',
'section',
'select',
'small',
'source',
'span',
'srch',
'strike',
'strong',
'style',
'sub',
'summary',
'sup',
'table',
'tbody',
'td',
'textarea',
'tfoot',
'th',
'thead',
'time',
'title',
'tr',
'track',
'tt',
'u',
'ul',
'utils',
'var',
'video',
'wbr',
'xmp'
]
```


3.4 engine

```
[
'BEAUTIFY',
'ContentHandler',
'DFS',
'WFS',
'ancestor',
'ancestors',
'beautify',
'between_levels_cond_func',
'breadth',
'child',
'children',
'copy',
'default_wfs_handler',
'depth',
'descendants',
'descendants_pls',
'dfs_traverse',
'dfspls',
'disconnect',
'elcl',
'extract_pls',
'following_sibs',
'grand_parent',
'html',
'init_cls_wfs_arguments',
'is_leaf',
'layer',
'layer_wfs_handler',
'lcin',
'leaf_descendants',
'leaf_descendants_pls',
'loc',
'loc2node',
'lsib',
'lxe',
'lxml',
'nonleaf_descendants',
'nonleaf_descendants_pls',
'parent',
'pathlist',
'plget',
'preceding_sibs',
'rcin',
're',
'rootnode',
'rsib',
'samepl_breadth',
'samepl_siblings',
'samepl_sibseq',
'siblings',
'sibseq',
'source',
'text_intag',
'txtize',
```

(continues on next page)

(continued from previous page)

```
'utils',  
'wfs_traverse',  
'wfspls',  
'xpath',  
'xpath_levels'  
]
```

CHAPTER 4

Package API

```
from lxml.etree import HTML as LXHTML
from lxml.etree import XML as LXML
import lxml.sax
from xdict.jprint import pdir, pobj

from nvhtml import txt
from nvhtml import lvsrch
from nvhtml import fs
from nvhtml import engine
from nvhtml import utils
```

4.1 engine

4.2 lvsrch

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given. You can contribute in many ways:

5.1 Types of Contributions

5.1.1 Report Bugs

Report bugs at <https://github.com/ihgazni2/nvhtml/issues>

5.1.2 Fix Bugs

5.1.3 Implement Features

5.1.4 Write Documentation

5.1.5 Submit Feedback

The best way to send feedback is to file an issue at <https://github.com/ihgazni2/nvhtml/issues>

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.

5.2 Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. The pull request should include tests.

6.1 Developers

- ihgazni2
- <286264978@qq.com>
- <https://github.com/ihgazni2>

6.2 Contributors

- ihgazni
- <terryinzaghi@163.com>
- <https://github.com/ihgazni>
- logo

CHAPTER 7

History

7.1 0.0.1 (2018-12-18)

- first commit

CHAPTER 8

TODO

- relations
- relationship
- relation_distance
- preceding_relations
- following_relations
- cousins
- preceding_cousins
- following_cousins
- cousin
- grandsons
- grandson

CHAPTER 9

Indices and tables

- `genindex`
- `modindex`
- `search`