Data Acquisition

Why is Machine Learning Difficult?

- The answer is that the data that required to train computers is most of the times **not available**.
- And in other cases, data is in **a very raw format that requires a lot of cleaning and feature engineering**.

What are some common ways to collect data?

1. Collection from an publicly available data repository (files). ✅
2. Web APIs (JSON)
3. Scraping a website (check for legality).
4. Databases (SQL / NoSQL)

Problem Statement: How to get all tweets with a tag `omicron`?

Get all the tweet for from twitter talking about `#omicron`

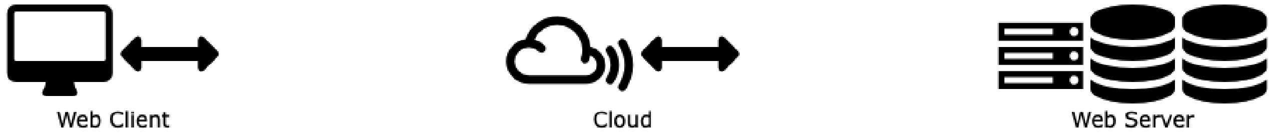
`Transition question` : **Does Twitter offer some sort of service that allows to query data in this manner?**

WWW

The World Wide Web is about communication between web clients and web servers.

Clients are often browsers (Chrome, Edge, Safari), but they can be any type of program or device.

Servers are most often computers in the cloud.



HTTP Request / Response

Communication between clients and servers is done by requests and responses:

A client (a browser) sends an HTTP request to the web A web server receives the request The server runs an application to process the request The server returns an HTTP response (output) to the browser The client (the browser) receives the response

The HTTP Request Circle

A typical HTTP request / response circle:

The browser requests an HTML page. The server returns an HTML file. The browser requests a style sheet. The server returns a CSS file. The browser requests an JPG image. The server returns a JPG file. The browser requests JavaScript code. The server returns a JS file The browser requests data. The server returns data (in XML or JSON).

In [ ]:

The GET Method

GET is used to request data from a specified resource.

Note that the query string (name/value pairs) is sent in the URL of a GET request:

```
/test/demo_form.php?name1=value1&name2=value2
```

Some notes on GET requests:

- GET requests can be cached
- GET requests remain in the browser history
- GET requests can be bookmarked
- GET requests should never be used when dealing with sensitive data
- GET requests have length restrictions
- GET requests are only used to request data (not modify)

The POST Method

POST is used to send data to a server to create/update a resource.

The data sent to the server with POST is stored in the request body of the HTTP request:

```
POST /test/demo_form.php HTTP/1.1
Host: w3schools.com

name1=value1&name2=value2
```

Some notes on POST requests:

- POST requests are never cached
- POST requests do not remain in the browser history
- POST requests cannot be bookmarked
- POST requests have no restrictions on data length

In [ ]:

What is an API?

Application Programming Interface (API), is a software that allows two applications to talk to each other (exchanging the data). Each time you check the weather on your phone, or using a Google Service, you're using an API.

APIs are just like the function calls, but those functions are sitting on the web server and API is the way to invoke those functions in your program.

1. We can send a request to the web server (to its API) to get the data.
2. In return if the call is successfully made the API returns us the data mostly in the **`json` format**.

There are some websites or APIs those offers are open to all and provide free data. Whereas mostly APIs are paid and require some sort of Authentication with the API Keys.

Use cases of APIs:

- APIs can be used to call a function on web server to perform a task and return the required response.
- APIs can also be used to get/send the data over the internet.

Let's first look at few free web APIs and then we will explore paid web api.

How to make these API calls?

Number of things that you should know while making a request on the web:

1. Protocol - HTTP
2. Authentication credentials for the API being called.
3. Functional requirements of the API (URL for the endpoint) - parameters, syntax, and sample response structure (is it JSON or something else) - using the language of your choice.
4. [Optional]: Are there 3rd party solutions available to make these requests easy? - e.g.: `yahoofinance` , `tweepy`

1. How to send an HTTP request using Python?

requests package

In [19]:

```
# !pip install requests
```

In [20]:

```
import requests
import json
```

In [21]:

```
url = "https://api.ipify.org"
```

In [22]:

```
response = requests.get(url)
response
```

Out[22]:

```
<Response [200]>
```

In [23]:

```
response.status_code
```

Out[23]:

```
200
```

In [24]:

```
response.text
```

Out[24]:

```
'122.161.82.151'
```

Getting `omicron` tagged tweets

In [25]:

```
!pip install tweepy
```

```
Requirement already satisfied: tweepy in /Users/harshit/miniconda3/env
s/dsml_env/lib/python3.9/site-packages (4.5.0)
Requirement already satisfied: requests<3,>=2.27.0 in /Users/harshit/m
iniconda3/envs/dsml_env/lib/python3.9/site-packages (from tweepy) (2.2
7.1)
Requirement already satisfied: requests-oauthlib<2,>=1.0.0 in /Users/h
arshit/miniconda3/envs/dsml_env/lib/python3.9/site-packages (from twee
py) (1.3.1)
Requirement already satisfied: idna<4,>=2.5 in /Users/harshit/minicond
a3/envs/dsml_env/lib/python3.9/site-packages (from requests<3,>=2.27.0
->tweepy) (3.3)
Requirement already satisfied: certifi>=2017.4.17 in /Users/harshit/mi
niconda3/envs/dsml_env/lib/python3.9/site-packages (from requests<3,>=
2.27.0->tweepy) (2021.10.8)
Requirement already satisfied: charset-normalizer~=2.0.0 in /Users/har
shit/miniconda3/envs/dsml_env/lib/python3.9/site-packages (from reques
ts<3,>=2.27.0->tweepy) (2.0.11)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/harshi
t/miniconda3/envs/dsml_env/lib/python3.9/site-packages (from requests<
3,>=2.27.0->tweepy) (1.26.8)
Requirement already satisfied: oauthlib>=3.0.0 in /Users/harshit/minic
onda3/envs/dsml_env/lib/python3.9/site-packages (from requests-oauthli
b<2,>=1.0.0->tweepy) (3.2.0)
```

In [26]:

```
import tweepy
```

In [27]:

```
## add bearer token

client = tweepy.Client(bearer_token="AAAAAAAAAAAAAAAAAAAAAPPSYwEAAAAARlorVWosZZfuZs5
```

Extract tweets from an account

In [28]:

```
query = 'from:scaler_official -is:retweet'


tweets = client.search_recent_tweets(query=query, tweet_fields = ['created_at', 'aut
                                     max_results=100)

print(len(tweets.data))
```

```
9
```

In [29]:

```
for tweet in tweets.data:
    print(tweet.text)
    print(tweet.author_id)
    print('----------')
```

```
Watch the full Episode:
https://t.co/5odWQ2gVCT (https://t.co/5odWQ2gVCT)
1194875574331699200
---------
You must've done thousands of #UPI transactions up till now. But do yo
u know how these transfers happen safely and efficiently?

Here's a peep at episode 4 of #OneByteAtATime, where we unfold the tec
hnicality and architecture of the working of UPI.

#SCALEROriginals https://t.co/KuEfntenf9 (https://t.co/KuEfntenf9)
1194875574331699200
---------
"I have witnessed many ups and downs in life. I have lost my two sons
 and my husband. I was completely devastated. But God has given me the
strength to continue to serve the people."
-Madam President

#DroupadiMurmu #CreateImpact https://t.co/rEPADEnvFf (https://t.co/rEP
ADEnvFf)
1194875574331699200
---------
What have we missed out on this list? Which 'friend' you would like to
thank, that gets you through the day? Comment below.

#SCALER #FriendshipDay https://t.co/Z1gd5QgSxi (https://t.co/Z1gd5QgSx
i)
1194875574331699200
---------
What are your favourite TV shows on #AI? Let us know in the comments b
elow.

#CreateImpact #ArtificialIntelligence
1194875574331699200
---------
These are the top 3 #ProgrammingLanguages that developers love by a su
rvey conducted by Stackoverflow.
What language do you think should've been in the top 3 list? Comment b
elow.

#SCALER #Programming https://t.co/KWyGgv0iNH (https://t.co/KWyGgv0iNH)
1194875574331699200
---------
The #FileOrganization describes the logical relationship among the var
ious stored records. In simple words, we can say that this technique d
efines how the file records are mapped onto disk blocks.

#CreateImpact #DBMS https://t.co/iHlyOdywUU (https://t.co/iHlyOdywUU)
1194875574331699200
---------
Full episode link: https://t.co/8YJIKw0FgK (https://t.co/8YJIKw0FgK)

@AnOpenLetter001
1194875574331699200
```

```
----------
What if someday the food that we eat, ends completely? No Biryani, no
 #foodforlife. That day might come sooner than we think.

Here's a nub of Ep.2 of Data Science for Humanity, where we explain ho
w #DataScience might stop food from being eradicated.

#SCALEROriginals https://t.co/YL6hgWVLig (https://t.co/YL6hgWVLig)
1194875574331699200
----------
```

In [30]:

```
##. getting twitter user id from username
client.get_user(username="elonmusk")
```

Out[30]:

```
Response(data=<User id=44196397 name=Elon Musk username=elonmusk>, inc
ludes={}, errors=[], meta={})
```

In [31]:

```
## get the following list of Elon Musk

fol_ing_list = client.get_users_following(id="44196397")
len(fol_ing_list.data)
```

Out[31]:

```
100
```

In [32]:

```
fol_ing_list.data[0]
```

Out[32]:

```
<User id=487833518 name=Nature Portfolio username=NaturePortfolio>
```

In [33]:

```
query = "#omicron -is:retweet"

data = {
    'text': [],
    'created_at': [],
    'author_id': []
}


for tweet in tweepy.Paginator(client.search_recent_tweets, query=query,
                             tweet_fields=['created_at', 'author_id'],
                             max_results=100).flatten(1000):
    data['text'].append(tweet.text)
    data['created_at'].append(tweet.created_at)
    data['author_id'].append(tweet.author_id)
```

In [34]:

```
data
```

Out[34]:

```
{'text': ['Cuba desarrolla candidato vacunal contra la variante #Omicr
on de la #Covid_19 (#Video)\n\nhttps://t.co/noTgpKDWT1\n\n🇨🇺 #Cuba #CO
VID19 #Salud #RadioCubana #CubaVive @Loypa2 https://t.co/dohKRt9AYp',
 (https://t.co/dohKRt9AYp',)
  'The latest The Health care Daily! https://t.co/2N22NSFQbC (https://
t.co/2N22NSFQbC) Thanks to @Crash20153 #omicron #covid19',
  'El 89% de estos participantes había recibido dos dosis de una vacun
a ARNm y ninguno había recibido una dosis de refuerzo.\nTos persistent
e, secreción nasal, cansancio, fatiga muscular, dolor de cabeza, dolor
de garganta, fiebre, estornudos. #Ómicron #COVID19 #COVID https://t.c
o/PF1z5uPnKJ', (https://t.co/PF1z5uPnKJ',)
  "What a nice bottle, isn't it?!\n#hk #ig #hkig #iger #igers #hkiger
 #nicebottle #pennybay #covid19 #covid_19 #omicron #isolationcentre #i
solationcamp @ 大嶼山竹篙灣隔離營中心地盤 https://t.co/cWGWUnc66g", (http
s://t.co/cWGWUnc66g",)
  'Biontech’ten yeni aşı açıklaması! Tarih verildi…\n\n#salı\n#BioNTec
h\n#Omicron\n\nhttps://t.co/aWZHywSKey',
  'महाराष्ट्राचा कोरोना रिपोर्ट: मंगळवार, ९ ऑगस्ट २०२२\n\nराज्यात १७८२ नवे रुग्ण, १८५४ बरे। मुंबई
```

In [35]:

```
import pandas as pd

df = pd.DataFrame(data)
df.head()
```

Out[35]:

	text	created_at	author_id
0	Cuba desarrolla candidato vacunal contra la va...	2022-08-09 14:56:09+00:00	305617100
1	The latest The Health care Daily! https://t.co...	2022-08-09 14:53:50+00:00	1019604048591253504
2	El 89% de estos participantes había recibido d...	2022-08-09 14:53:11+00:00	145010613
3	What a nice bottle, isn't it?!\n#hk #ig #hkig ...	2022-08-09 14:50:41+00:00	2320664202
4	Biontech’ten yeni aşı açıklaması! Tarih verild...	2022-08-09 14:42:09+00:00	1487410657607704585

In [36]:

```
df.to_csv("tweets.csv", index=False)
```

OpenWeathermap API

In [37]:

```
url = "https://api.openweathermap.org/data/2.5/weather?q=delhi&appid=9b199c2b6cd2fbc
```

In [38]:

```
r = requests.get(url)
r.text
```

Out[38]:

```
'{"coord":{"lon":77.2167,"lat":28.6667},"weather":[{"id":721,"main":"H
aze","description":"haze","icon":"50n"}],"base":"stations","main":{"te
mp":306.2,"feels_like":313.2,"temp_min":304.99,"temp_max":306.2,"press
ure":999,"humidity":66},"visibility":3000,"wind":{"speed":1.54,"deg":1
00},"clouds":{"all":75},"dt":1660056483,"sys":{"type":1,"id":9165,"cou
ntry":"IN","sunrise":1660004210,"sunset":1660052198},"timezone":1980
0,"id":1273294,"name":"Delhi","cod":200}'
```

In [39]:

```
weather = json.loads(r.text)
weather
```

Out[39]:

```
{'coord': {'lon': 77.2167, 'lat': 28.6667},
 'weather': [{'id': 721,
   'main': 'Haze',
   'description': 'haze',
   'icon': '50n'}],
 'base': 'stations',
 'main': {'temp': 306.2,
  'feels_like': 313.2,
  'temp_min': 304.99,
  'temp_max': 306.2,
  'pressure': 999,
  'humidity': 66},
 'visibility': 3000,
 'wind': {'speed': 1.54, 'deg': 100},
 'clouds': {'all': 75},
 'dt': 1660056483,
 'sys': {'type': 1,
  'id': 9165,
  'country': 'IN',
  'sunrise': 1660004210,
  'sunset': 1660052198},
 'timezone': 19800,
 'id': 1273294,
 'name': 'Delhi',
 'cod': 200}
```

In [ ]:

```

```

Scraping information from webpages

In [40]:

```
!pip install beautifulsoup4
```

```
Requirement already satisfied: beautifulsoup4 in /Users/harshit/minico
nda3/envs/dsml_env/lib/python3.9/site-packages (4.10.0)
Requirement already satisfied: soupsieve>1.2 in /Users/harshit/minicon
da3/envs/dsml_env/lib/python3.9/site-packages (from beautifulsoup4)
(2.3.1)
```

In [41]:

```
baseurl = "http://books.toscrape.com/"

r = requests.get(baseurl)

r.content
```

Out[41]:

```
b'<!DOCTYPE html>\n<!--[if lt IE 7]>      <html lang="en-us" class="no
-js lt-ie9 lt-ie8 lt-ie7"> <![endif]-->\n<!--[if IE 7]>         <html
lang="en-us" class="no-js lt-ie9 lt-ie8"> <![endif]-->\n<!--[if IE 8]>
<html lang="en-us" class="no-js lt-ie9"> <![endif]-->\n<!--[if gt IE
8]><!--> <html lang="en-us" class="no-js"> <!--<![endif]-->\n    <head
>\n        <title>\n    All products | Books to Scrape - Sandbox\n</ti
tle>\n\n        <meta http-equiv="content-type" content="text/html; ch
arset=UTF-8" />\n        <meta name="created" content="24th Jun 2016 0
9:29" />\n        <meta name="description" content="" />\n        <met
a name="viewport" content="width=device-width" />\n        <meta name
="robots" content="NOARCHIVE,NOCACHE" />\n\n        <!-- Le HTML5 shi
m, for IE6-8 support of HTML elements -->\n        <!--[if lt IE 9]>\n
<script src="//html5shim.googlecode.com/svn/trunk/html5.js"></script>
\n        <![endif]-->\n\n        \n            <link rel="shortcut ic
on" href="static/oscar/favicon.ico" />\n        \n\n        \n
\n    \n    \n        <link rel="stylesheet" type="text/css" href="sta
tic/oscar/css/styles.css" />\n    \n    <link rel="stylesheet" href="s
tatic/oscar/js/bootstrap-datetimepicker/bootstrap-datetimepicker.css"
```

In [42]:

```
from bs4 import BeautifulSoup
```

In [43]:

```
soup = BeautifulSoup(r.content)
soup
```

Out[43]:

```
<!DOCTYPE html>
<!--[if lt IE 7]>      <html lang="en-us" class="no-js lt-ie9 lt-ie8 l
t-ie7"> <![endif]--><!--[if IE 7]>         <html lang="en-us" class="n
o-js lt-ie9 lt-ie8"> <![endif]--><!--[if IE 8]>         <html lang="en
-us" class="no-js lt-ie9"> <![endif]--><!--[if gt IE 8]><!--><html cla
ss="no-js" lang="en-us"> <!--<![endif]-->
<head>
<title>
    All products | Books to Scrape - Sandbox
</title>
<meta content="text/html; charset=utf-8" http-equiv="content-type"/>
<meta content="24th Jun 2016 09:29" name="created"/>
<meta content="" name="description"/>
<meta content="width=device-width" name="viewport"/>
<meta content="NOARCHIVE,NOCACHE" name="robots"/>
<!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
<!--[if lt IE 9]>
        <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></
```

In [44]:

```
ul_list = soup.find('ul', class_="nav-list")
```

In [45]:

```
ul_li_items = ul_list.ul.find_all('li')
baseurl + ul_li_items[0].a['href']
```

Out[45]:

```
'http://books.toscrape.com/catalogue/category/books/travel_2/index.htm
l'
```

In [46]:

```
def extract_categories(baseurl):
    categories = {}
    r = requests.get(baseurl)
    soup = BeautifulSoup(r.content)
    categories_list = soup.find('ul', class_="nav-list").ul.find_all("li")
    for li in categories_list:
        categories.update({li.text.strip(): baseurl + li.a['href']})
    return categories
```

In [47]:

```
extract_categories(baseurl)
```

Out[47]:

```
{'Travel': 'http://books.toscrape.com/catalogue/category/books/travel_
2/index.html',
 'Mystery': 'http://books.toscrape.com/catalogue/category/books/myster
y_3/index.html',
 'Historical Fiction': 'http://books.toscrape.com/catalogue/category/b
ooks/historical-fiction_4/index.html',
 'Sequential Art': 'http://books.toscrape.com/catalogue/category/book
s/sequential-art_5/index.html',
 'Classics': 'http://books.toscrape.com/catalogue/category/books/class
ics_6/index.html',
 'Philosophy': 'http://books.toscrape.com/catalogue/category/books/phi
losophy_7/index.html',
 'Romance': 'http://books.toscrape.com/catalogue/category/books/romanc
e_8/index.html',
 'Womens Fiction': 'http://books.toscrape.com/catalogue/category/book
s/womens-fiction_9/index.html',
 'Fiction': 'http://books.toscrape.com/catalogue/category/books/fictio
n_10/index.html',
 'Childrens': 'http://books.toscrape.com/catalogue/category/books/chil
drens_11/index.html',
 'Religion': 'http://books.toscrape.com/catalogue/category/books/relig
ion_12/index.html',
 'Nonfiction': 'http://books.toscrape.com/catalogue/category/books/non
fiction_13/index.html',
 'Music': 'http://books.toscrape.com/catalogue/category/books/music_1
4/index.html',
 'Default': 'http://books.toscrape.com/catalogue/category/books/defaul
t_15/index.html',
 'Science Fiction': 'http://books.toscrape.com/catalogue/category/book
s/science-fiction_16/index.html',
 'Sports and Games': 'http://books.toscrape.com/catalogue/category/boo
ks/sports-and-games_17/index.html',
 'Add a comment': 'http://books.toscrape.com/catalogue/category/books/
add-a-comment_18/index.html',
 'Fantasy': 'http://books.toscrape.com/catalogue/category/books/fantas
y_19/index.html',
 'New Adult': 'http://books.toscrape.com/catalogue/category/books/new-
adult_20/index.html',
 'Young Adult': 'http://books.toscrape.com/catalogue/category/books/yo
ung-adult_21/index.html',
 'Science': 'http://books.toscrape.com/catalogue/category/books/scienc
e_22/index.html',
 'Poetry': 'http://books.toscrape.com/catalogue/category/books/poetry_
23/index.html',
 'Paranormal': 'http://books.toscrape.com/catalogue/category/books/par
anormal_24/index.html',
 'Art': 'http://books.toscrape.com/catalogue/category/books/art_25/ind
ex.html',
 'Psychology': 'http://books.toscrape.com/catalogue/category/books/psy
chology_26/index.html',
 'Autobiography': 'http://books.toscrape.com/catalogue/category/books/
autobiography_27/index.html',
 'Parenting': 'http://books.toscrape.com/catalogue/category/books/pare
nting_28/index.html',
 'Adult Fiction': 'http://books.toscrape.com/catalogue/category/books/
adult-fiction_29/index.html',
 'Humor': 'http://books.toscrape.com/catalogue/category/books/humor_3
0/index.html',
 'Horror': 'http://books.toscrape.com/catalogue/category/books/horror_
31/index.html',
 'History': 'http://books.toscrape.com/catalogue/category/books/histor
y_32/index.html',
 'Food and Drink': 'http://books.toscrape.com/catalogue/category/book
s/food-and-drink_33/index.html',
 'Christian Fiction': 'http://books.toscrape.com/catalogue/category/bo
oks/christian-fiction_34/index.html',
 'Business': 'http://books.toscrape.com/catalogue/category/books/busin
ess_35/index.html',
 'Biography': 'http://books.toscrape.com/catalogue/category/books/biog
raphy_36/index.html',
 'Thriller': 'http://books.toscrape.com/catalogue/category/books/thril
ler_37/index.html',
 'Contemporary': 'http://books.toscrape.com/catalogue/category/books/c
ontemporary_38/index.html',
 'Spirituality': 'http://books.toscrape.com/catalogue/category/books/s
pirituality_39/index.html',
 'Academic': 'http://books.toscrape.com/catalogue/category/books/acade
mic_40/index.html',
 'Self Help': 'http://books.toscrape.com/catalogue/category/books/self
-help_41/index.html',
 'Historical': 'http://books.toscrape.com/catalogue/category/books/his
torical_42/index.html',
 'Christian': 'http://books.toscrape.com/catalogue/category/books/chri
stian_43/index.html',
 'Suspense': 'http://books.toscrape.com/catalogue/category/books/suspe
nse_44/index.html',
 'Short Stories': 'http://books.toscrape.com/catalogue/category/books/
short-stories_45/index.html',
 'Novels': 'http://books.toscrape.com/catalogue/category/books/novels_
46/index.html',
 'Health': 'http://books.toscrape.com/catalogue/category/books/health_
47/index.html',
 'Politics': 'http://books.toscrape.com/catalogue/category/books/polit
ics_48/index.html',
 'Cultural': 'http://books.toscrape.com/catalogue/category/books/cultu
ral_49/index.html',
 'Erotica': 'http://books.toscrape.com/catalogue/category/books/erotic
a_50/index.html',
 'Crime': 'http://books.toscrape.com/catalogue/category/books/crime_5
1/index.html'}
```

In [ ]:

```

```

In [ ]:

```

```

Extracting all book info

In [48]:

```
url = "http://books.toscrape.com/catalogue/category/books/mystery_3/index.html"
res = requests.get(url)

soup = BeautifulSoup(res.content)
```

In [49]:

```
soup
```

```
<meta content="

" name="description"/>
<meta content="width=device-width" name="viewport"/>
<meta content="NOARCHIVE,NOCACHE" name="robots"/>
<!-- Le HTML5 shim, for IE6-8 support of HTML elements -->
<!--[if lt IE 9]>
        <script src="//html5shim.googlecode.com/svn/trunk/html5.js"></
script>
        <![endif]-->
<link href="../../../../static/oscar/favicon.ico" rel="shortcut icon"/
>
<link href="../../../../static/oscar/css/styles.css" rel="stylesheet"
type="text/css"/>
<link href="../../../../static/oscar/js/bootstrap-datetimepicker/boots
trap-datetimepicker.css" rel="stylesheet"/>
<link href="../../../../static/oscar/css/datetimepicker.css" rel="styl
esheet" type="text/css"/>
</head>
<body class="default" id="default">
```

In [73]:

```
data = {
    'product_page_url': [],
    'title': [],
    'price_including_tax': [],
    'number_available': []
}
```

In [50]:

```
book_page_list = soup.find('ol', class_="row").find_all("li")
book_page_list
```

Out[50]:

```
[<li class="col-xs-6 col-sm-4 col-md-3 col-lg-3">
 <article class="product_pod">
 <div class="image_container">
 <a href="../../../sharp-objects_997/index.html"><img alt="Sharp Objec
ts" class="thumbnail" src="../../../../media/cache/32/51/3251cf3a3412f
53f339e42cac2134093.jpg"/></a>
 </div>
 <p class="star-rating Four">
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 <i class="icon-star"></i>
 </p>
 <h3><a href="../../../sharp-objects_997/index.html" title="Sharp Obje
cts">Sharp Objects</a></h3>
 <div class="product_price">
 <p class="price color">£47.82</p>
```

In [59]:

```
book_page_list[0].a['href'].split("/")[3:]
```

Out[59]:

```
['sharp-objects_997', 'index.html']
```

In [74]:

```
for li in book_page_list:
    href_split = li.a['href'].split("/")[3:]
    book_page_link = "http://books.toscrape.com/catalogue/" + "/".join(href_split)

    print(book_page_link)

    req = requests.get(book_page_link)
    soup = BeautifulSoup(req.content, features="html.parser")

    data['product_page_url'].append(book_page_link)

    rows = soup.find("table").find_all("tr")

    title = soup.find('title').text.strip()
    data['title'].append(title)

    data['price_including_tax'].append(rows[3].find('td').text.strip())
    data['number_available'].append(rows[5].find('td').text.strip())
```

```
http://books.toscrape.com/catalogue/sharp-objects_997/index.html (htt
p://books.toscrape.com/catalogue/sharp-objects_997/index.html)
http://books.toscrape.com/catalogue/in-a-dark-dark-wood_963/index.html
(http://books.toscrape.com/catalogue/in-a-dark-dark-wood_963/index.htm
l)
http://books.toscrape.com/catalogue/the-past-never-ends_942/index.html
(http://books.toscrape.com/catalogue/the-past-never-ends_942/index.htm
l)
http://books.toscrape.com/catalogue/a-murder-in-time_877/index.html (h
ttp://books.toscrape.com/catalogue/a-murder-in-time_877/index.html)
http://books.toscrape.com/catalogue/the-murder-of-roger-ackroyd-hercul
e-poirot-4_852/index.html (http://books.toscrape.com/catalogue/the-mur
der-of-roger-ackroyd-hercule-poirot-4_852/index.html)
http://books.toscrape.com/catalogue/the-last-mile-amos-decker-2_754/in
dex.html (http://books.toscrape.com/catalogue/the-last-mile-amos-decke
r-2_754/index.html)
http://books.toscrape.com/catalogue/that-darkness-gardiner-and-renner-
1_743/index.html (http://books.toscrape.com/catalogue/that-darkness-ga
rdiner-and-renner-1_743/index.html)
http://books.toscrape.com/catalogue/tastes-like-fear-di-marnie-rome-3_
742/index.html (http://books.toscrape.com/catalogue/tastes-like-fear-d
i-marnie-rome-3_742/index.html)
http://books.toscrape.com/catalogue/a-time-of-torment-charlie-parker-1
4_657/index.html (http://books.toscrape.com/catalogue/a-time-of-tormen
t-charlie-parker-14_657/index.html)
http://books.toscrape.com/catalogue/a-study-in-scarlet-sherlock-holmes
-1_656/index.html (http://books.toscrape.com/catalogue/a-study-in-scar
let-sherlock-holmes-1_656/index.html)
http://books.toscrape.com/catalogue/poisonous-max-revere-novels-3_627/
index.html (http://books.toscrape.com/catalogue/poisonous-max-revere-n
ovels-3_627/index.html)
http://books.toscrape.com/catalogue/murder-at-the-42nd-street-library-
raymond-ambler-1_624/index.html (http://books.toscrape.com/catalogue/m
urder-at-the-42nd-street-library-raymond-ambler-1_624/index.html)
http://books.toscrape.com/catalogue/most-wanted_623/index.html (htt
p://books.toscrape.com/catalogue/most-wanted_623/index.html)
http://books.toscrape.com/catalogue/hide-away-eve-duncan-20_620/index.
html (http://books.toscrape.com/catalogue/hide-away-eve-duncan-20_620/
index.html)
http://books.toscrape.com/catalogue/boar-island-anna-pigeon-19_613/ind
ex.html (http://books.toscrape.com/catalogue/boar-island-anna-pigeon-1
9_613/index.html)
http://books.toscrape.com/catalogue/the-widow_609/index.html (http://b
ooks.toscrape.com/catalogue/the-widow_609/index.html)
http://books.toscrape.com/catalogue/playing-with-fire_602/index.html
 (http://books.toscrape.com/catalogue/playing-with-fire_602/index.htm
l)
http://books.toscrape.com/catalogue/what-happened-on-beale-street-secr
ets-of-the-south-mysteries-2_506/index.html (http://books.toscrape.co
m/catalogue/what-happened-on-beale-street-secrets-of-the-south-mysteri
es-2_506/index.html)
http://books.toscrape.com/catalogue/the-bachelor-girls-guide-to-murder
-herringford-and-watts-mysteries-1_491/index.html (http://books.toscra
pe.com/catalogue/the-bachelor-girls-guide-to-murder-herringford-and-wa
tts-mysteries-1_491/index.html)
http://books.toscrape.com/catalogue/delivering-the-truth-quaker-midwif
e-mystery-1_464/index.html (http://books.toscrape.com/catalogue/delive
ring-the-truth-quaker-midwife-mystery-1_464/index.html)
```

In [ ]:

In [68]:

```
data
```

Out[68]:

```
{'product_page_url': ['http://books.toscrape.com/catalogue/sharp-objec
ts_997/index.html',
  'http://books.toscrape.com/catalogue/in-a-dark-dark-wood_963/index.h
tml',
  'http://books.toscrape.com/catalogue/the-past-never-ends_942/index.h
tml',
  'http://books.toscrape.com/catalogue/a-murder-in-time_877/index.htm
l',
  'http://books.toscrape.com/catalogue/the-murder-of-roger-ackroyd-her
cule-poirot-4_852/index.html',
  'http://books.toscrape.com/catalogue/the-last-mile-amos-decker-2_75
4/index.html',
  'http://books.toscrape.com/catalogue/that-darkness-gardiner-and-renn
er-1_743/index.html',
  'http://books.toscrape.com/catalogue/tastes-like-fear-di-marnie-rome
-3_742/index.html',
  'http://books.toscrape.com/catalogue/a-time-of-torment-charlie-parke
r-14_657/index.html',
  'http://books.toscrape.com/catalogue/a-study-in-scarlet-sherlock-hol
mes-1_656/index.html',
  'http://books.toscrape.com/catalogue/poisonous-max-revere-novels-3_6
27/index.html',
  'http://books.toscrape.com/catalogue/murder-at-the-42nd-street-libra
ry-raymond-ambler-1_624/index.html',
  'http://books.toscrape.com/catalogue/most-wanted_623/index.html',
  'http://books.toscrape.com/catalogue/hide-away-eve-duncan-20_620/ind
ex.html',
  'http://books.toscrape.com/catalogue/boar-island-anna-pigeon-19_613/
index.html',
  'http://books.toscrape.com/catalogue/the-widow_609/index.html',
  'http://books.toscrape.com/catalogue/playing-with-fire_602/index.htm
l',
  'http://books.toscrape.com/catalogue/what-happened-on-beale-street-s
ecrets-of-the-south-mysteries-2_506/index.html',
  'http://books.toscrape.com/catalogue/the-bachelor-girls-guide-to-mur
der-herringford-and-watts-mysteries-1_491/index.html',
  'http://books.toscrape.com/catalogue/delivering-the-truth-quaker-mid
wife-mystery-1_464/index.html'],
 'title': [],
 'price_including_tax': ['£47.82',
  '£19.63',
  '£56.50',
  '£16.64',
  '£44.10',
  '£54.21',
  '£13.92',
  '£10.69',
  '£48.35',
  '£16.73',
  '£26.80',
  '£54.36',
  '£35.28',
  '£11.84',
  '£59.48',
  '£27.26',
  '£13.71',
  '£25.37',
  '£52.30',
  '£20.89'],
 'number_available': ['In stock (20 available)',
  'In stock (18 available)',
  'In stock (16 available)',
  'In stock (16 available)',
  'In stock (15 available)',
  'In stock (14 available)',
  'In stock (14 available)',
  'In stock (14 available)',
  'In stock (14 available)',
  'In stock (14 available)',
  'In stock (12 available)',
  'In stock (12 available)',
  'In stock (12 available)',
  'In stock (12 available)',
  'In stock (12 available)',
  'In stock (11 available)',
  'In stock (11 available)',
  'In stock (7 available)',
  'In stock (7 available)',
  'In stock (7 available)']}
```

In [75]:

```
df = pd.DataFrame(data)
df
```

Out[75]:

	product_page_url	title	price_including_tax	number_availal
0	http://books.toscrape.com/catalogue/sharp-obje...	Sharp Objects | Books to Scrape - Sandbox	£47.82	In stock availal
1	http://books.toscrape.com/catalogue/in-a-dark-...	In a Dark, Dark Wood | Books to Scrape - Sandbox	£19.63	In stock availal
2	http://books.toscrape.com/catalogue/the-past-n...	The Past Never Ends | Books to Scrape - Sandbox	£56.50	In stock availal
3	http://books.toscrape.com/catalogue/a-murder-i...	A Murder in Time | Books to Scrape - Sandbox	£16.64	In stock availal
4	http://books.toscrape.com/catalogue/the-murder...	The Murder of Roger Ackroyd (Hercule Poirot #4...	£44.10	In stock availal
5	http://books.toscrape.com/catalogue/the-last-m...	The Last Mile (Amos Decker #2) | Books to Scra...	£54.21	In stock availal
6	http://books.toscrape.com/catalogue/that-darkn...	That Darkness (Gardiner and Renner #1) | Books...	£13.92	In stock availal
7	http://books.toscrape.com/catalogue/tastes-lik...	Tastes Like Fear (DI Marnie Rome #3) | Books t...	£10.69	In stock availal
8	http://books.toscrape.com/catalogue/a-time-of-...	A Time of Torment (Charlie Parker #14) | Books...	£48.35	In stock availal
9	http://books.toscrape.com/catalogue/a-study-in...	A Study in Scarlet (Sherlock Holmes #1) | Book...	£16.73	In stock availal

	product_page_url	title	price_including_tax	number_availal
10	http://books.toscrape.com/catalogue/poisonous-...	Poisonous (Max Revere Novels #3) | Books to Sc...	£26.80	In stock availal
11	http://books.toscrape.com/catalogue/murder-at-...	Murder at the 42nd Street Library (Raymond Amb...	£54.36	In stock availal
12	http://books.toscrape.com/catalogue/most-wante...	Most Wanted | Books to Scrape - Sandbox	£35.28	In stock availal
13	http://books.toscrape.com/catalogue/hide-away-...	Hide Away (Eve Duncan #20) | Books to Scrape -...	£11.84	In stock availal
14	http://books.toscrape.com/catalogue/boar-islan...	Boar Island (Anna Pigeon #19) | Books to Scrap...	£59.48	In stock availal
15	http://books.toscrape.com/catalogue/the-widow_...	The Widow | Books to Scrape - Sandbox	£27.26	In stock availal
16	http://books.toscrape.com/catalogue/playing-wi...	Playing with Fire | Books to Scrape - Sandbox	£13.71	In stock availal
17	http://books.toscrape.com/catalogue/what-happe...	What Happened on Beale Street (Secrets of the ...	£25.37	In stocl availal
18	http://books.toscrape.com/catalogue/the-bachel...	The Bachelor Girl's Guide to Murder (Herringfo...	£52.30	In stocl availal
19	http://books.toscrape.com/catalogue/delivering...	Delivering the Truth (Quaker Midwife Mystery #...	£20.89	In stocl availal

In [ ]: