›lished in Towards Data Science

›onghao Wu
ul 15, 2020 · 9 min read · ✦ Member-only · Listen

Scraping Basics

scrape data from a website in Python

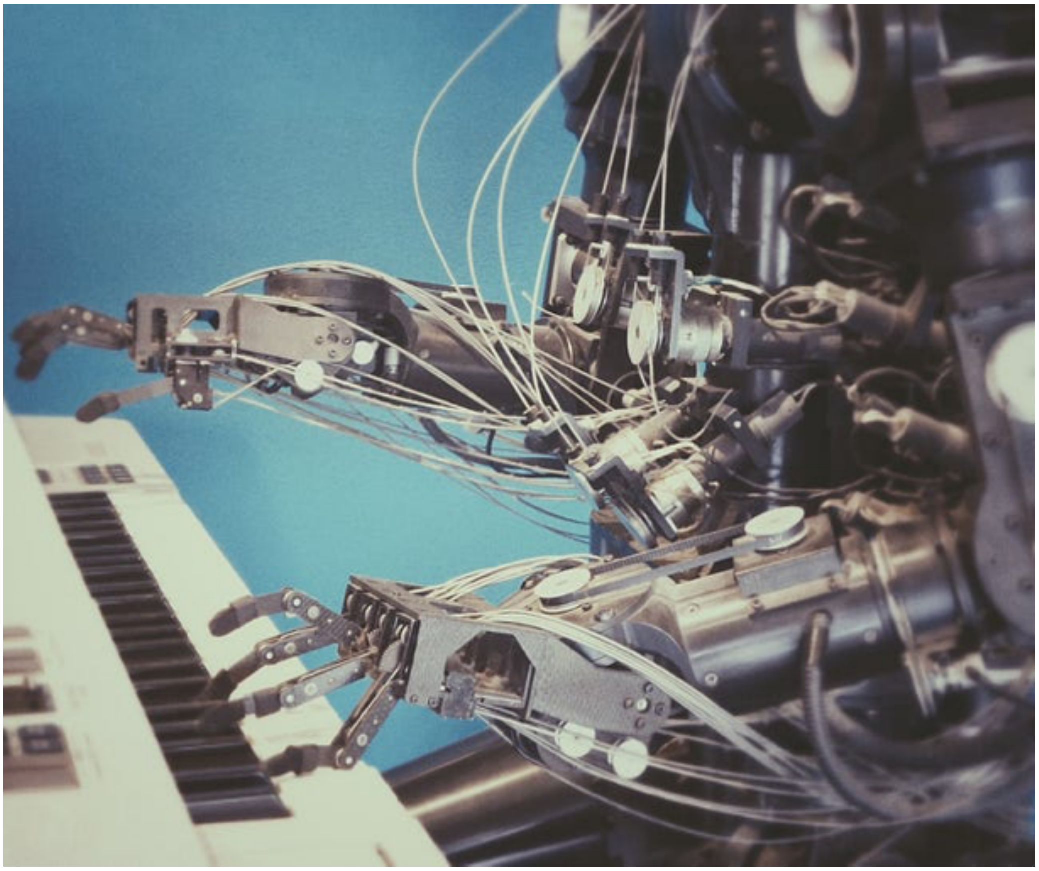


Photo by Franck V from Unsplash

ıys say “Garbage in Garbage out” in data science. If you do not have ality and quantity of data, most likely you would not get many out of it. Web Scraping is one of the important methods to retrieve ırty data automatically. In this article, I will be covering the basics of aping and use two examples to illustrate the 2 different ways to do it ɔn.

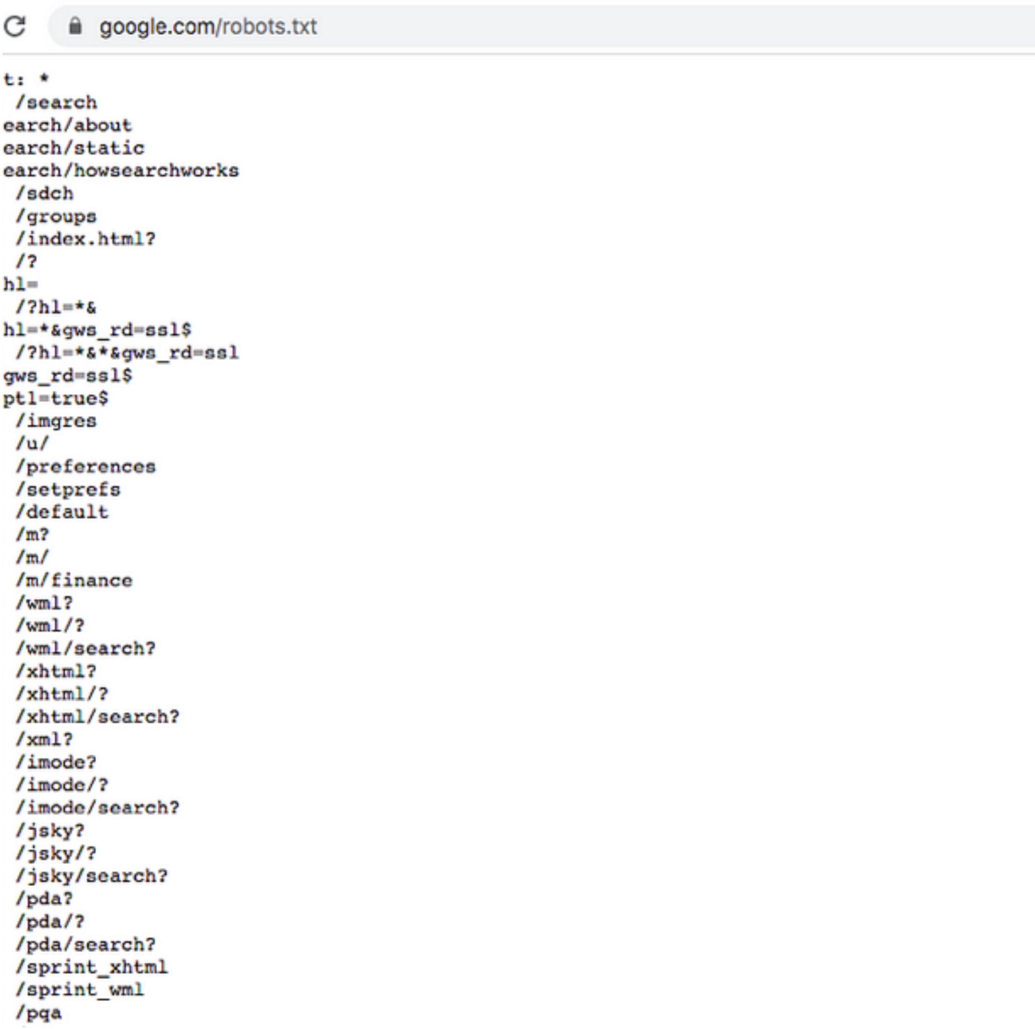
Web Scraping

ɾaping is an automatic way to retrieve unstructured data from a and store them in a structured format. For example, if you want to what kind of face mask can sell better in Singapore, you may want to ıll the face mask information on an E-Commerce website like Lazada.

ı scrape from all the websites?

ʒ makes the website traffic spike and may cause the breakdown of site server. Thus, not all websites allow people to scrape. How do you hich websites are allowed or not? You can look at the ‘robots.txt’ file ɛbsite. You just simply put robots.txt after the URL that you want to ınd you will see information on whether the website host allows you e the website.

ɔogle.com for an example



```
t: *
 /search
earch/about
earch/static
earch/howsearchworks
 /sdch
 /groups
 /index.html?
 /?
hl=
 /?hl=*&
hl=*&gws_rd=ssl$
 /?hl=*&*&gws_rd=ssl
gws_rd=ssl$
ptl=true$
 /imgres
 /u/
 /preferences
 /setprefs
 /default
 /m?
 /m/
 /m/finance
 /wml?
 /wml/?
 /wml/search?
 /xhtml?
 /xhtml/?
 /xhtml/search?
 /xml?
 /imode?
 /imode/?
 /imode/search?
 /jsky?
 /jsky/?
 /jsky/search?
 /pda?
 /pda/?
 /pda/search?
 /sprint_xhtml
 /sprint_wml
 /pqa
```

robots.txt file of Google.com

see that Google does not allow web scraping for many of its sub-
s. However, it allows certain paths like '/m/finance' and thus if you
collect information on finance then this is a completely legal place to

: note is that you can see from the first row on User-agent. Here
specifies the rules for all of the user-agents but the website may give
user-agent special permission so you may want to refer to
ttion there.

es web scraping work?

aping just works like a bot person browsing different pages website y pastedown all the contents. When you run the code, it will send a to the server and the data is contained in the response you get. What ı do is parse the response data and extract out the parts you want.

we do web scraping?

finally we are here. There are 2 different approaches for web ʒ depending on how does website structure their contents.

ch 1: *If website stores all their information on the HTML front end, directly use code to download the HTML contents and extract out ıformation.*

re roughly 5 steps as below:

ect the website HTML that you want to crawl

ɛss URL of the website using code and download all the HTML ents on the page

nat the downloaded content into a readable format

act out useful information and save it into a structured format

nformation displayed on multiple pages of the website, you may l to repeat steps 2–4 to have the complete information.

d Cons for this approach: It is simple and direct. However, if the s front-end structure changes then you need to adjust your code ıgly.

ch 2: *If website stores data in API and the website queries the API each ıen user visit the website, you can simulate the request and directly ata from the API* ect the XHR network section of the URL that you want to crawl

out the request-response that gives you the data that you want

ending on the type of request(post or get) and also the request header yload, simulate the request in your code and retrieve the data from Usually, the data got from API is in a pretty neat format.

act out useful information that you need

API with a limit on query size, you will need to use 'for loop' to atedly retrieve all the data

d Cons for this approach: It is definitely a preferred approach if you the API request. The data you receive will be more structured and This is because compared to the website front end, it is less likely for pany to change its backend API. However, it is a bit more cated than the first approach especially if authentication or token is d.

t tools and library for web scraping

re many different scraping tools available that do not require any However, most people still use the Python library to do web scraping it is easy to use and also you can find an answer in its big nity.

st commonly used library for web scraping in Python is **Beautiful equests, and Selenium**.

l Soup: It helps you parse the HTML or XML documents into a e format. It allows you to search different elements within the nts and help you retrieve required information faster.

s: It is a Python module in which you can send HTTP requests to contents. It helps you to access website HTML contents or API by Get or Post requests.

m: It is widely used for website testing and it allows you to automate t events(clicking, scrolling, etc) on the website to get the results you

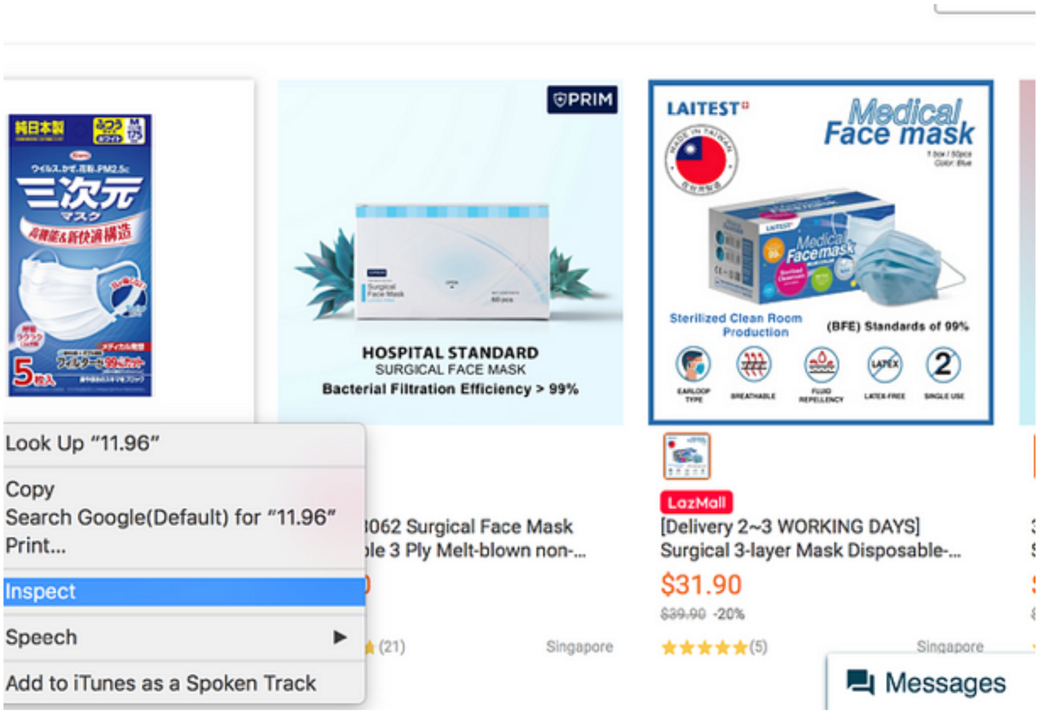
either use Requests + Beautiful Soup or Selenium to do web g. **Selenium is preferred if you need to interact with the (JavaScript events) and if not I will prefer Requests + Beautiful Soup e it's faster and easier.**

raping Example:

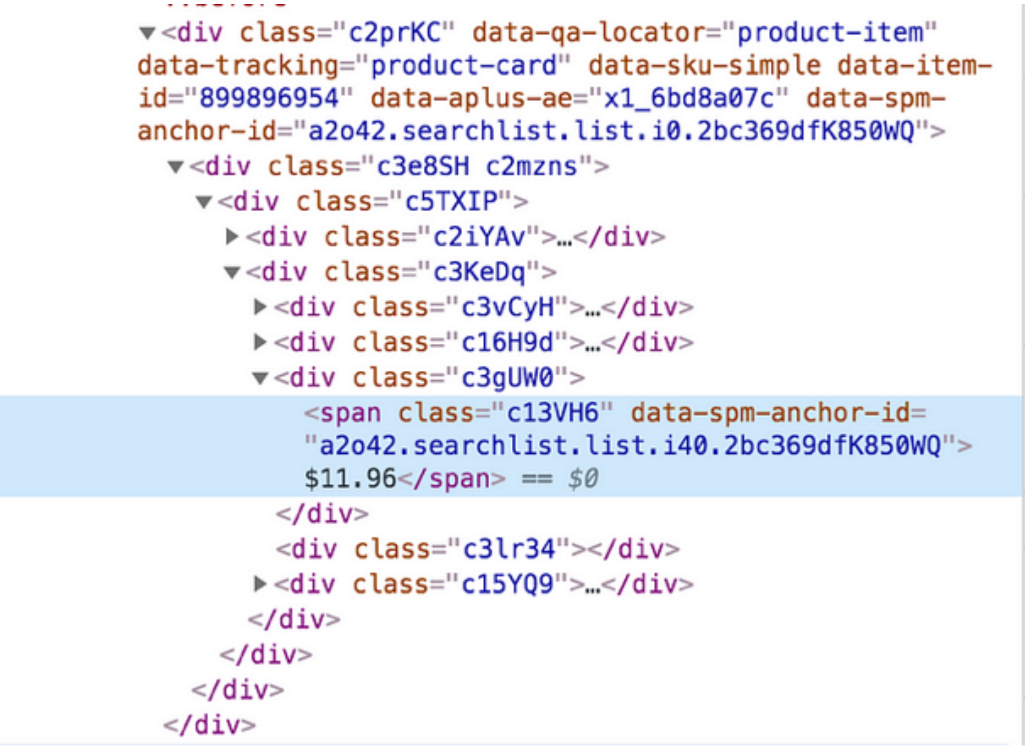
n statement: *I want to find out about the local market for face mask. I rested on online face mask price, discount, ratings, sold quantity etc.*

proach 1 Example(Download HTML for all pages) Lazada:

nspect the website(if using Chrome you can right-click and select)


Look Up "11.96"
Copy
Search Google(Default) for "11.96"
Print...
Inspect
Speech
Add to iTunes as a Spoken Track
HOSPITAL STANDARD
SURGICAL FACE MASK
Bacterial Filtration Efficiency > 99%
062 Surgical Face Mask
ble 3 Ply Melt-blown non-...
LAITEST
Medical Face mask
LazMall
[Delivery 2~3 WORKING DAYS]
Surgical 3-layer Mask Disposable-...
$31.90
$39.90 -20%
(21)
Singapore
(5)
Singapore
Messages

Inspect Lazada page on Chrome



```
▼<div class="c2prKC" data-qa-locator="product-item"
data-tracking="product-card" data-sku-simple data-item-
id="899896954" data-aplus-ae="x1_6bd8a07c" data-spm-
anchor-id="a2o42.searchlist.list.i0.2bc369dfK850WQ">
 ▼<div class="c3e8SH c2mzns">
  ▼<div class="c5TXIP">
   ▶<div class="c2iYAv">…</div>
   ▼<div class="c3KeDq">
    ▶<div class="c3vCyH">…</div>
    ▶<div class="c16H9d">…</div>
    ▼<div class="c3gUW0">
       <span class="c13VH6" data-spm-anchor-id=
       "a2o42.searchlist.list.i40.2bc369dfK850WQ">
       $11.96</span> == $0
     </div>
     <div class="c3lr34"></div>
    ▶<div class="c15YQ9">…</div>
    </div>
   </div>
  </div>
 </div>
```

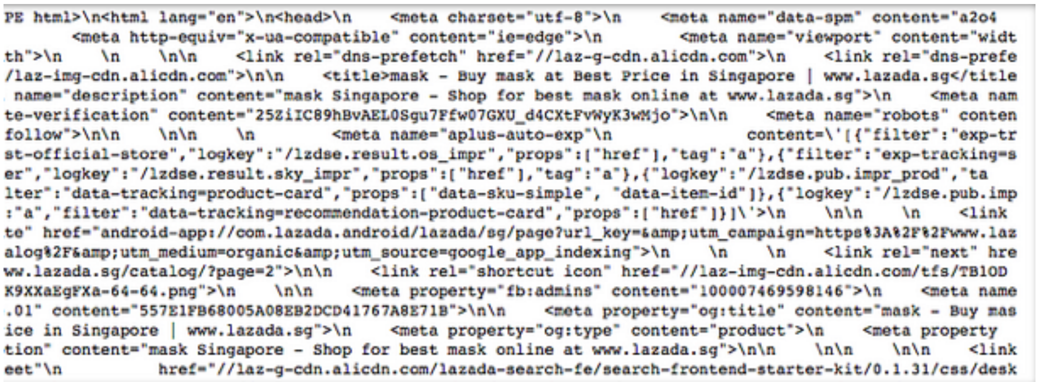
HTML result for price on Lazada

e that data I need are all wrap in the HTML element with the unique me.

ccess URL of the website using code and download all the HTML s on the page

```
ort library
os4 import BeautifulSoup
t requests

uest to website and download HTML contents
https://www.lazada.sg/catalog/?_keyori=ss&from=input&q=mask'
equests.get(url)
nt=req.text
```



```
PE html>\n<html lang="en">\n<head>\n    <meta charset="utf-8">\n    <meta name="data-spm" content="a2o4
       <meta http-equiv="x-ua-compatible" content="ie=edge">\n        <meta name="viewport" content="widt
th">\n    \n    \n\n    <link rel="dns-prefetch" href="//laz-g-cdn.alicdn.com">\n    <link rel="dns-prefe
/laz-img-cdn.alicdn.com">\n\n    <title>mask - Buy mask at Best Price in Singapore | www.lazada.sg</title
 name="description" content="mask Singapore - Shop for best mask online at www.lazada.sg">\n    <meta nam
te-verification" content="25ZiIC89hBvAEL0Sgu7Ffw07GXU_d4CXtFvWyK3wMjo">\n\n    <meta name="robots" conten
follow">\n\n    \n\n    \n        <meta name="aplus-auto-exp"\n            content=\'[{"filter":"exp-tr
st-official-store","logkey":"/lzdse.result.os_impr","props":["href"],"tag":"a"},{"filter":"exp-tracking=s
er","logkey":"/lzdse.result.sky_impr","props":["href"],"tag":"a"},{"logkey":"/lzdse.pub.impr_prod","ta
lter":"data-tracking=product-card","props":["data-sku-simple", "data-item-id"]},{"logkey":"/lzdse.pub.imp
:"a","filter":"data-tracking=recommendation-product-card","props":["href"]}]\'>\n    \n\n    \n    <link
te" href="android-app://com.lazada.android/lazada/sg/page?url_key=&amp;utm_campaign=https%3A%2F%2Fwww.laz
alog%2F&amp;utm_medium=organic&amp;utm_source=google_app_indexing">\n    \n    \n    <link rel="next" hre
ww.lazada.sg/catalog/?page=2">\n\n    <link rel="shortcut icon" href="//laz-img-cdn.alicdn.com/tfs/TB1OD
K9XXaEgFXa-64-64.png">\n    \n\n    <meta property="fb:admins" content="100007469598146">\n    <meta name
.01" content="557E1FB68005A08EB2DCD41767A8E71B">\n\n    <meta property="og:title" content="mask - Buy mas
ice in Singapore | www.lazada.sg">\n    <meta property="og:type" content="product">\n    <meta property
tion" content="mask Singapore - Shop for best mask online at www.lazada.sg">\n\n    \n\n    \n\n    <link
eet"\n         href="//laz-g-cdn.alicdn.com/lazada-search-fe/search-frontend-starter-kit/0.1.31/css/desk
```

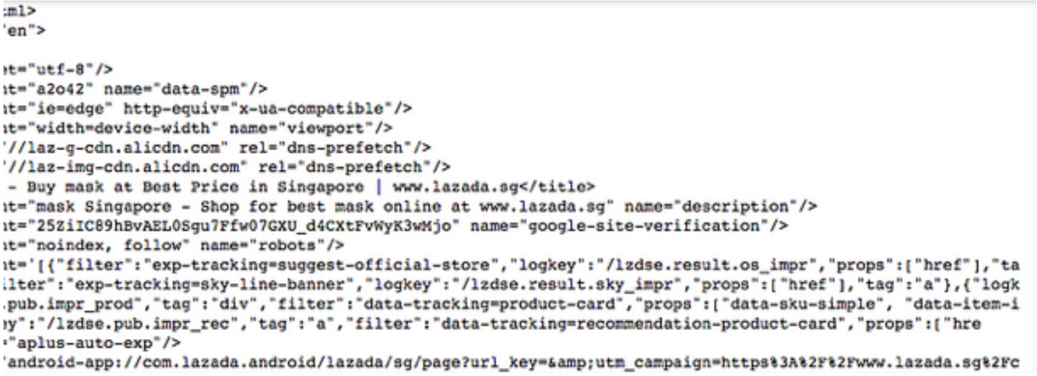
Request content before applying Beautiful Soup

ne requests library to get data from a website. You can see that so far have is unstructured text.

ormat the downloaded content into a readable format

```
BeautifulSoup(content)
```

p is very straightforward and what we do is just parse unstructured Beautiful Soup and what you get is as below.



```
:ml>
'en">

:t="utf-8"/>
:t="a2o42" name="data-spm"/>
:t="ie=edge" http-equiv="x-ua-compatible"/>
:t="width=device-width" name="viewport"/>
'//laz-g-cdn.alicdn.com" rel="dns-prefetch"/>
'//laz-img-cdn.alicdn.com" rel="dns-prefetch"/>
 - Buy mask at Best Price in Singapore | www.lazada.sg</title>
:t="mask Singapore - Shop for best mask online at www.lazada.sg" name="description"/>
:t="25ZiIC89hBvAEL0Sgu7Ffw07GXU_d4CXtFvWyK3wMjo" name="google-site-verification"/>
:t="noindex, follow" name="robots"/>
:t='[{"filter":"exp-tracking=suggest-official-store","logkey":"/lzdse.result.os_impr","props":["href"],"ta
.lter":"exp-tracking=sky-line-banner","logkey":"/lzdse.result.sky_impr","props":["href"],"tag":"a"},{"logk
.pub.impr_prod","tag":"div","filter":"data-tracking=product-card","props":["data-sku-simple", "data-item-i
:y":"/lzdse.pub.impr_rec","tag":"a","filter":"data-tracking=recommendation-product-card","props":["hre
="aplus-auto-exp"/>
'android-app://com.lazada.android/lazada/sg/page?url_key=&amp;utm_campaign=https%3A%2F%2Fwww.lazada.sg%2Fc
```

HTML content after using Beautiful Soup

put is a much more readable format and you can search different elements or classes in it.

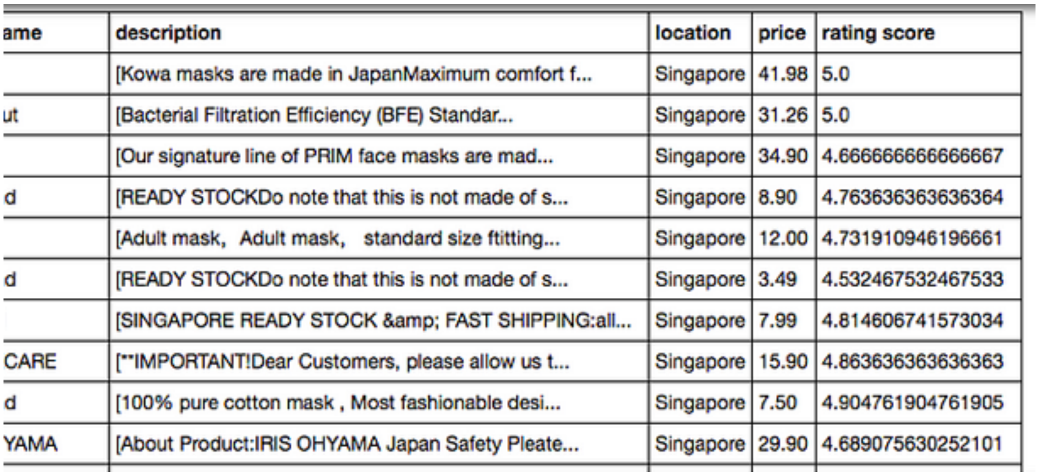
Extract out useful information and save it into a structured format

p requires some time to understand website structure and find out he data is stored exactly. For the Lazada case, it is stored in a Script in JSON format.

```
oup.findAll('script')[3].text
od.read_json(raw.split("window.pageData=")[1],orient='records')
e data
tem in page.loc['listItems','mods']:
rand_name.append(item['brandName'])
rice.append(item['price'])
ocation.append(item['location'])
escription.append(ifnull(item['description'],0))
ating_score.append(ifnull(item['ratingScore'],0))
```

d 5 different lists to store the different fields of data that I need. I e for loop here to loop through the list of items in the JSON nts inside. After that, I combine the 5 columns into the output file.

```
 data into an output
t=pd.DataFrame({'brandName':brand_name,'price':price,'location':
ion,'description':description,'rating score':rating_score})
```



ame	description	location	price	rating score
	[Kowa masks are made in JapanMaximum comfort f...	Singapore	41.98	5.0
ut	[Bacterial Filtration Efficiency (BFE) Standar...	Singapore	31.26	5.0
	[Our signature line of PRIM face masks are mad...	Singapore	34.90	4.666666666666667
d	[READY STOCKDo note that this is not made of s...	Singapore	8.90	4.763636363636364
	[Adult mask, Adult mask, standard size ftitting...	Singapore	12.00	4.731910946196661
d	[READY STOCKDo note that this is not made of s...	Singapore	3.49	4.532467532467533
	[SINGAPORE READY STOCK &amp; FAST SHIPPING:all...	Singapore	7.99	4.814606741573034
CARE	[**IMPORTANT!Dear Customers, please allow us t...	Singapore	15.90	4.863636363636363
d	[100% pure cotton mask , Most fashionable desi...	Singapore	7.50	4.904761904761905
YAMA	[About Product:IRIS OHYAMA Japan Safety Pleate...	Singapore	29.90	4.689075630252101

Final output in Python DataFrame format

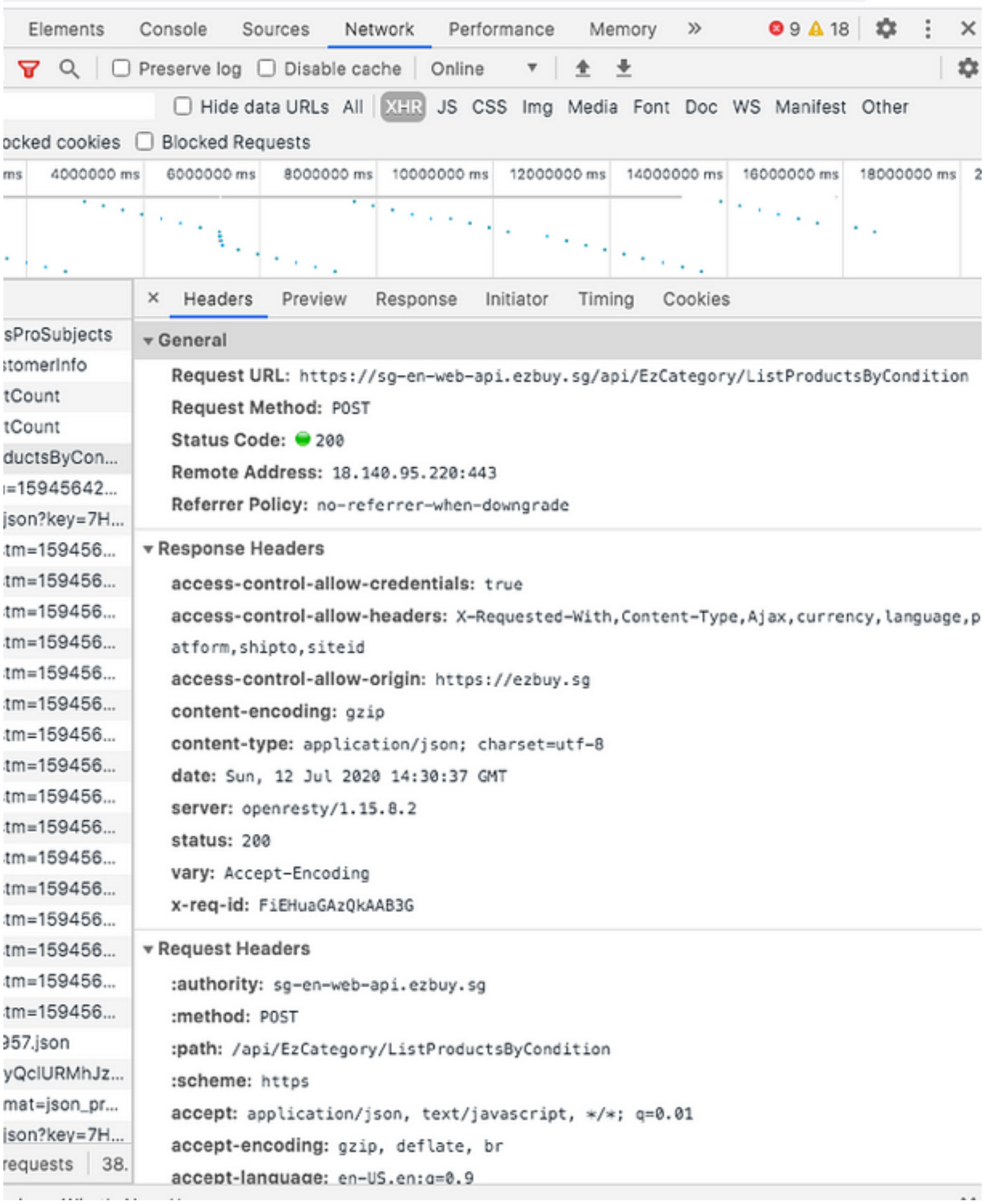
or information displayed on multiple pages of the website, you may repeat steps 2–4 to have the complete information.

ant to scrape all the data. Firstly you should find out about the total sellers. Then you should loop through pages by passing in ental page numbers using payload to URL. Below is the full code that scrape and I loop through the first 50 pages to get content on those

```
 in range(1,50):
ime.sleep(max(random.gauss(5,1),2))
rint('page'+str(i))
ayload['page']=i
eq=requests.get(url,params=payload)
ontent=req.text
oup=BeautifulSoup(content)
aw=soup.findAll('script')[3].text
age=pd.read_json(raw.split("window.pageData=")
rient='records')
or item in page.loc['listItems','mods']:
   brand_name.append(item['brandName'])
   price.append(item['price'])
   location.append(item['location'])
   description.append(ifnull(item['description'],0))
   rating_score.append(ifnull(item['ratingScore'],0))
```

ɔroach 2 example(Query data directly from API) — ɔuy:

ɩnspect the XHR network section of the URL that you want to crawl ɩ out the request-response that gives you the data that you want



```
sProSubjects
ıtomerInfo
tCount
tCount
ductsByCon...
ı=15945642...
json?key=7H...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɩtm=159456...
ɜ57.json
yQcIURMhJz...
mat=json_pr...
json?key=7H...
requests | 38.
```

```
▾ General
  Request URL: https://sg-en-web-api.ezbuy.sg/api/EzCategory/ListProductsByCondition
  Request Method: POST
  Status Code: ● 200
  Remote Address: 18.140.95.220:443
  Referrer Policy: no-referrer-when-downgrade
▾ Response Headers
  access-control-allow-credentials: true
  access-control-allow-headers: X-Requested-With,Content-Type,Ajax,currency,language,p
  atform,shipto,siteid
  access-control-allow-origin: https://ezbuy.sg
  content-encoding: gzip
  content-type: application/json; charset=utf-8
  date: Sun, 12 Jul 2020 14:30:37 GMT
  server: openresty/1.15.8.2
  status: 200
  vary: Accept-Encoding
  x-req-id: FiEHuaGAzQkAAB3G
▾ Request Headers
  :authority: sg-en-web-api.ezbuy.sg
  :method: POST
  :path: /api/EzCategory/ListProductsByCondition
  :scheme: https
  accept: application/json, text/javascript, */*; q=0.01
  accept-encoding: gzip, deflate, br
  accept-language: en-US,en;q=0.9
```

XHR section under Network — Product list API request and response

ɘ from the Network that all product information is listed in this API ist Product by Condition’. The response gives me all the data I need a POST request.

Depending on the type of request(post or get) and also the request & payload, simulate the request in your code and retrieve the data PI. Usually, the data got from API is in a pretty neat format.

```
uests.session()

he API url
earch='https://sg-en-web-
zbuy.sg/api/EzCategory/ListProductsByCondition'

he header for the post request
rs={'user-agent':'Mozilla/5.0 (Macintosh; Intel Mac OS X
_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.116
i/537.36'}

he payload for the request form
{
searchCondition":
   {"categoryId":0,"freeShippingType":0,"filter:
eyWords":"mask"},
   "limit":100,
   "offset":0,
   "language":"en",
   "dataType":"new"


.post(url_search,headers=headers,json=data)
```

reate the HTTP POST request using the requests library. For post , you need to define the request header(setting of the request) and (data you are sending with this post request). Sometimes token or ication is required here and you will need to request for token first ending your POST request. Here there is no need to retrieve the id usually just follow what’s in the request payload in Network and user-agent’ for the header.

thing to note here is that inside the payload, I specified limit as 100 et as 0 because I found out it only allows me to query 100 data rows me. Thus, what we can do later is to use for loop to change offset and ıore data points.

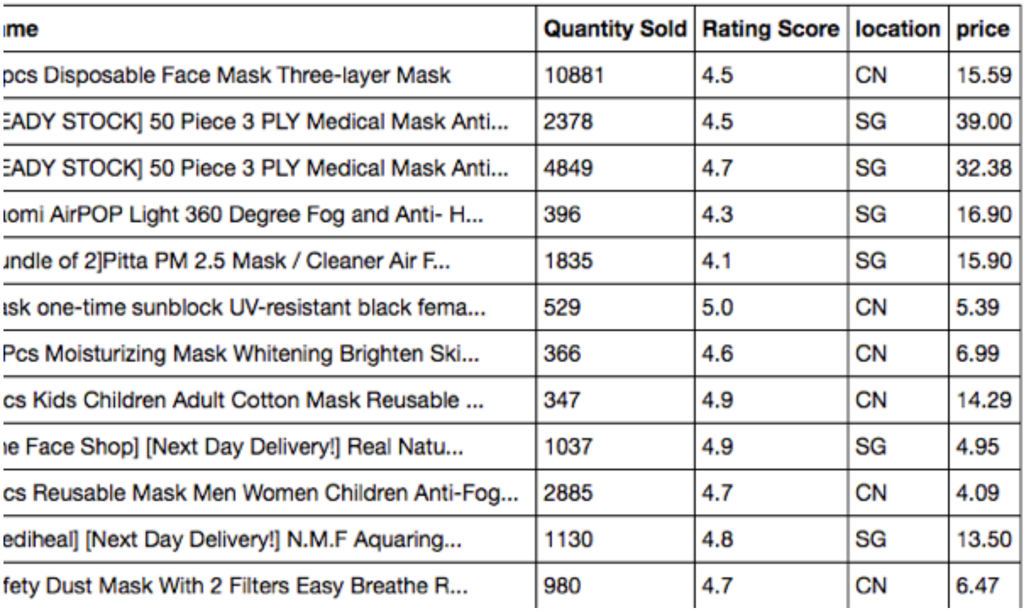
:xtract out useful information that you need

```
 the data back as json file
.json()

re data into the fields
tem in j['products']:
rice.append(item['price'])
ocation.append(item['originCode'])
ame.append(item['name'])
atingScore.append(item['leftView']['rateScore'])
uantity.append(item['rightView']['text'].split(' Sold')[0]

ine all the columns together
t=pd.DataFrame({'Name':name,'price':price,'location':location,'R
 Score':ratingScore,'Quantity Sold':quantity})
```

ım API is usually quite neat and structured and thus what I did was ead it in JSON format. After that, I extract the useful data into t columns and combine them together as output. You can see the :put below.



ıme	Quantity Sold	Rating Score	location	price
pcs Disposable Face Mask Three-layer Mask	10881	4.5	CN	15.59
EADY STOCK] 50 Piece 3 PLY Medical Mask Anti...	2378	4.5	SG	39.00
EADY STOCK] 50 Piece 3 PLY Medical Mask Anti...	4849	4.7	SG	32.38
ıomi AirPOP Light 360 Degree Fog and Anti- H...	396	4.3	SG	16.90
ındle of 2]Pitta PM 2.5 Mask / Cleaner Air F...	1835	4.1	SG	15.90
ısk one-time sunblock UV-resistant black fema...	529	5.0	CN	5.39
Pcs Moisturizing Mask Whitening Brighten Ski...	366	4.6	CN	6.99
cs Kids Children Adult Cotton Mask Reusable ...	347	4.9	CN	14.29
ıe Face Shop] [Next Day Delivery!] Real Natu...	1037	4.9	SG	4.95
cs Reusable Mask Men Women Children Anti-Fog...	2885	4.7	CN	4.09
ediheal] [Next Day Delivery!] N.M.F Aquaring...	1130	4.8	SG	13.50
fety Dust Mask With 2 Filters Easy Breathe R...	980	4.7	CN	6.47

EZbuy face mask data output

For API with a limit on query size, you will need to use 'for loop' to dly retrieve all the data

```
ne API url
earch='https://sg-en-web-
zbuy.sg/api/EzCategory/ListProductsByCondition'

ne header for the post request
rs={'user-agent':'Mozilla/5.0 (Macintosh; Intel Mac OS X
_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/83.0.4103.116
i/537.36'}
```

```
ime.sleep(max(random.gauss(3,1),2))
rint(i)
ata={
   "searchCondition":
  {"categoryId":0,"freeShippingType":0,"filters":
   [],"keyWords":"mask"},
   "limit":100,
   "offset":i,
   "language":"en",
   "dataType":"new"

eq=s.post(url_search,headers=headers,json=data)
=req.json()
or item in j['products']:
   price.append(item['price'])
   location.append(item['originCode'])
   name.append(item['name'])
   ratingScore.append(item['leftView']['rateScore'])
   quantity.append(item['rightView']['text'].split(' Sold')[0])

ine all the columns together
t=pd.DataFrame({'Name':name,'price':price,'location':location,'R
 Score':ratingScore,'Quantity Sold':quantity})
```

the complete code to scrape all rows of face mask data in Ezbuy. I nat the total number of rows is 14k and thus I write a for loop to loop incremental offset number to query all the results. Another nt thing to note here is that I put a **random timeout** at the start of op. This is because I do not want very frequent HTTP requests to e traffic of the website and get spotted out by the website.

Recommendation

ant to scrape a website, I would suggest **checking the existence of**

Songhao Wu

433 Followers

Data Enthusiast | Let's have this data j together! | linkedin.com/in/songhaowu

Follow

More from Medium

Matt Chap... in Towards Data Sc...

The Portfolio that Got Me a Data Scientist Job

The PyCoach in Artificial Corner

You're Using ChatGPT Wrong! Here's How to Be Ahead of 99% of ChatGPT Users

Josep Ferrer in Geek Culture

Stop doing this on ChatGPT and get ahead of the 99% of its users

Albers Uzila in Level Up Coding

Wanna Break into Data

t **in the network section** using inspect. If you can find the response to st that gives you all the data you need, you can build a stable and neat .. If you cannot find the data in-network, you should try using s or Selenium to download HTML content and use Beautiful Soup to :he data. Lastly, please use a timeout to avoid a too frequent visits to site or API. This may prevent you from being blocked by the website elps to alleviate the traffic for the good of the website.

re interested to know more about web scraping using Scrapy in can refer to my latest article below

:raping using Scrapy in Python

retrieve second-hand cars information in Singapore

wu.medium.com

ence Web Scraping Machine Learning Data Engineering Python

for The Variable

Data Science

day, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge original features you don't want to miss. Take a look.

you will create a Medium account if you don’t already have one. Review licy for more information about our privacy practices.

Get this newsletter

Science in 2023? Think Twice!

Help Status Writers Blog Careers Privacy Te
Text to speech