JUL 15, 2016

Beautiful Soup 4 Cheatsheet

Basic

```
# https://www.crummy.com/software/BeautifulSoup/bs4/doc/
from bs4 import BeautifulSoup
soup = BeautifulSoup(html_doc, 'html.parser')



soup.title # <title>The Dormouse's story</title>
soup.title.name # u'title'
soup.title.string # u'The Dormouse's story'
soup.title.parent.name # u'head'

#various finder
css_soup.select("p.strikeout.body") # css finder
soup.p # <p class="title"><b>The Dormouse's story</b></p>
soup.p['class'] # u'title'
soup.a # <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>
soup.find_all('a') # [<a ..>, ..]
soup.find(id="link3") # <a class="sister" href="http://example.com/tillie" id="link3">Tillie</a>
```

```
for link in soup.find_all('a'):
    print(link.get('href')) # http://example.com/elsi, # http://example.com/lacie
```

Make soup

```

from bs4 import BeautifulSoup


soup = BeautifulSoup(open("index.html"))
soup = BeautifulSoup("<html>data</html>")
```

Output

```

# HTML
soup.prettify() #pretty print
str(soup) # non-pretty print

# String
soup.get_text() #all text under the element
```

Search

```
soup.select('a[href="http://example.com/elsie"]') # exact attribute
soup.select('a[href^="http://example.com/"]') # negative match
soup.select('a[href$="tillie"]') # end match
soup.select('a[href*=".com/el"]') # middle match
```

```python
#-------------------------
# basic
#-------------------------
soup.find_all('b') # match by tag
soup.find_all(re.compile("^b")) # match by tag using regex
soup.find_all(["a", "b"]) # match by tag in list

# function (complex condition)
def has_class_but_no_id(tag):
  return tag.has_attr('class') and not tag.has_attr('id')
soup.find_all(has_class_but_no_id)



#-------------------------
# find_all_api
#-------------------------
find_all(name, attrs, recursive, string, limit, **kwargs)

soup.find_all("title") # tag condition
soup.find_all("p", "title") # tag and attr
# [<p class="title"><b>The Dormouse's story</b></p>]
soup.find_all("a")

# keyword arguments
soup.find_all(id="link2")
soup.find_all(href=re.compile("elsie"), id='link1')
soup.find(string=re.compile("sisters")) # text contain sisters

# css class (class is researved keyword)
```

Navigation

```
#----------------------------
# going up/down/side
#----------------------------
# ----- going down -----
soup.head# <head><title>The Dormouse's story</title></head>
soup.title# <title>The Dormouse's story</title>
soup.body.b # <b>The Dormouse's story</b>
soup.a # <a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>
soup.find_all('a')
# [<a class="sister" href="http://example.com/elsie" id="link1">Elsie</a>,
# <a class="sister" href="http://example.com/lacie" id="link2">Lacie</a>,
# <a class="sister" href="http:

# children = contents
head_tag.contents # [<title>The Dormouse's story</title>]
head_tag.children # [<title>The Dormouse's story</title>]

# descendants (all of a tag’s children, recursively)
for child in head_tag.descendants:
  print(child)

# .string is tricky
head_tag.contents # [<title>The Dormouse's story</title>]
head_tag.string # u'The Dormouse's story' (because head tag has only one child)
print(soup.html.string) # None (because html has many children)

# whitespace removed strings
```

```
for string in soup.stripped_strings:
  print(repr(string))



# ----- going up -----
title_tag.parent # <head><title>The Dormouse's story</title></head>
# going up recursively
```

Edit

```

#---------------------------
# change exisitng tag
#---------------------------
tag.name = "blockquote" # modify tag name
tag['class'] = 'verybold' # modify tag attribute
del tag['class'] # delete attribute
tag.string= 'not too bold' # modify tag contents string
tag.append(" but bolder than usual") # append tag contents

#---------------------------
# insert tag
#---------------------------
new_tag = soup.new_tag("a", href="http://www.example.com")
original_tag.append(new_tag) # create child
new_tag.string = "Link text." # can edit element after creating child

soup.b.string.insert_before(tag)
soup.b.i.insert_after(soup.new_string(" ever "))

#---------------------------
```

```
# delete tag
#----------------------------
soup.i.clear() # removes the contents
i_tag = soup.i.extract() # completely removes a tag from tree and returns the element
soup.i.decompose() # completely removes a tag from tree and discard the tag

#----------------------------
# replace/wrap/unwrap tag
#----------------------------
a_tag.i.replace_with(soup.new_tag("b"))
a_tag.i.replace_with(Beautifulsoup("<b>bold element</b>")) # replace inner html
soup.p.string.wrap(soup.new_tag("b"))
a_tag.i.unwrap()
```

Encoding

```

#output
soup.prettify("latin-1")
tag.encode("utf-8")
tag.encode("latin-1")
tag.encode("ascii")
```

Parse only part

```

# The SoupStrainer class allows you to choose which parts of an
# incoming document are parsed
from bs4 import SoupStrainer
```

```python
# conditions
only_a_tags = SoupStrainer("a")
only_tags_with_id_link2 = SoupStrainer(id="link2")

def is_short_string(string):
  return len(string) < 10
only_short_strings = SoupStrainer(string=is_short_string)

# execute parse
BeautifulSoup(html_doc, "html.parser", parse_only=only_a_tags)
BeautifulSoup(html_doc, "html.parser", parse_only=only_tags_with_id_link2)
BeautifulSoup(html_doc, "html.parser", parse_only=only_short_strings)
```

The link to these cheatsheet can be found here.

If you liked this article and think others should read it, please share it on Twitter or Facebook.

Comments

1 Comment Blog 1 Login

Recommend 6 Tweet Share Sort by Best

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Deepak rai • 5 months ago

How can I get the following html element as I am getting it using js with beautifulsoup in python--
document.querySelector('body > table > tbody > tr > td > table > tbody > tr:nth-child(2) > td:nth-child(2) > div:nth-child(3) > table > tbody > tr:nth-child(11) > td > table > tbody > tr:nth-child(4) > td:nth-child(5) > input[type="hidden"]:nth-child(1)').getAttribute("name")

• Reply • Share ›

Subscribe Add Disqus to your siteAdd DisqusAdd Disqus' Privacy PolicyPrivacy PolicyPrivacy