

Web Crawling

April 21, 2021

```
[72]: import urllib.request
from bs4 import BeautifulSoup

url = 'https://www.google.de/search?
↳q=coronavirus+bochum&sxsrf=ALeKk02BiGMdcH7JgfQCMGJvx17cQxxF1g:
↳1596055195309&source=lnms&tbm=wns&sa=X&ved=2ahUKEwi9-dT3qPPqAhWjsaQKHUBRD_cQ_AUoAXoECB8QAw&
req = urllib.request.Request(url, headers = {'User-Agent': 'Mozilla/5.0 (X11;
↳Linux x86_64) AppleWebKit/537.11 (KHTML, like Gecko) Chrome/23.0.1271.64
↳Safari/537.11',
      'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.
↳8',
      'Accept-Charset': 'ISO-8859-1,utf-8;q=0.7,*;q=0.3',
      'Accept-Encoding': 'none',
      'Accept-Language': 'en-US,en;q=0.8',
      'Connection': 'keep-alive'})
html = urllib.request.urlopen(req).read()
soup = BeautifulSoup(html,'html.parser')

titles = soup.find_all(class_ = 'JheGif nDgy9d')
urls = soup.find_all(class_ = 'dbsr')

url_list = []
title_list= []

for url in urls :
    url_list.append(url.find("a")['href'])

for title in titles:
    title_list.append(title.text)

for i in range(len(url_list)) :
    print(title_list[i])
    print(url_list[i])
    print()
```

Bochum: Corona-Ausbruch in Altenheim ist schlimmer als befürchtet

<https://www.waz.de/staedte/bochum/bochum-corona-ausbruch-in-altenheim-ist-schlimmer-als-befuerchtet-id229617920.html>

Corona: Infektionszahlen in Bochum steigen weiter deutlich

<https://www.waz.de/staedte/bochum/corona-infektionszahlen-in-bochum-steigen-weiter-deutlich-id229596492.html>

Die Besorgnis der Bürger wächst: Corona Neuerkrankungen :Bochum auf Platz 3 in NRW - Bochum

https://www.lokalkompass.de/bochum/c-ratgeber/corona-neuerkrankungen-bochum-auf-platz-3-in-nrw_a1410507

Weiter angespannte Corona-Lage in Bochum

<https://www.radiobochem.de/artikel/weiter-angespannte-corona-lage-in-bochem-666083.html>

Steigende Fallzahlen: Drive-In-Corona-Teststelle wieder auf

<https://www.sueddeutsche.de/gesundheit/gesundheit-bochem-steigende-fallzahlen-drive-in-corona-teststelle-wieder-auf-dpa.urn-newsml-dpa-com-20090101-200728-99-951492>

Bochumer Coronatest-Drive-In öffnet wieder - Ruhrgebiet - Nachrichten

<https://www1.wdr.de/nachrichten/ruhrgebiet/bochumer-corona-drive-in-wieder-geoeffnet-100.html>

Coronavirus News: Covid -19 legt Polizeiwache lahm - Bochum

https://www.lokalkompass.de/bochem/c-blaulicht/covid-19-legt-polizeiwache-lahm_a1410450

Corona: In Bochum wächst die Sorge vor neuen Infektionen

<https://www.waz.de/staedte/bochem/corona-in-bochem-waechst-die-sorge-vor-neuen-infektionen-id229612572.html>

Corona Bochum: Steigende Zahlen am Wochenende

<https://www.radiobochem.de/artikel/corona-in-bochem-auch-am-wochenende-steigende-zahlen-663812.html>

Aktiv gegen Brustkrebs: Yoga-Spaziergang in Corona-Zeiten für Brustkrebspatientinnen

https://www.lokalkompass.de/bochem/c-ratgeber/yoga-spaziergang-in-corona-zeiten-fuer-brustkrebspatientinnen_a1410306

```
[21]: from urllib.parse import quote_plus
      from bs4 import BeautifulSoup
      from selenium import webdriver
```

```

base = 'https://www.google.de/search?q='
plusurl = input('      ? : ')
n = input('      ? : ')
if(int(n) == 1): num = 1
elif(int(n) > 1) : num = (int(n)-1)*10

page = "&start="
url = baseurl + plusurl+ page + str(num)

driver = webdriver.Chrome()
driver.get(url)

html = driver.page_source
soup = BeautifulSoup(html)

r = soup.select('.r')
for i in r :
    print(i.select_one("a")['href'])
    print(i.select_one("h3").text)
    print()

```

```

      ? : corona bochum
      ? : 3
https://www.lg-
bochum.nrw.de/beh_toppmeldungen_archiv/Toppmeldung_02_Corona/index.php
Landgericht Bochum: CORONA

https://www.augusta-kliniken.de/pflege-betreuung/informationen-zum-coronavirus-
sars-cov-2.html
Informationen zum Coronavirus (SARS-CoV-2) - Augusta ...

https://www.starlight-express.de/coronavirus.html
Informationen zum Coronavirus - STARLIGHT EXPRESS ...

https://www.hno-bochum-witten.de/aktuell/wichtige-informationen-zum-
coronavirus.php
Wichtige Informationen zum Coronavirusaktuelle meldungen ...

https://www.sueddeutsche.de/sport/fussball-bochum-corona-positiv-blum-fehlt-
bochum-zum-start-der-vorbereitung-dpa.urn-newsml-dpa-
com-20090101-200803-99-26296
Fußball - Bochum - Corona-positiv: Blum fehlt Bochum zum ...

https://www.nrz.de/region/rhein-und-ruhr/coronavirus-stadt-meldet-die-erste-
infektion-in-bochum-id228613971.html
Coronavirus in Bochum: Anrufe bei Corona-Hotline auf ... - Nrz

```

<https://news.google.de/news/section?hl=de&q=Bochum>
Suchen - Google News

https://www.planetarium-bochum.de/de_DE/corona
Informationen zu COVID-19 (Coronavirus) - Planetarium ...

<https://www.usb-bochum.de/presse/aktuellen/>
Aktuelles - USB Bochum GmbH

<https://else-hirsch-schule.de/umgang-mit-dem-corona-virus-an-schulen-info-stadt-bochum/>
Umgang mit dem Corona-Virus an Schulen - Info Stadt Bochum

```
[ ]: from urllib.request import urlopen
from urllib.parse import quote_plus
from bs4 import BeautifulSoup

baseUrl = 'https://search.naver.com/search.naver?
↳sm=top_h ty&fbm=0&ie=utf8&query='
plusUrl = input('      .')
url = baseUrl+quote_plus(plusUrl)

html = urlopen(url).read()
soup = BeautifulSoup(html, 'html.parser')

img = soup.find_all(class_ = '_img')
n=1
#for i in img :
#    imgUrl = i['data-source']
#    with urlopen(imgUrl) as f:
#        with open('./img'+plusUrl+str(n) + '.jpg', 'wb') as h:
#            img = f.read()
#            h.write(img)
#            n=n+1
print('Download Complete')
```

```
[ ]: import csv
from urllib.request import urlopen
from urllib.request import quote_plus
from bs4 import BeautifulSoup

search = input('      ')
url = f'https://search.naver.com/search.naver?
↳sm=top_h ty&fbm=0&ie=utf8&query={quote_plus(search)}'
```

```

html = urlopen(url).read()
soup = BeautifulSoup(html, 'html.parser')

total = soup.select('api_txt_lines.total_tit')
searchlist = []

for i in total:
    temp = []
    temp.append(i.text)
    temp.append(i.attrs['href'])
    searchlist.append(temp)

f = open(f'{search}.csv', 'w', encoding = 'utf-8', newline = '')
csvWriter = csv.writer(f)
for i in searchlist :
    csvWriter.writerow(i)

f.close
print('complete')

```