Web Scraping - Parse part of the document

August 19, 2022

[1]:
```
#import the required library
from bs4 import BeautifulSoup
```

[2]:
```
#sample web document from www.simplilearn.com website
data_SL = """<ul class="content-col_discover">
        <h5>Discover</h5>
        <li><a href="/resources" id="free_resources">Free resources</a></li>
        <li><a href="http://community.simplilearn.com/"␣
 ↪id="community">Simplilearn community</a></li>
        <li><a href="/career-data-labs" id="lab">Career data labs</a></li>
        <li><a href="/scholarships-for-veterans" id="scholarship">Veterans␣
 ↪scholarship</a></li>
        <li><a href="http://www.simplilearn.com/feed/" id="rss">RSS feed</a></
 ↪li>
    </ul>"""
```

[3]:
```
soup=BeautifulSoup(data_SL,"html.parser")
```

[4]:
```
print(soup.get_text()) # retrive only the text data in html doc
```

```
Discover
Free resources
Simplilearn community
Career data labs
Veterans scholarship
RSS feed
```

[5]:
```
# part of the parser in the webdocument

#SoupStrainer--------->used to parse onlt part of the data using Id

from bs4 import SoupStrainer

tagcomm=SoupStrainer(id=["community","scholarship"])
```

[6]:
```
#print the value
soup1=BeautifulSoup(data_SL,"html.parser",parse_only=tagcomm)
print(soup1)
```

```
<a href="http://community.simplilearn.com/" id="community">Simplilearn
community</a><a href="/scholarships-for-veterans" id="scholarship">Veterans
scholarship</a>
```

[7]:
```
# prettify
print(soup1.prettify())
```

```
<a href="http://community.simplilearn.com/" id="community">
 Simplilearn community
</a>
<a href="/scholarships-for-veterans" id="scholarship">
 Veterans scholarship
</a>
```

[ ]: