

Forum:[Hack Forums](#)

Thread: [Beautiful soup](#)

SIM-574702 - September 9, 2017, 10:03 am

I'm trying to create a web crawler and I need to get links from a website. First I planned to use regex but I found lots of comments about BS and making everything easier. Currently I'm stuck because website I want to parse has more complicated structure than tutorials and I can't find a good way to get only the links I'm interested in.

The webpage I'm trying to parse: [HERE](#)

What's making it hard is the fact that I don't want to get links from premium ads (they have yellow background). To go around that I've used title at top of regular ads.

Also links don't start with http but just /rel/link

```
Code from bs4 import BeautifulSoup
import requests
import re
```

```
url = 'http://www.njuskalo.hr/prijenosna-racunala/'
page = requests.get(url)
prefix = "http://www.njuskalo.hr"
soup = BeautifulSoup(page.content, 'html.parser')
```

```
#title = soup.find('title').text
```

```
letters = soup.find_all("h3", class_="EntityList-groupTitle")
```

```
lista = letters[2].next_sibling
lista = lista.next_sibling
print(lista)
```

This returns the part of website I'm interested in but it's type is `<class 'bs4.element.Tag'>` and I have no idea what to do next. I guess I'm missing something really simple that would extract links from that but any solution is welcome. I would try to find regex to extract it but that would cause lots of duplicates because link is on img and title.

vorticalbox - September 10, 2017, 6:35 am

PHP Code

```
<?php
from bs4 import BeautifulSoup as bs
import requests
```

```
url = 'http://www.njuskalo.hr/prijenosna-racunala'
with requests.Session() as s:
    r = s.get(url)
    soup = bs(r.text, 'html.parser')
    letters = soup.find_all("h3", class_="entity-subitem-title")
```

```
for item in letters:
    for link in item.find_all('a'):
        print('http://www.njuskalo.hr+link.get("href"))
```

Gives

PHP Code

```
<?php
```

```
http:
//www.njuskalo.hr/asus-prijenosnici/asus-notebook-asus-x556uq-nov-jamstvo-r-1-dostava-oglas-22763632
http:
//www.njuskalo.hr/hp-prijenosnici/hp-x2-210-g2-odvojiva-tipkovnica-win10-pro-128gb-ssd-touchscreen-oglas-22935904
http:
//www.njuskalo.hr/hp-prijenosnici/hp-x2-210-g2-odvojiva-tipkovnica-win10-home-128gb-ssd-touchscreen-oglas-22935933
http:
//www.njuskalo.hr/acer-prijenosnici/acer-aspire-es1-15.6-hd-intel-pentium-quad-8gb-1tb-geforce-920mx-oglas-23382969
http:
//www.njuskalo.hr/asus-prijenosnici/asus-k550vx-15.6-fhd-intel-i7-8gb-256gb-ssd-gtx950m-4gb-win10-oglas-23383893
http:
//www.njuskalo.hr/dell-prijenosnici/linija-dell-poslovnih-prijenosnika-dell-vostro-5568-oglas-23384051
http:
//www.njuskalo.hr/ostali-prijenosnici/microsoft-surfacebook-128gb-intel-core-i5-8gb-ram-izdavanje-r1-oglas-22590426
http:
//www.njuskalo.hr/dell-prijenosnici/dell-inspiron-15-5578-1777-2-in-1-15.6-full-hd-izdavanje-r1-oglas-22677476
http:
//www.njuskalo.hr/ostali-prijenosnici/microsoft-surface-laptop-128gb-intel-core-i5-4gb-izdavanje-r1-oglas-22828591
```

SIM-574702 - September 10, 2017, 10:29 pm

Exactly what I was looking for. I did the same thing with three times as much code.

vorticalbox - September 11, 2017, 1:45 am

(Sep 10, 2017 10:29 PM)matemistic Wrote: [\[CITE\]](#) Exactly what I was looking for. I did the same thing with three times as much code.

It just finding the right classes and tags

PHP Code

```
<?php
```

```
soup.find('a') #returns the first link
links = soup.find_all('a') # returns a list of all links
for link in links:
    print(link)#print the html tag
    print(link.get('href')) #print just the href
```