Python BeautifulSoup

```
# index.html
<!DOCTYPE html>
<html>
    <head>
        <title>Header</title>
        <meta charset="utf-8">
    </head>

    <body>
        <h2>Operating systems</h2>

        <ul id="mylist" style="width:150px">
            <li>Solaris</li>
            <li>FreeBSD</li>
            <li>Debian</li>
            <li>NetBSD</li>
            <li>Windows</li>
        </ul>

        <p>
          FreeBSD is an advanced computer operating system used to
          power modern servers, desktops, and embedded platforms.
        </p>

        <p>
          Debian is a Unix-like computer operating system that is
          composed entirely of free software.
        </p>

    </body>
</html>

##############################################################
# simple.py

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()
    soup = BeautifulSoup(contents, 'lxml')

    print(soup.h2)
    print(soup.head)
    print(soup.li)

####################################################################
tags_names.py

#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    print("HTML: {0}, name: {1}, text: {2}".format(soup.h2,
        soup.h2.name, soup.h2.text))
################################################################
traverse_tree.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    for child in soup.recursiveChildGenerator():

        if child.name:

            print(child.name)
########################################################################
#
```

```
get_children.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    root = soup.html

    root_childs = [e.name for e in root.children if e.name is not None]
    print(root_childs)


#####################################################################
get_descendants.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    root = soup.body

    root_childs = [e.name for e in root.descendants if e.name is not None]
    print(root_childs)
#####################################################################
scraping.py
#!/usr/bin/python3

from bs4 import BeautifulSoup
import requests as req

resp = req.get("http://www.something.com")
```

```
soup = BeautifulSoup(resp.text, 'lxml')

print(soup.title)
print(soup.title.text)
print(soup.title.parent)


##############################################
prettify.py

#!/usr/bin/python3

from bs4 import BeautifulSoup
import requests as req

resp = req.get("http://www.something.com")

soup = BeautifulSoup(resp.text, 'lxml')

print(soup.prettify())

#################################################
find_by_id.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    #print(soup.find("ul", attrs={ "id" : "mylist"}))
    print(soup.find("ul", id="mylist"))

###################################################
regex.py
#!/usr/bin/python3
```

```
import re

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    strings = soup.find_all(string=re.compile('BSD'))

    for txt in strings:

        print(" ".join(txt.split()))

################################################
select_nth_tag.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')

    print(soup.select("li:nth-of-type(3)"))
############################################################
select_by_id.py
#!/usr/bin/python3

from bs4 import BeautifulSoup

with open("index.html", "r") as f:

    contents = f.read()

    soup = BeautifulSoup(contents, 'lxml')
```

```
print(soup.select_one("#mylist"))
```