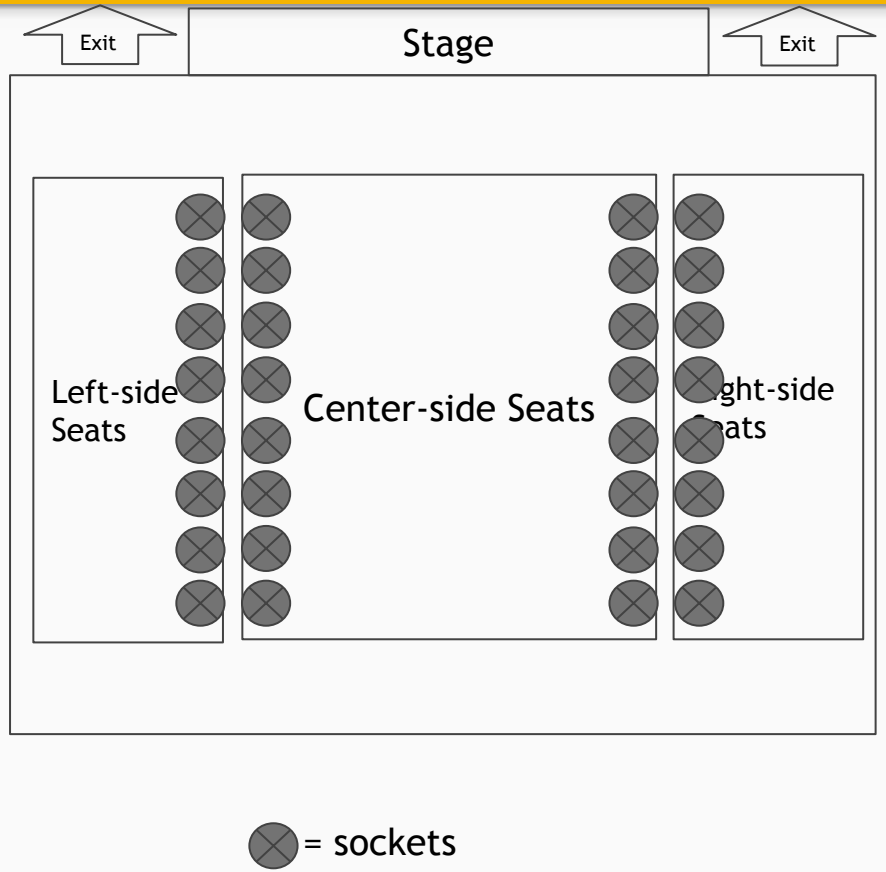
Yasumoto International Academic Park - YIA LT6

Limited Mobile Signal. Please use on-campus wifi.
32 Sockets. But please bring your own charger.

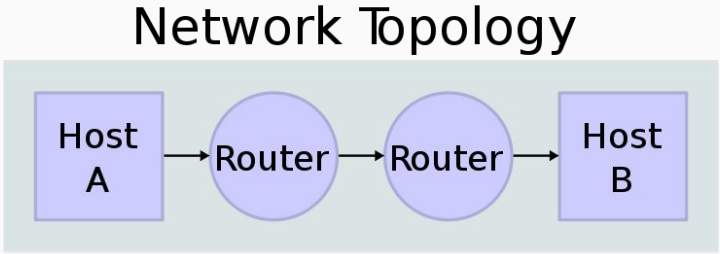


Web Scraping, API

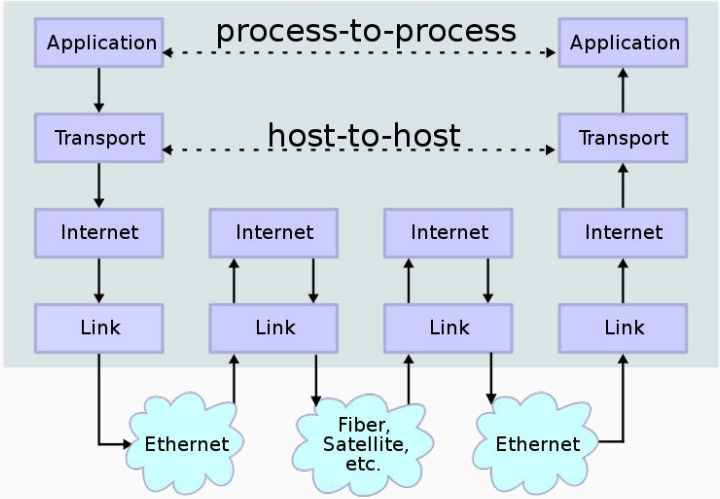
CUHK MSc Data Science & Biz Stat. Program
STAT5106 - Programming Techniques for Data Science
Week 5 @ 10 Oct 2024

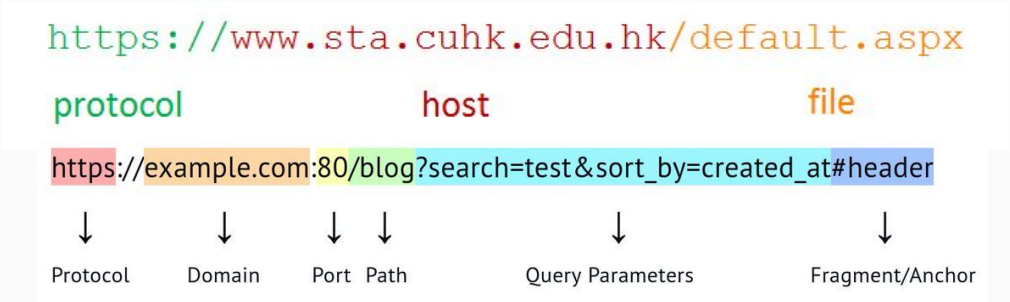
Some knowledge on internet

- Reference: Dr. Chuck free courses
  - Internet History, Tech, Security
  - Python for Everybody Ch. 13, 14



- TCP/IP
- Protocol
  A set of rules that all parties follow so we can predict each other's behavior
  - HTTP - HyperText Transfer Protocol
  - HTTPS - (+) Secure
  - Request - GET / PUT / DELETE …
- Protocol + Host + Files





tracert / traceroute

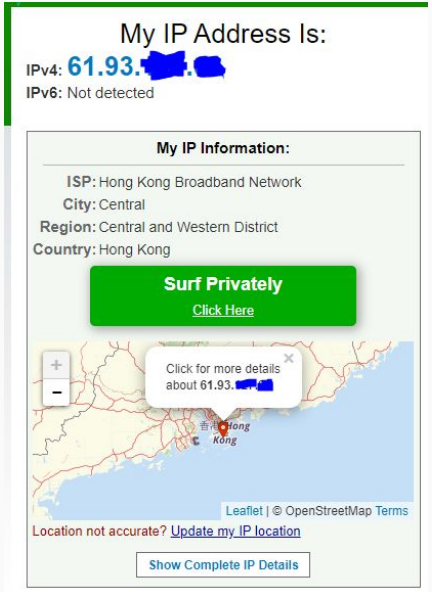
P.S. 你果度我冇睇錯係黃大仙黎，要搵你住邊計下角度就知，IP 都唔洗check

<= what does this mean?

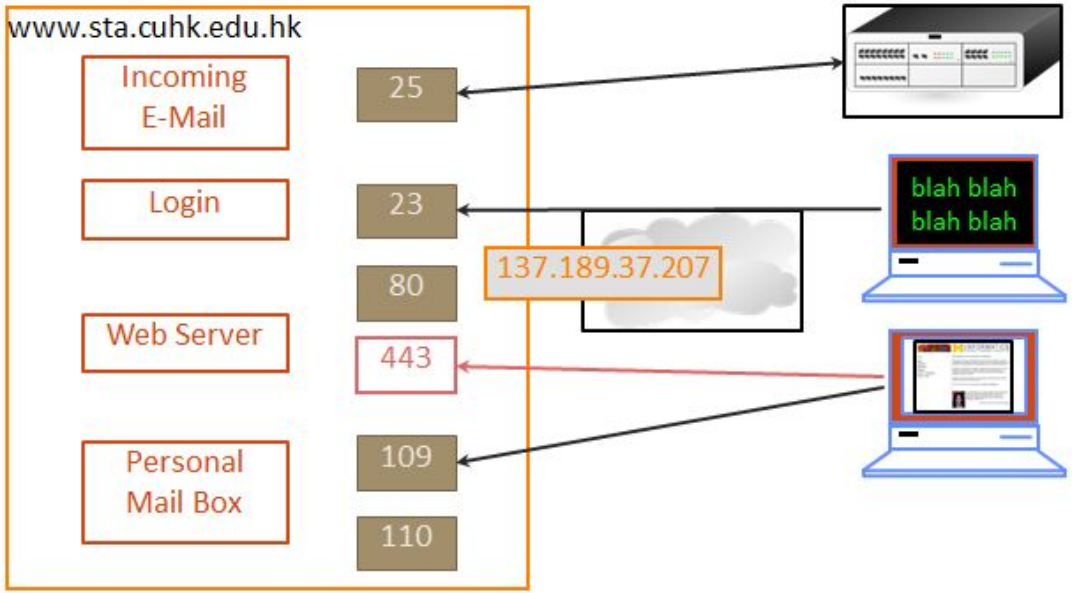
```
C:\Users\kyala>tracert www.wikipedia.org

在上限 30 個躍點上
追蹤 dyna.wikimedia.org [103.102.166.224] 的路由:

  1     4 ms     2 ms     2 ms  192.168.0.1
  2     6 ms     6 ms     6 ms  061093127001.ctinets.com [61.93.127.1]
  3     7 ms     7 ms     6 ms  10.239.98.1
  4    10 ms     9 ms     9 ms  014199253001.ctinets.com [14.199.253.1]
  5     7 ms     7 ms     6 ms  061244224065.ctinets.com [61.244.224.65]
  6    41 ms    44 ms    42 ms  14907.sgw.equinix.com [27.111.228.186]
  7    42 ms    42 ms    41 ms  text-lb.eqsin.wikimedia.org [103.102.166.224]
```

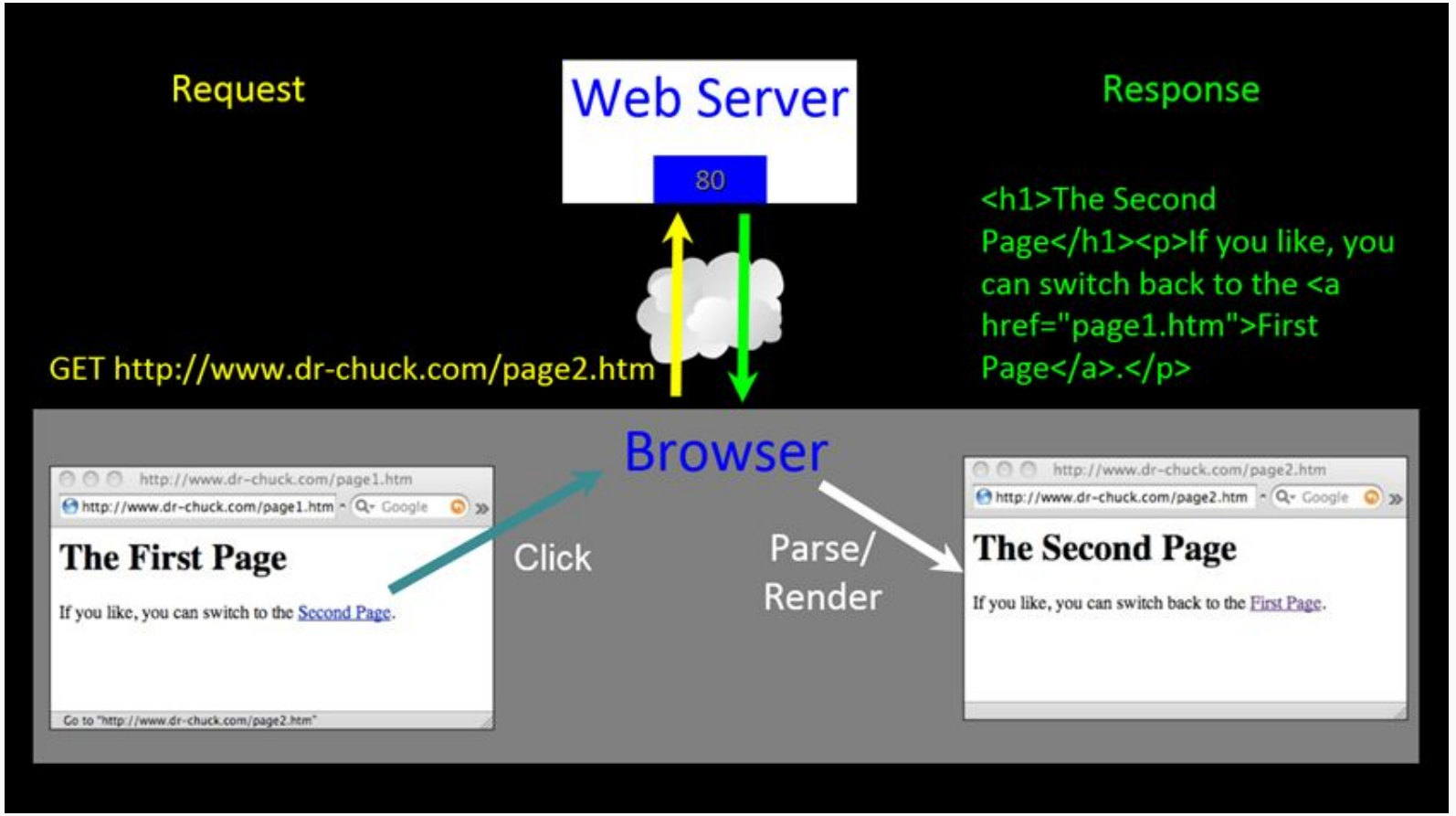


Port



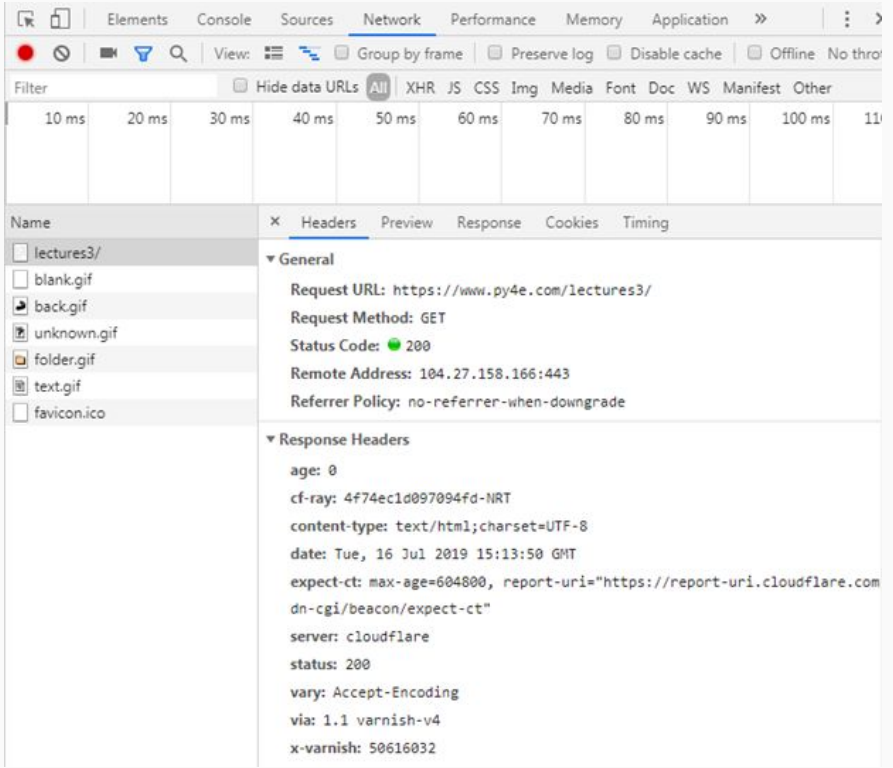
Telnet (23) – Login	SSH (22) – Secure Login
HTTP (80)	HTTPS (443)
SMTP (25) (Mail)	FTP (21) – File Transfer

How to get something on the web



Ctrl+U: View Source Code / F12 – Network Information

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<html>
 <head>
  <title>Index of /lectures3</title>
 </head>
 <body>
<h1>Index of /lectures3</h1>
  <table>
    <tr><th valign="top"><img src="/icons/blank.gif" alt="[ICO]"></th><th><a href="?C=N;O=D">Name</a></th><th><a href="?
C=M;O=A">Last modified</a></th><th><a href="?C=S;O=A">Size</a></th><th><a href="?C=D;O=A">Description</a></th></tr>
    <tr><th colspan="5"><hr></th></tr>
<tr><td valign="top"><img src="/icons/back.gif" alt="[PARENTDIR]"></td><td><a href="/">Parent Directory</a></td><td> 
</td><td align="right">  - </td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="00-Master.ppt">00-Master.ppt</a></td><td
align="right">2018-01-04 18:11  </td><td align="right"> 45K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-01-Intro.pptx">Pythonlearn-01-
Intro.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">2.1M</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/folder.gif" alt="[DIR]"></td><td><a href="Pythonlearn-01/">Pythonlearn-01/</a></td>
<td align="right">2018-01-04 18:11  </td><td align="right">  - </td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-02-
Expressions.pptx">Pythonlearn-02-Expressions.pptx</a></td><td align="right">2018-01-04 18:11  </td><td
align="right">236K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-03-
Conditional.pptx">Pythonlearn-03-Conditional.pptx</a></td><td align="right">2018-01-04 18:11  </td><td
align="right">266K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-04-
Functions.pptx">Pythonlearn-04-Functions.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">126K</td>
<td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-05-
Iterations.pptx">Pythonlearn-05-Iterations.pptx</a></td><td align="right">2018-01-04 18:11  </td><td
align="right">380K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-06-Strings.pptx">Pythonlearn-
06-Strings.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">345K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-07-Files.pptx">Pythonlearn-07-
Files.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">185K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-08-Lists.pptx">Pythonlearn-08-
Lists.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">162K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-09-
Dictionaries.pptx">Pythonlearn-09-Dictionaries.pptx</a></td><td align="right">2018-01-04 18:11  </td><td
align="right">1.9M</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-10-Tuples.pptx">Pythonlearn-
10-Tuples.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">  88K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/text.gif" alt="[TXT]"></td><td><a href="Pythonlearn-11-Regex-
Handout.txt">Pythonlearn-11-Regex-Handout.txt</a></td><td align="right">2018-01-04 18:11  </td><td align="right">730 </td>
<td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-11-Regex.pptx">Pythonlearn-11-
Regex.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">345K</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-12-HTTP.pptx">Pythonlearn-12-
HTTP.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">4.7M</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/folder.gif" alt="[DIR]"></td><td><a href="Pythonlearn-12-HTTP/">Pythonlearn-12-
HTTP/</a></td><td align="right">2018-01-04 18:11  </td><td align="right">  - </td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-13-
WebServices.pptx">Pythonlearn-13-WebServices.pptx</a></td><td align="right">2018-01-04 18:11  </td><td
align="right">2.9M</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/unknown.gif" alt="[   ]"></td><td><a href="Pythonlearn-14-Objects.pptx">Pythonlearn-
14-Objects.pptx</a></td><td align="right">2018-01-04 18:11  </td><td align="right">3.6M</td><td> </td></tr>
<tr><td valign="top"><img src="/icons/text.gif" alt="[TXT]"></td><td><a href="Pythonlearn-15-Database-
```



HTTP response

100 Continue
101 Switching Protocols
103 Early Hints

200 OK
201 Created
202 Accepted
203 Non-Authoritative Information
204 No Content
205 Reset Content
206 Partial Content

300 Multiple Choices
301 Moved Permanently
302 Found
303 See Other
304 Not Modified

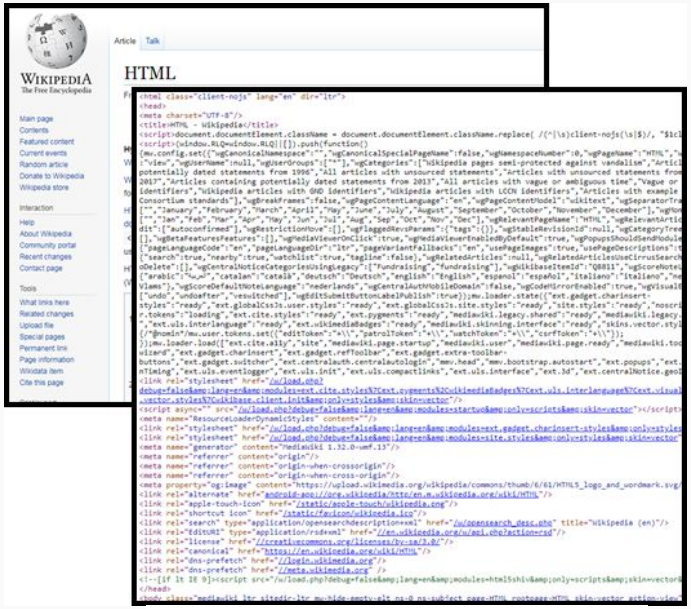
400 Bad Request
401 Unauthorized
402 Payment Required
403 Forbidden
404 Not Found
405 Method Not Allowed
406 Not Acceptable
407 Proxy Authentication Required
408 Request Timeout

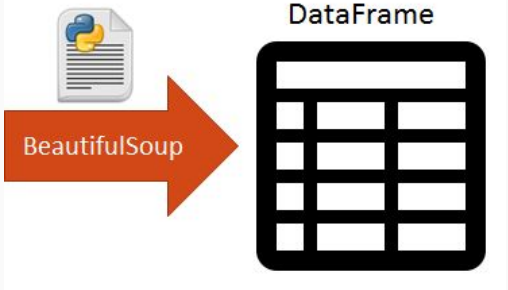
500 Internal Server Error
501 Not Implemented
502 Bad Gateway
503 Service Unavailable
504 Gateway Timeout
505 HTTP Version Not Supported
511 Network Authentication Required

Start Coding...

Please access Week 5 - Web Scraping...

Web Scraping: using urllib + BeautifulSoup





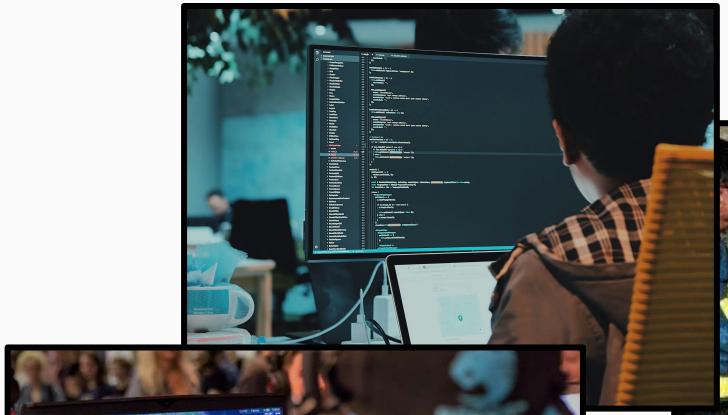

```
import urllib.request
from bs4 import BeautifulSoup


html = urllib.request.urlopen('https://www.sta.cuhk.edu.hk').read()
# " <html  lang="en-US">
# <head id="Head">
# <title>
#         Statistics > Home
# </title>… "
soup = BeautifulSoup(html, 'html.parser')


# Retrieve all of the anchor tags
tags = soup('a')
for tag in tags:
    print(tag.get('href', None))
```

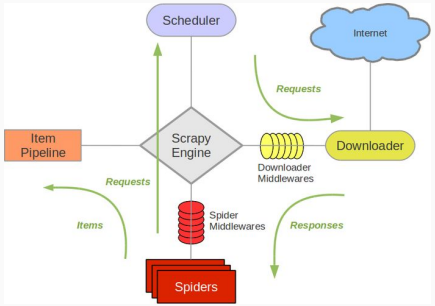
The extreme example (unwire)

A fellow who goes by the handle beaston02 wanted to see how unlimited Amazon's "unlimited" cloud storage plan was, so he uploaded 293 years' worth (2 million gigabytes or 2 petabytes) of 😃😃😃 videos to his account.

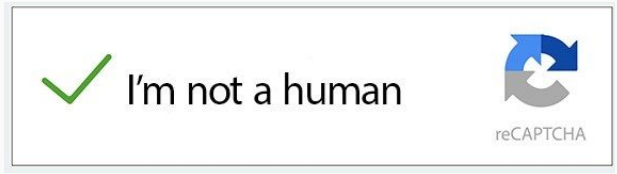


More Reference

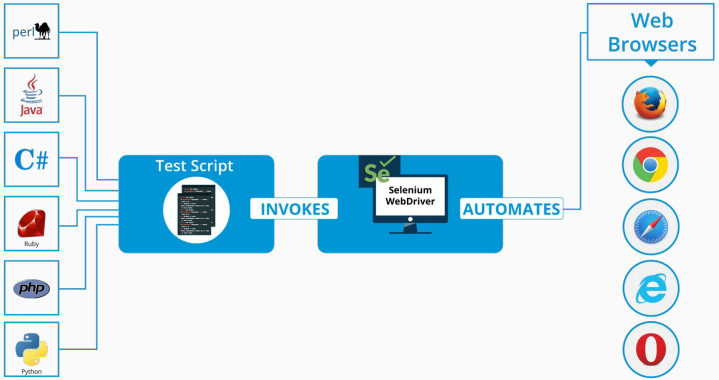
- scrapy



- reCAPTCHA (ref)

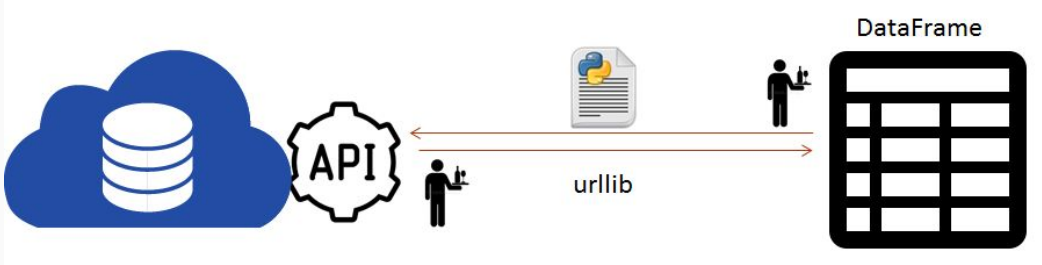


- Selenium (Ref)



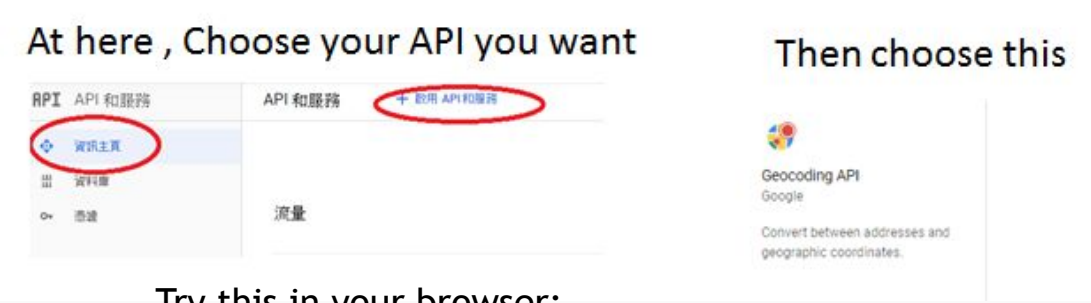
API

- Application programming interface
  https://youtu.be/zvKadd9Cflc



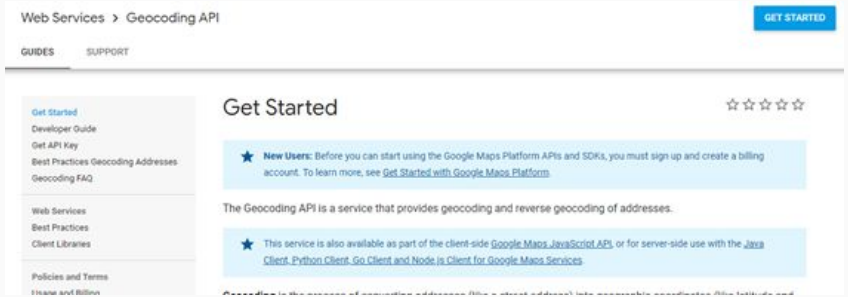
- Google Map API
  - Get API key (Reference)
  - At here , Create the credential = your API Key (time limit = 24 hours)

API



Try this in your browser:
https://maps.googleapis.com/maps/api/geocode/json?address=Lady+Shaw+Building%2C+Hong+Kong&key=YOUR_API_KEY

- The API documentation



- Usage and Billing

$200 USD Google Maps Platform credit is available each month

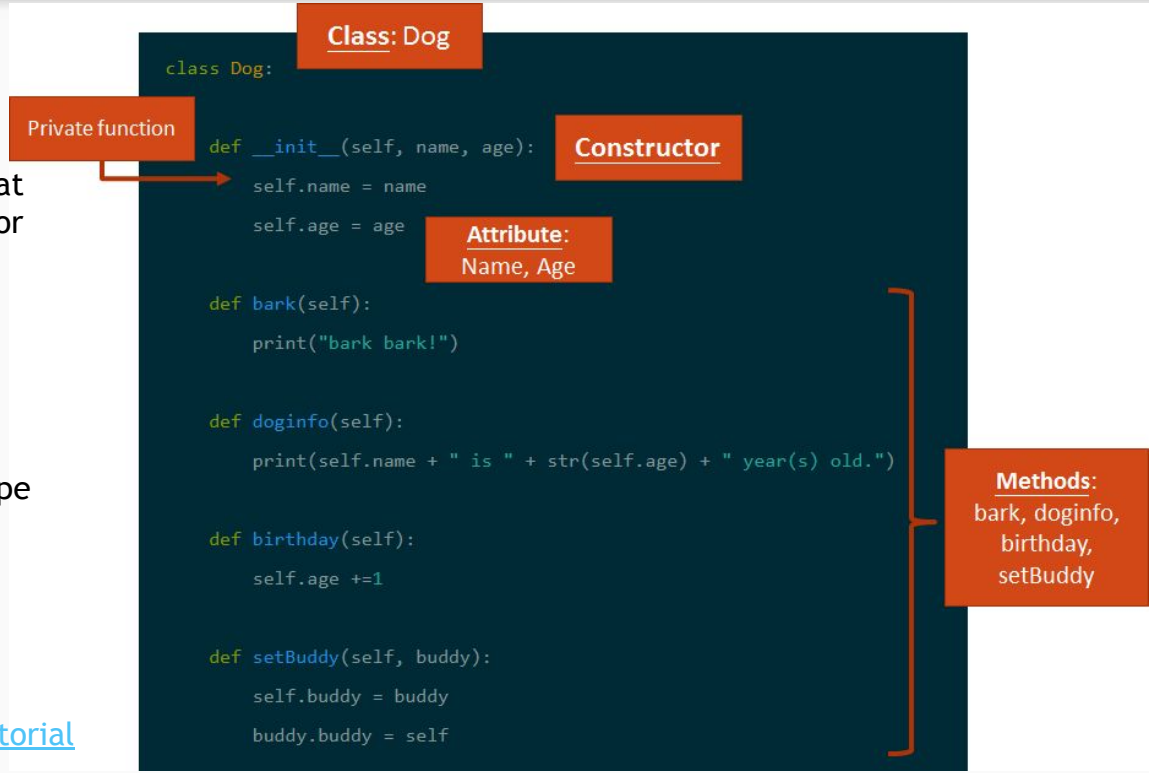
MONTHLY VOLUME RANGE (Price per REQUEST)		
0–100,000	100,001–500,000	500,000+
0.005 USD per each (5.00 USD per 1000)	0.004 USD per each (4.00 USD per 1000)	Contact Sales for volume pricing

Start Coding...

Please access Week 5 - Google API Example...

Object-Oriented Programming

- **Object-oriented programming (OOP)** is a programming paradigm based on the concept of "objects",
- **Classes** – the definitions for the data format and available procedures for a given type or class of object
- **Objects**, instances of classes, may contain
  - data, in the form of fields, often known as *attributes*
  - code, in the form of procedures, often known as *methods*
- Objects are created by calling a special type of method in the class known as a ***constructor***.



Reference:

- https://www.py4e.com/lessons/Objects
- https://www.datacamp.com/community/tutorials/python-oop-tutorial
- https://ithelp.ithome.com.tw/articles/10161285
- http://teddy-chen-tw.blogspot.com/2012/01/2object-class-instance.html

API

[Data.gov.hk]

[CSDI]

To be continue…

if we have no time for API...