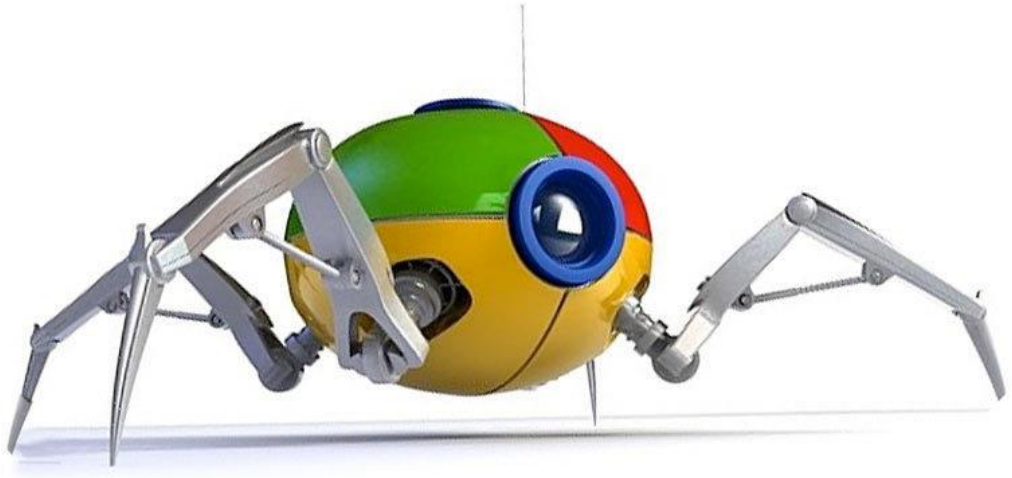
Python Crawler

Outline

- HTML Introduction
  - html/css/javascript(jquery)
- Crawler Basic
  - python request module
  - beautifulsoup + regular expression
- Practical issues
  - try/except and max retry mechanism
  - CAPTCHA solver
  - crawler automation
- Brief DB introduction
  - database(sqlite) manipulation

What's Crawler

- A crawler can automatically retrieve target information on Internet
  - Grab all hotel tickets and choose the cheapest one
  - Grab financial information to make better buy/sell decision



Crawl Overview

- Observe HTML file
- Crawl target information
- Store data
- Analysis data
- Predict

HTML Introduction

Common browsers

What human see on a webpage

What human see on a webpage

What computer see on a webpage

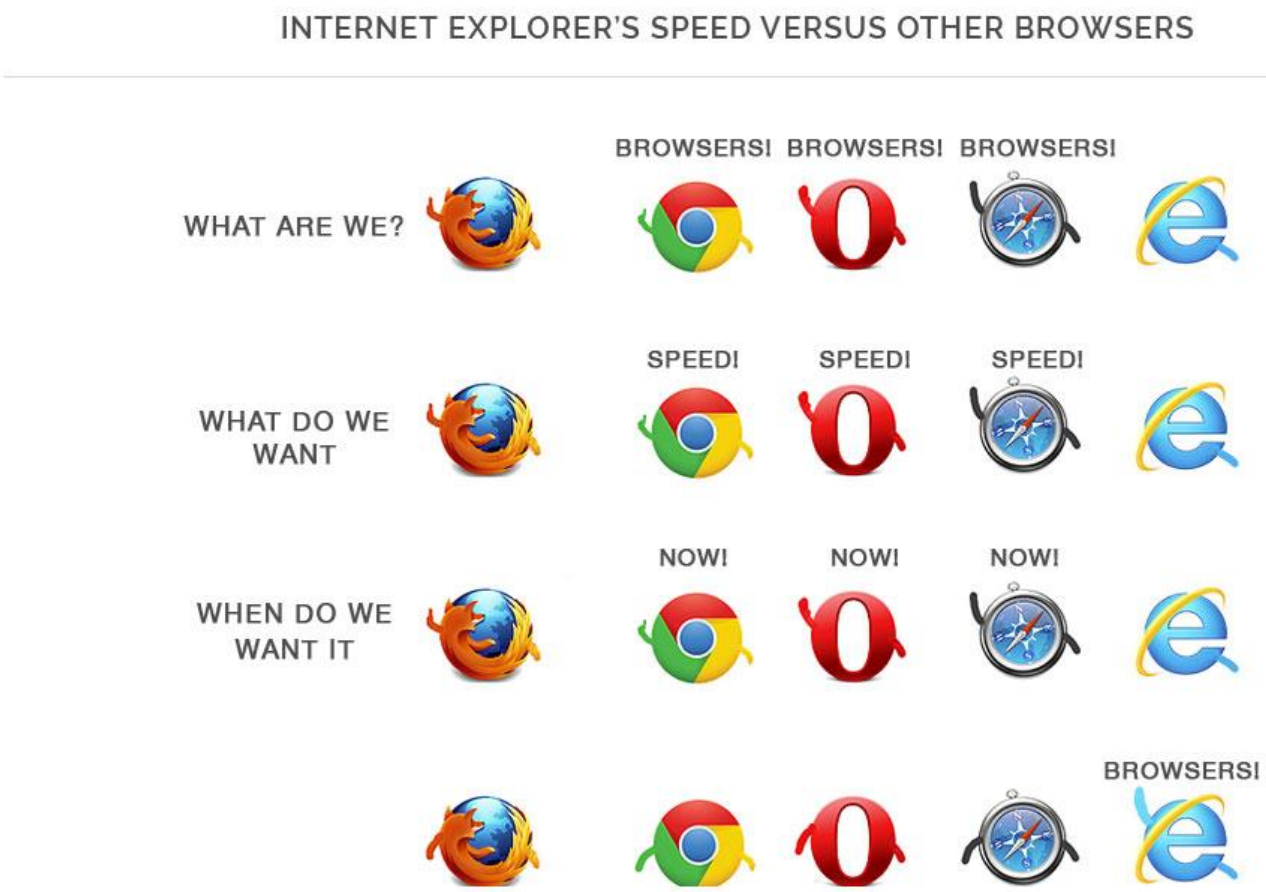
```
<title>學術單位 - 國立臺灣大學</title>
</head>
<body>
<header data-collapse="off" id="Mheader">
  <noscript>
  <div id="topbar">
    <div class="container">
      <div id="student">
        <ul>
        <div class="accesskey"><a accesskey="U" title="上方導覽連結區" href="#">:::</a></div>
          <li><a href="../sitemap.html" title="網站導覽">網站導覽</a></li>
          <li><a href="http://reg.aca.ntu.edu.tw/newstu" target="_blank" title="新生(另開視窗)">新生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=5,6" target="_blank" title="在校學生(另開視窗)">在校學生</a></li>
          <li><a href="http://www.oia.ntu.edu.tw" target="_blank" title="國際生(另開視窗)">國際生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=1,2,4" target="_blank" title="教職員(另開視窗)">教職員</a></li>
          <li><a href="http://visitorcenter.cloud.ntu.edu.tw/" target="_blank" title="訪客(另開視窗)">訪客</a></li>
          <li><a href="http://homepage.ntu.edu.tw/~ntualumni/" target="_blank" title="校友(另開視窗)">校友</a></li>
        </ul>
      </div>
      <div id="search">
        <ul>
          <li><a href="http://www.ntu.edu.tw/oldchinese/" target="_blank" title="回舊站(另開視窗)">回舊站</a></li>
          <li><a href="../contact.html" title="聯絡我們">聯絡我們</a></li>
          <!--<li>站內搜尋</li>-->
          <li>
            <form action="http://www.google.com/cse" id="cse-search-box" style="width: 200px;">
              <label for="keyword">站內搜尋：</label>
              <input type="hidden" name="cx" value="011987122760880416627:zsghqazzvk4">
              <input type="hidden" name="ie" value="utf-8">
              <input id="searchInput" type="text" value="請輸入關鍵字" onkeypress="if(this.value=='請輸入關鍵字'){this.style.color='';this.value='';}" onclick="if(this.value=='請輸入關鍵字')
{this.style.color='';this.value='';}" onblur="if(!this.value.length){this.style.color='#999999';this.value='請輸入關鍵字';}" maxlength="2048" name="q">
              <input type="image" alt="開始搜尋" title="開始搜尋" src="../images/search.png" name="go" class="searchBtn">
              <input type="hidden" value="www.google.com/cse/home?cx=011987122760880416627:zsghqazzvk4" name="siteurl">
              <input type="hidden" value="www.google.com/cse/panel/basics?cx=011987122760880416627:zsghqazzvk4&amp;sig=__ojBF1PpI1CTk4OCeWf1DlJ4PyGE=" name="ref">
            </form>
          </li>
        </ul>
      </div>
      <div id="webType">
        <ul>
          <li class="mobileWebType2"><a id="EngLink" href="../english/academics/academics.html" title="English">English</a></li>
        </ul>
```

What computer see on a webpage

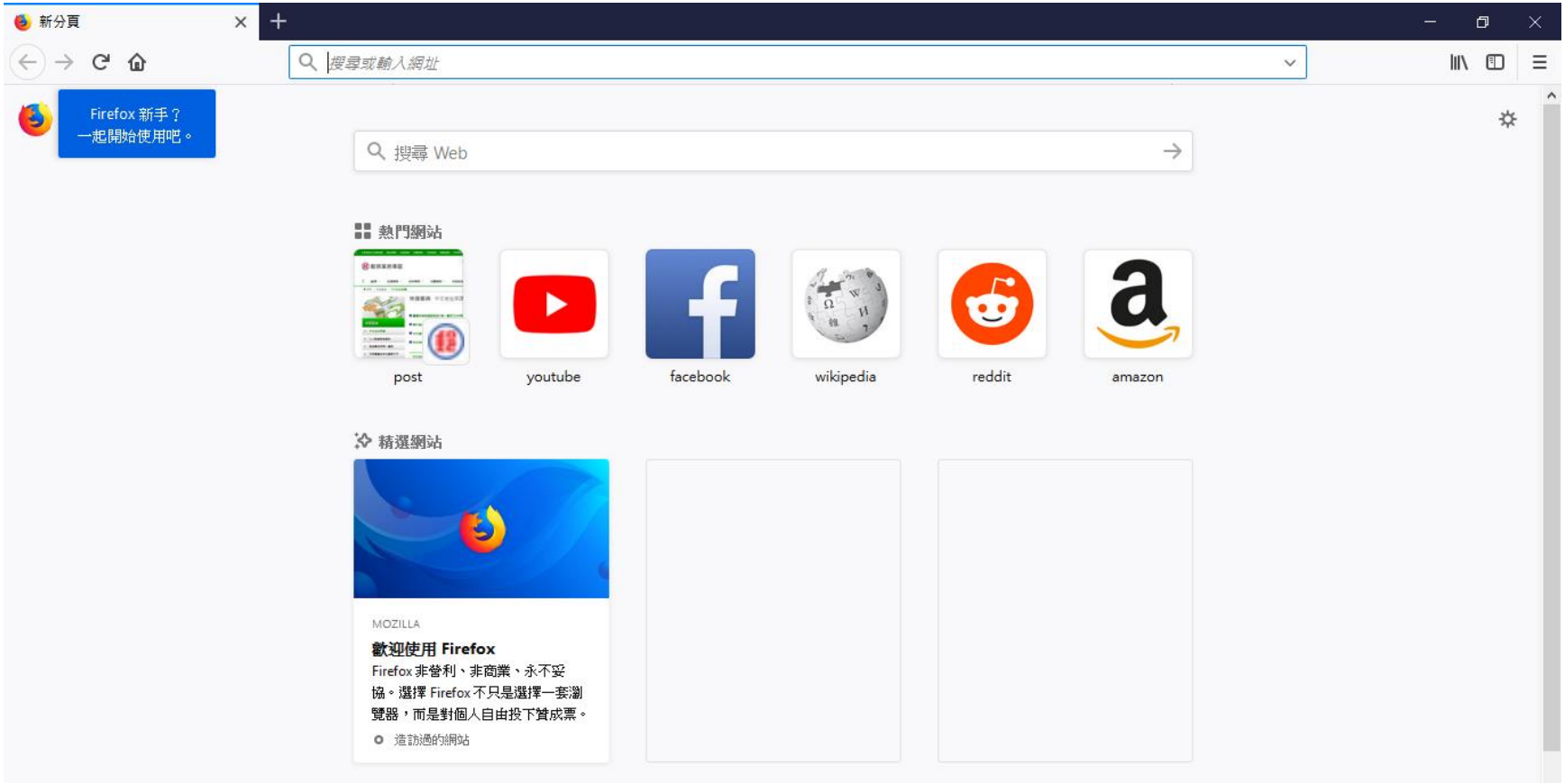
```
<title>學術單位 - 國立臺灣大學</title>
</head>
<body>
<header data-collapse="off" id="Mheader">
  <noscript>
  <div id="topbar">
    <div class="container">
      <div id="student">
        <ul>
        <div class="accesskey"><a accesskey="U" title="上方導覽連結區" href="#">:::</a></div>
          <li><a href="../sitemap.html" title="網站導覽">網站導覽</a></li>
          <li><a href="http://reg.aca.ntu.edu.tw/newstu" target="_blank" title="新生(另開視窗)">新生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=5,6" target="_blank" title="在校學生(另開視窗)">在校學生</a></li>
          <li><a href="http://www.oia.ntu.edu.tw" target="_blank" title="國際生(另開視窗)">國際生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=1,2,4" target="_blank" title="教職員(另開視窗)">教職員</a></li>
          <li><a href="http://visitorcenter.cloud.ntu.edu.tw/" target="_blank" title="訪客(另開視窗)">訪客</a></li>
          <li><a href="http://homepage.ntu.edu.tw/~ntualumni/" target="_blank" title="校友(另開視窗)">校友</a></li>
        </ul>
      </div>
      <div id="search">
        <ul>
          <li><a href="http://www.ntu.edu.tw/oldchinese/" target="_blank" title="回舊站(另開視窗)">回舊站</a></li>
          <li><a href="../contact.html" title="聯絡我們">聯絡我們</a></li>
          <!--<li>站內搜尋</li>-->
          <li>
            <form action="http://www.google.com/cse" id="cse-search-box" style="width: 200px;">
              <label for="keyword">站內搜尋：</label>
              <input type="hidden" name="cx" value="011987122760880416627:zsghqazzvk4">
              <input type="hidden" name="ie" value="utf-8">
              <input id="searchInput" type="text" value="請輸入關鍵字" onkeypress="if(this.value=='請輸入關鍵字'){this.style.color='';this.value='';}" onclick="if(this.value=='請輸入關鍵字')
{this.style.color='';this.value='';}" onblur="if(!this.value.length){this.style.color='#999999';this.value='請輸入關鍵字';}" maxlength="2048" name="q">
              <input type="image" alt="開始搜尋" title="開始搜尋" src="../images/search.png" name="go" class="searchBtn">
              <input type="hidden" value="www.google.com/cse/home?cx=011987122760880416627:zsghqazzvk4" name="siteurl">
              <input type="hidden" value="www.google.com/cse/panel/basics?cx=011987122760880416627:zsghqazzvk4&amp;sig=__ojBF1PpI1CTk4OCeWf1DlJ4PyGE=" name="ref">
            </form>
          </li>
        </ul>
      </div>
      <div id="webType">
        <ul>
          <li class="mobileWebType2"><a id="EngLink" href="../english/academics/academics.html" title="English">English</a></li>
        </ul>
```

This is why we need to know HTML code……

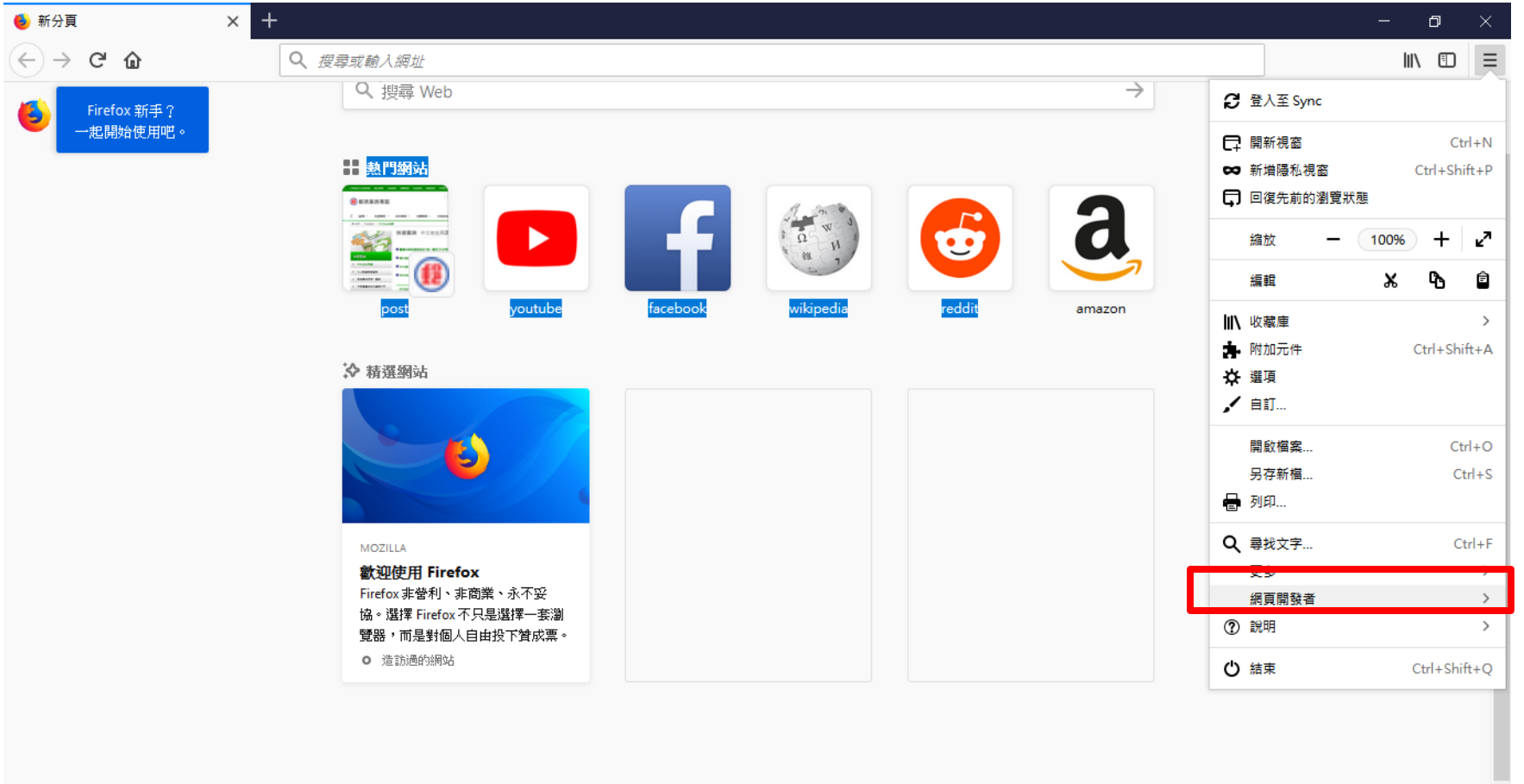
A Joke on IE


INTERNET EXPLORER'S SPEED VERSUS OTHER BROWSERS
WHAT ARE WE?
BROWSERS! BROWSERS! BROWSERS!
WHAT DO WE WANT
SPEED! SPEED! SPEED!
WHEN DO WE WANT IT
NOW! NOW! NOW!
BROWSERS!

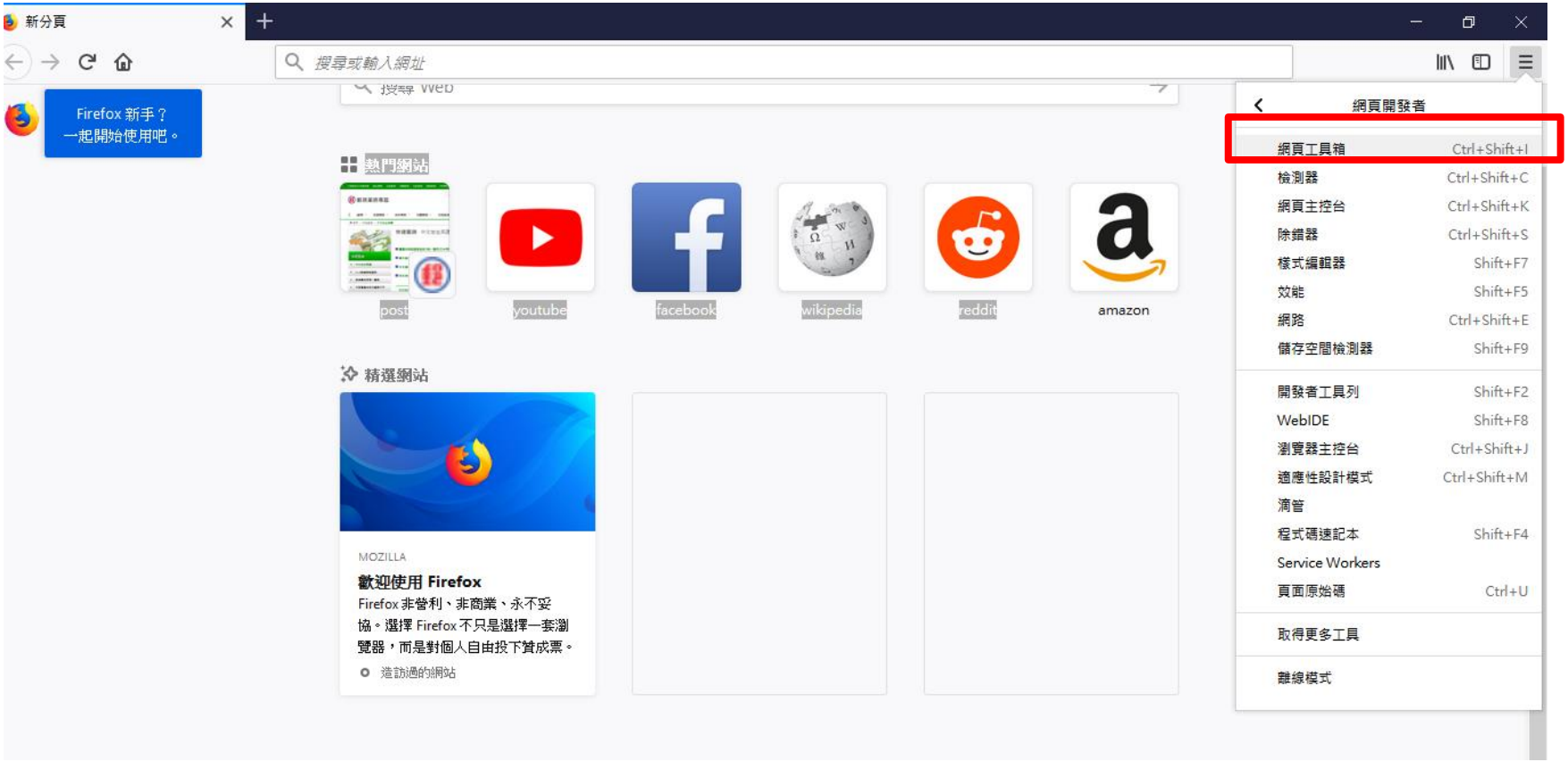
Firefox(recommend)



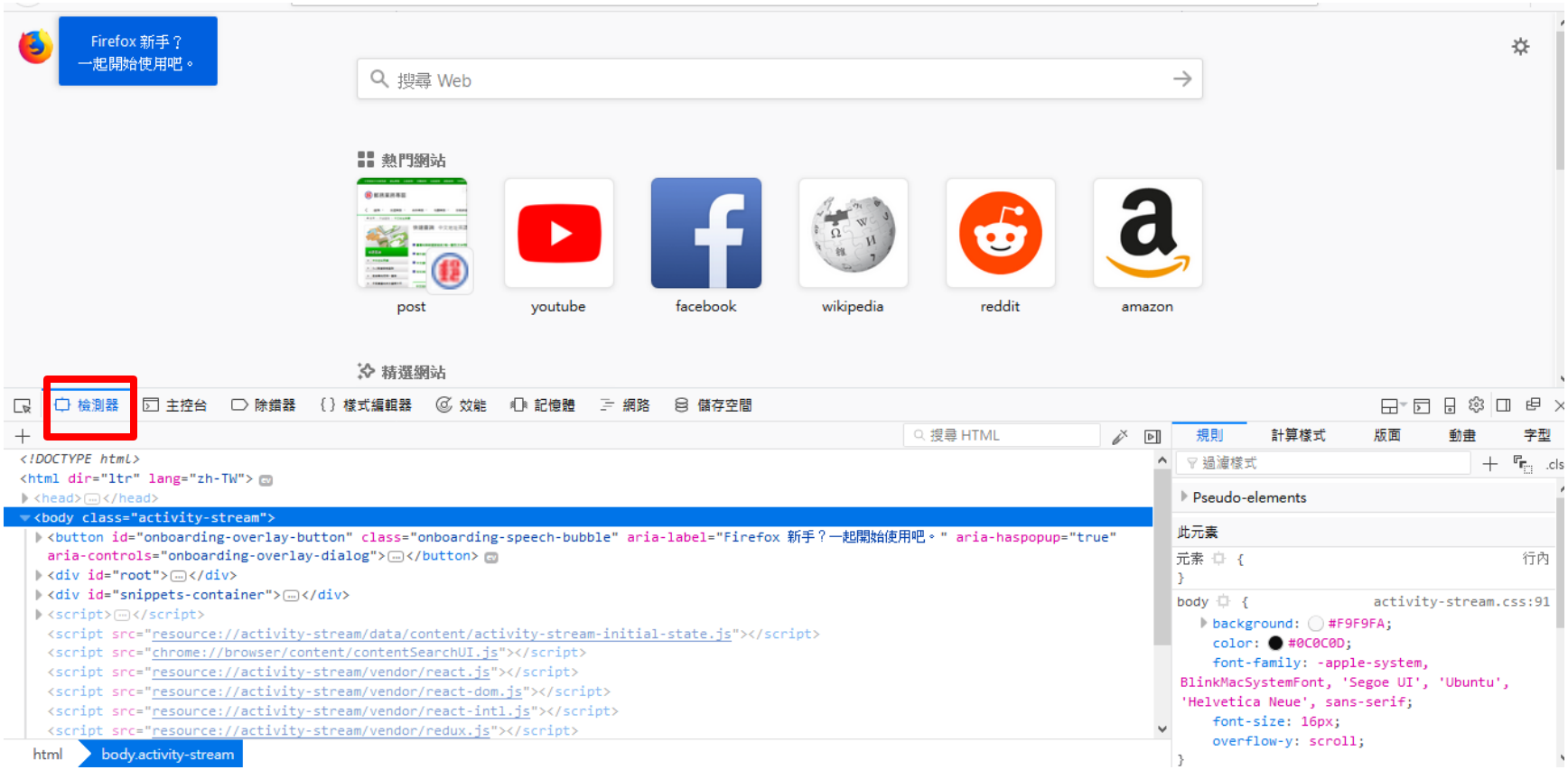
Firefox(recommend)



Firefox(recommend)



Firefox(recommend)



Use “F12” shortcut to open developer tool

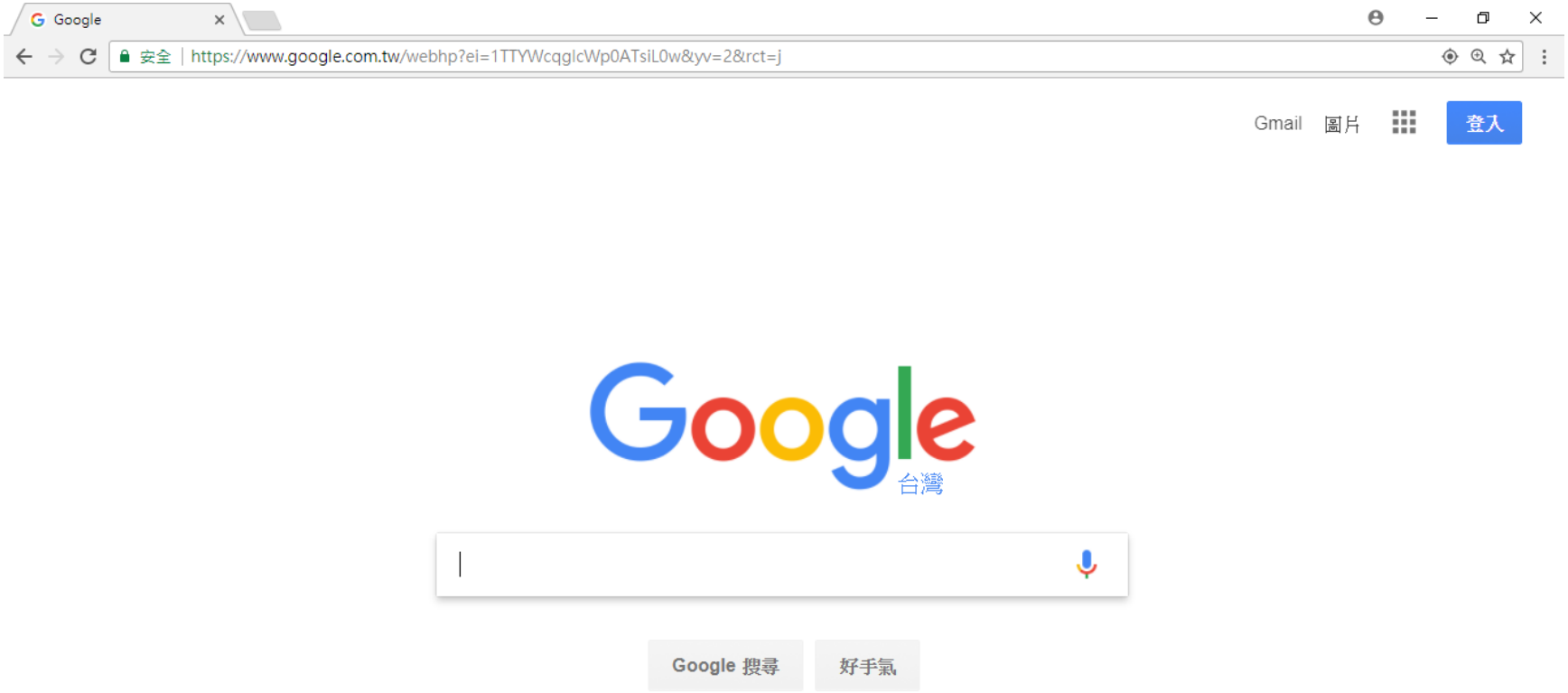
Firefox(recommend)

Click “select” button to observe any elements on the webpage

Firefox(recommend)

Directly click html code can also observe elements in the website

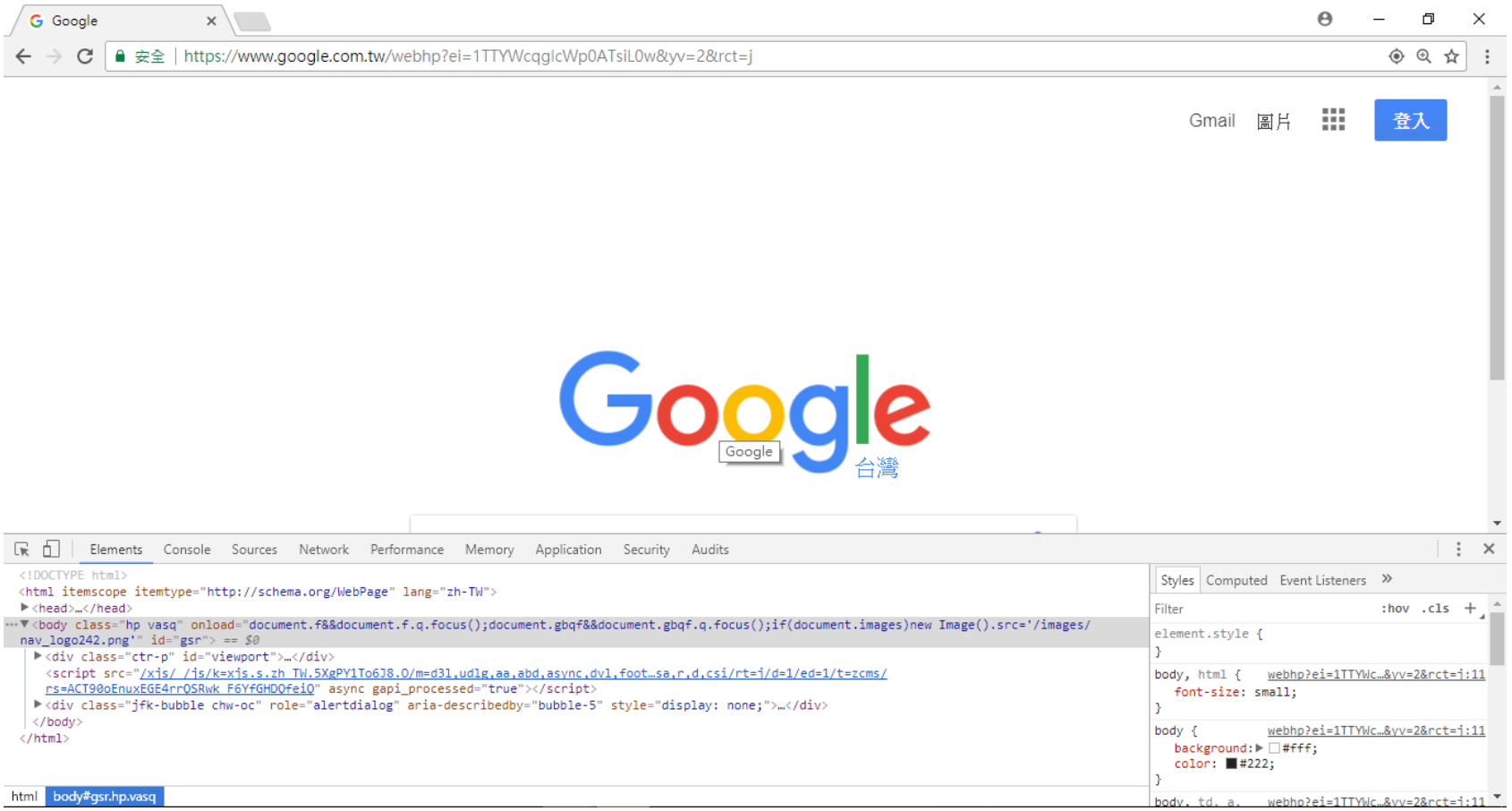
Chrome


Google
安全 | https://www.google.com.tw/webhp?ei=1TTYWcqgIcWp0ATsiL0w&yv=2&rct=j
Gmail
圖片
登入
Google
台灣
Google 搜尋
好手氣

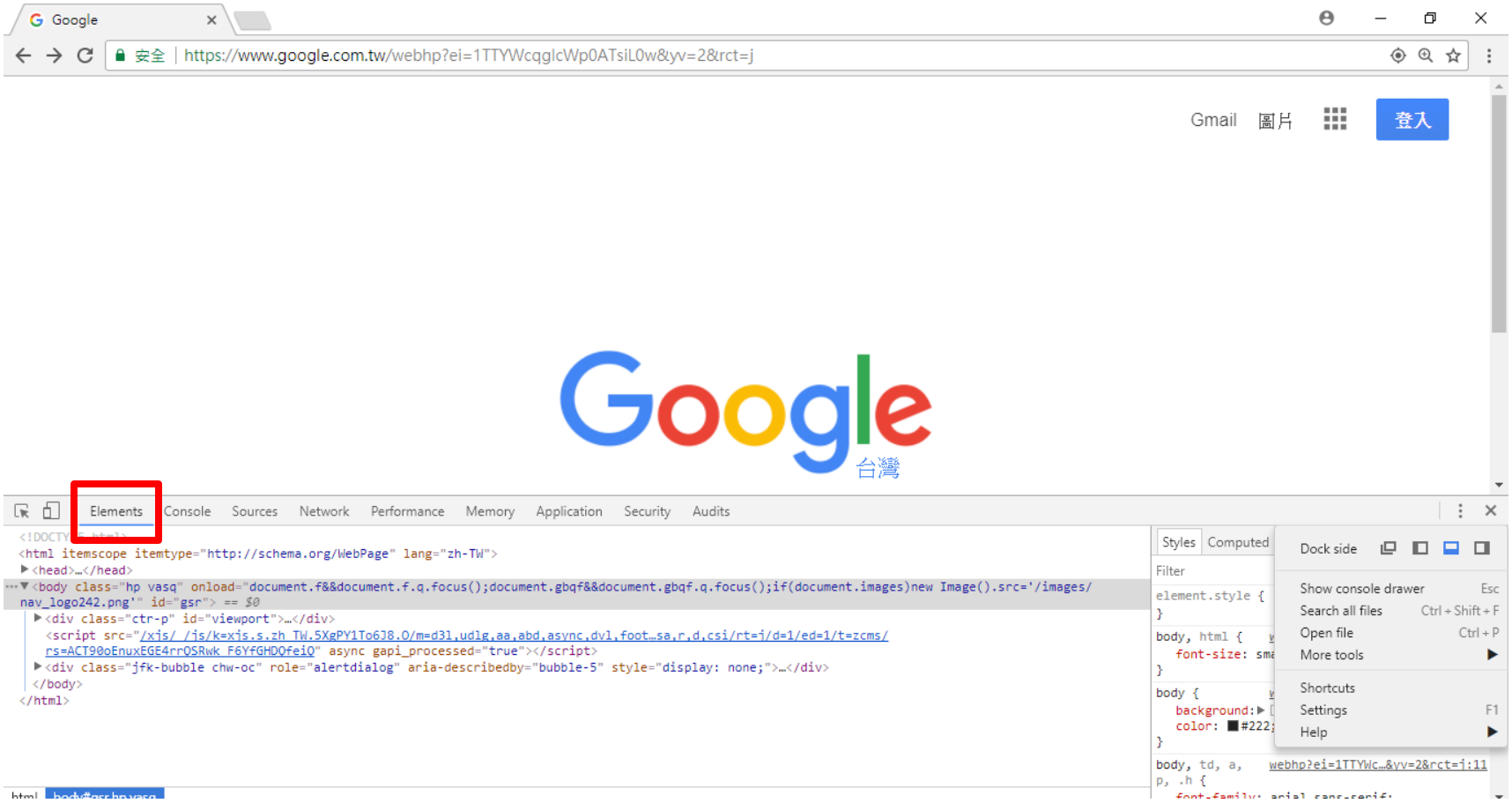
Chrome


Google
安全 | https://www.google.com.tw/webhp?ei=1TTYWcqgIcWp0ATsiL0w&yv=2&rct=j
新增分頁(T) Ctrl+T
新增視窗(N) Ctrl+N
新增無痕式視窗(I) Ctrl+Shift+N
記錄(H)
下載(D) Ctrl+J
書籤(B)
縮放 − 125% +
列印(P)... Ctrl+P
投放(C)...
更多工具(L)
編輯 剪下(T) 複製(C) 貼上(P)
設定(S)
說明(E)
結束(X) Ctrl+Shift+Q
另存網頁為(A)... Ctrl+S
加到桌面...
清除瀏覽資料(C)... Ctrl+Shift+Del
擴充功能(E)
工作管理員(T) Shift+Esc
開發人員工具(D) Ctrl+Shift+I
Google 搜尋
好手氣

Chrome


Google
安全 | https://www.google.com.tw/webhp?ei=1TTYWcqgIcWp0ATsiL0w&yv=2&rct=j
Gmail
圖片
登入
Google
台灣
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Audits
<!DOCTYPE html>
<html itemscope itemtype="http://schema.org/WebPage" lang="zh-TW">
<head>…</head>
<body class="hp vasq" onload="document.f&&document.f.q.focus();document.gbqf&&document.gbqf.q.focus();if(document.images)new Image().src='/images/
nav_logo242.png'" id="gsr"> == $0
…
<script src="/xjs/_/js/k=xjs.s.zh_TW.5XgPY1To6J8.O/m=d31,udlg,aa,abd,async,dvl,foot…sa,r,d,csi/rt=j/d=1/ed=1/t=zcms/
rs=ACT90oEnuxEGE4rrOSRwk_F6YfGHDQfeiQ" async gapi_processed="true"></script>
…
</body>
</html>
html
body#gsr.hp.vasq
Styles
Computed
Event Listeners
Filter
:hov .cls
element.style {
}
body, html {
font-size: small;
}
body {
background: #fff;
color: #222;
}

Chrome



Use “F12” shortcut to open developer tool

Example and Exercise (5 min)

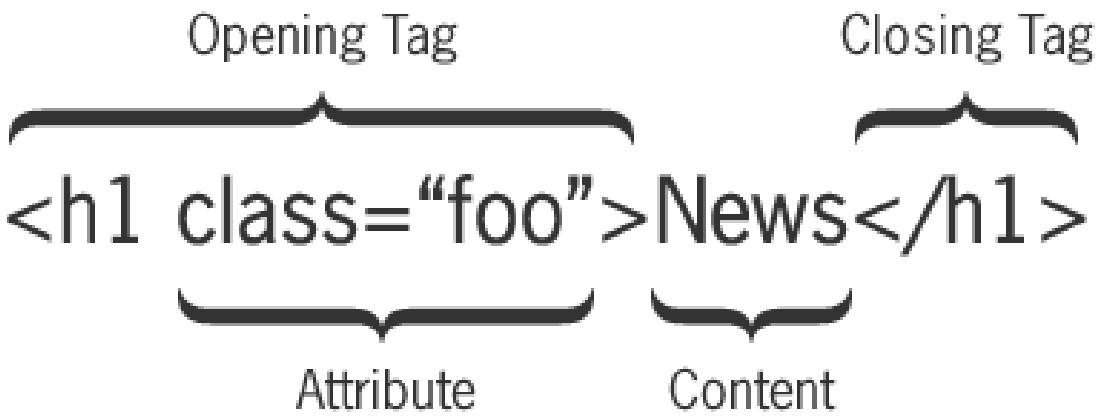
- Please install Firefox
  - https://www.mozilla.org/zh-TW/firefox/new/
- Please observe different elements in different web
  - http://www.tse.com.tw/zh/
  - http://rate.bot.com.tw/xrt?Lang=zh-TW
  - http://www.cnyes.com/usastock/

Modern web

- Web consist of three different part
  - **HTML** define the content of web pages
  - **CSS** specify the layout of web pages
  - **JavaScript** program the behavior of web pages

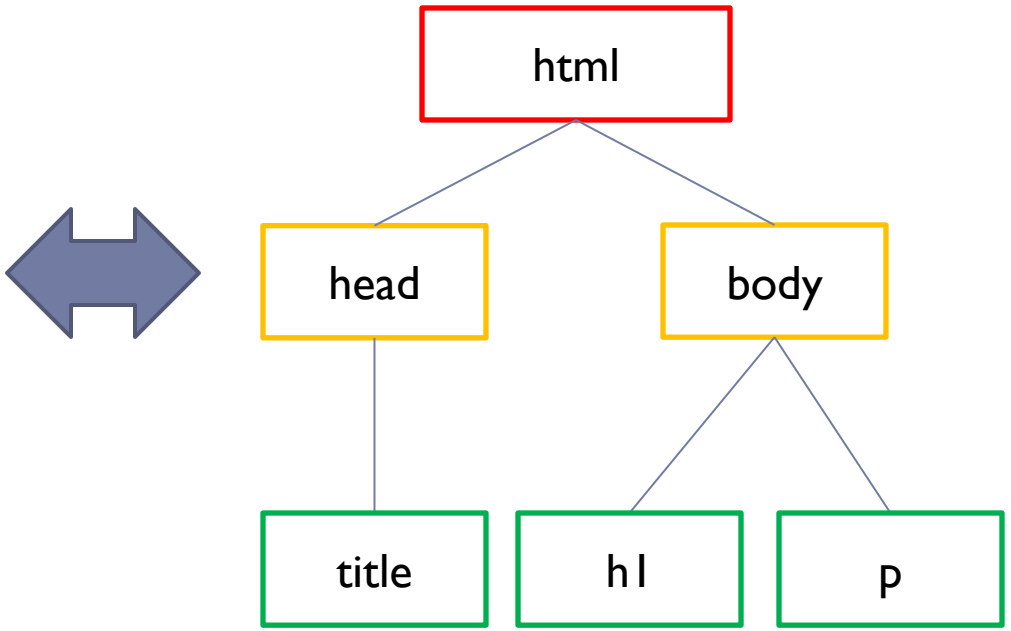
What's HTML

- **H**yper**T**ext **M**arkup **L**anguage
  - Consist of many elements
- Each elements contain opening/closing tag, attribute and its content



Html Concept

```
<!DOCTYPE html>
<html>
  <head>
    <title>hello html</title>
  </head>
  <body>
    <h1>This is h1 tag</h1>
    <p>This is p tag</p>
  </body>
</html>
```



Common tags

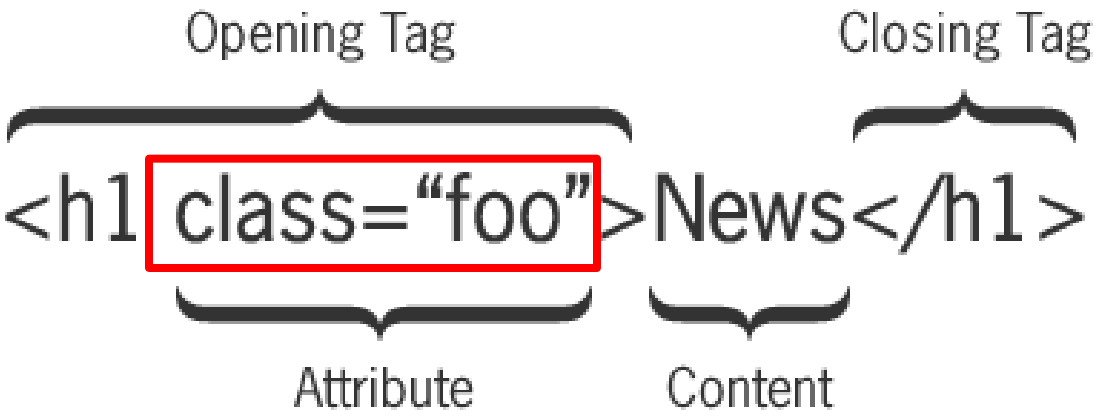
Tag Name	Purpose
<h1> - <h6>	Title
<p>	Paragraph
<a>	Hyper link
<table>	Table
<tr>	Row in table
<td>	Cell in table
<img>	image
</br>	Line break(no end tag)

Example and Exercise (6 min)

- html hello word example
  - example\html_css_js\hello_html.html
- html common tags
  - example\html_css_js\common_tags.html
- If you would like to know more tags, please reference
  - https://www.w3schools.com/Tags/default.asp
- Please try to add any tags/change any text and refresh your browser (F5)
- Please google "textarea" tag and add it to your file

HTML Attributes

- You can describe attributes on each tags
  - Usually, it is inside each start tag
- One tag can have multiple attributes



Common Attributes

Attribute Name	Purpose
Id	Id of a tag (can not duplicate in a file)
class	Class of a tag(can not duplicate in a file)
href	Hyper link
src	path of image

Example and Exercise (3 min)

- html common attributes
  - example\html_css_js\common_attributes.html
- If you would like to know more attributes, please reference
  - https://www.w3schools.com/html/html_attributes.asp
- Please try to add/modify any attributes

CSS Introduction

- Cascading Style Sheets
  - how HTML elements are to be displayed (color, layout, font size, ……)
  - Usually use "style" tag
  - Some save external stylesheets as CSS files

Example and Exercise (3 min)

- css example
  - example\html_css_js\css_example.html
- External css file example
  - example\html_css_js\external_css_example.html
- If you would like to know more css style, please reference
  - https://www.w3schools.com/css/
- Please try to add/modify any css style

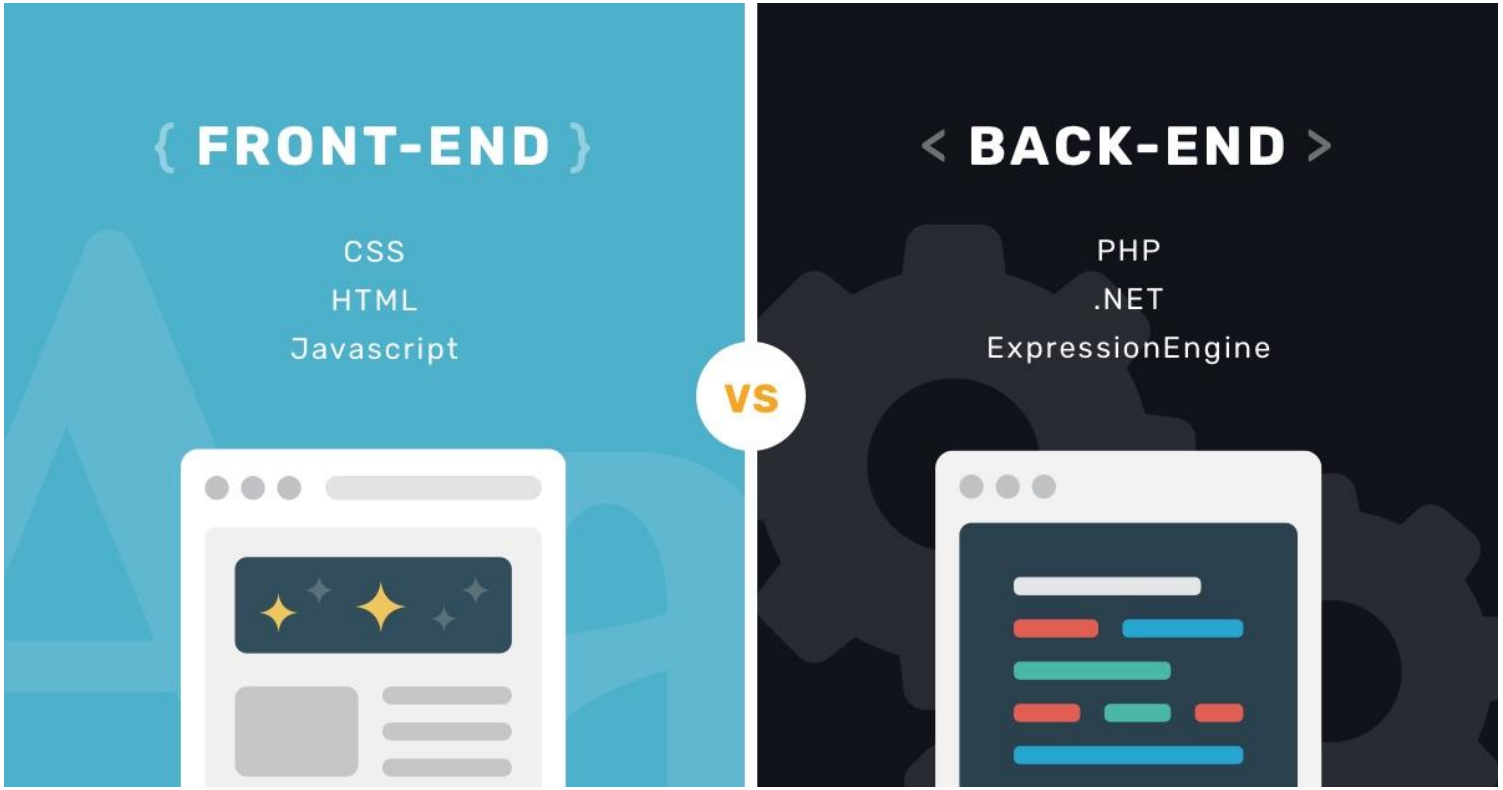
Introduction to Javascript

- JavaScript is the programming language of HTML and the web
- It make webpages interactive
- jQuery is a fast, small, and feature-rich JavaScript library
  - Usually be used by web developers because of it is more elegant

Example and Exercise (3 min)

- jquery example
  - example\html_css_js\jquery_example.html
- External jquery file example
  - example\html_css_js\external_jquery_example.html
- If you would like to know more jquery, please reference
  - https://www.w3schools.com/jquery/

Front-End and Back-End



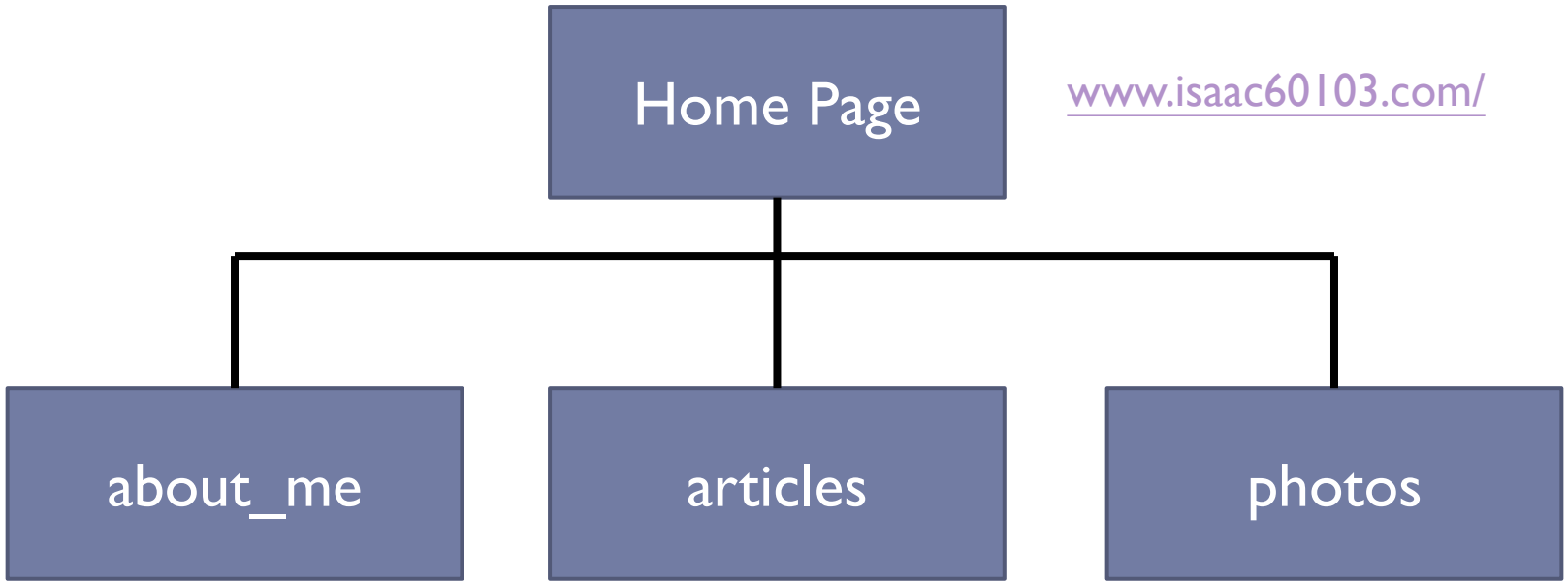
Crawler Introduction

Relative url and absolute url

- A Uniform Resource Locator (URL) is also called web address
- Absolute url
  - http://www.ntu.edu.tw/highlights/2017/he20171024.html
- Relative url
  - highlights/2017/he20171024.html

We should use absolute url when crawling!

Web Structure



www.isaac60103.com/

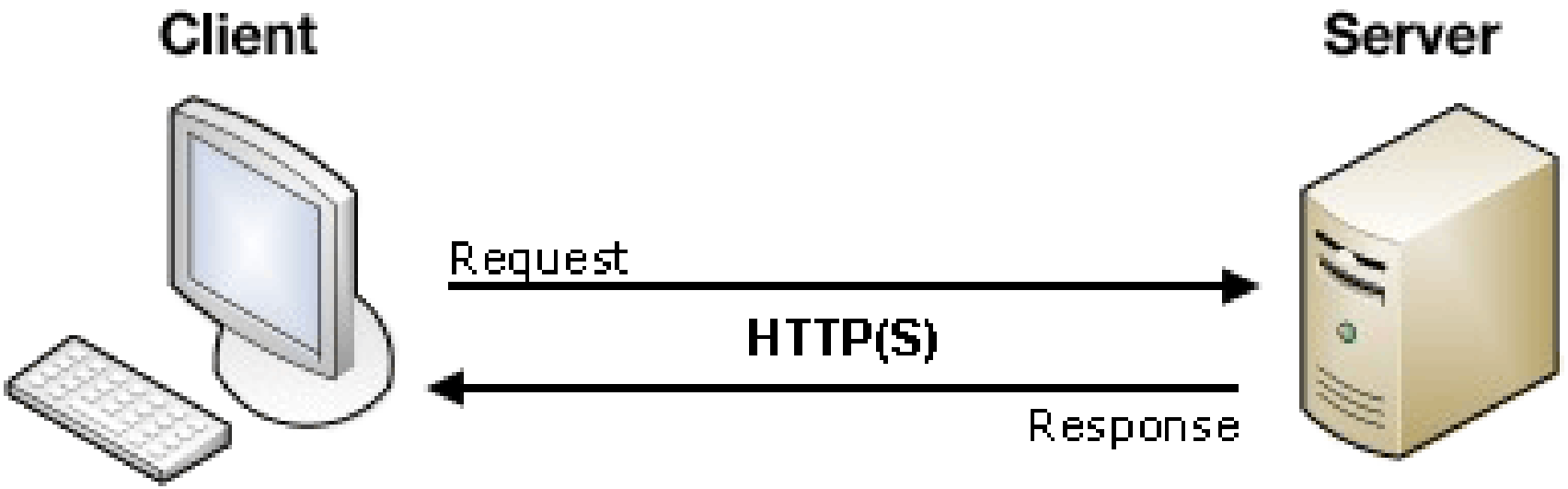
www.isaac60103.com/**about_me** www.isaac60103.com/**articles** www.isaac60103.com/**photos**

Relative url

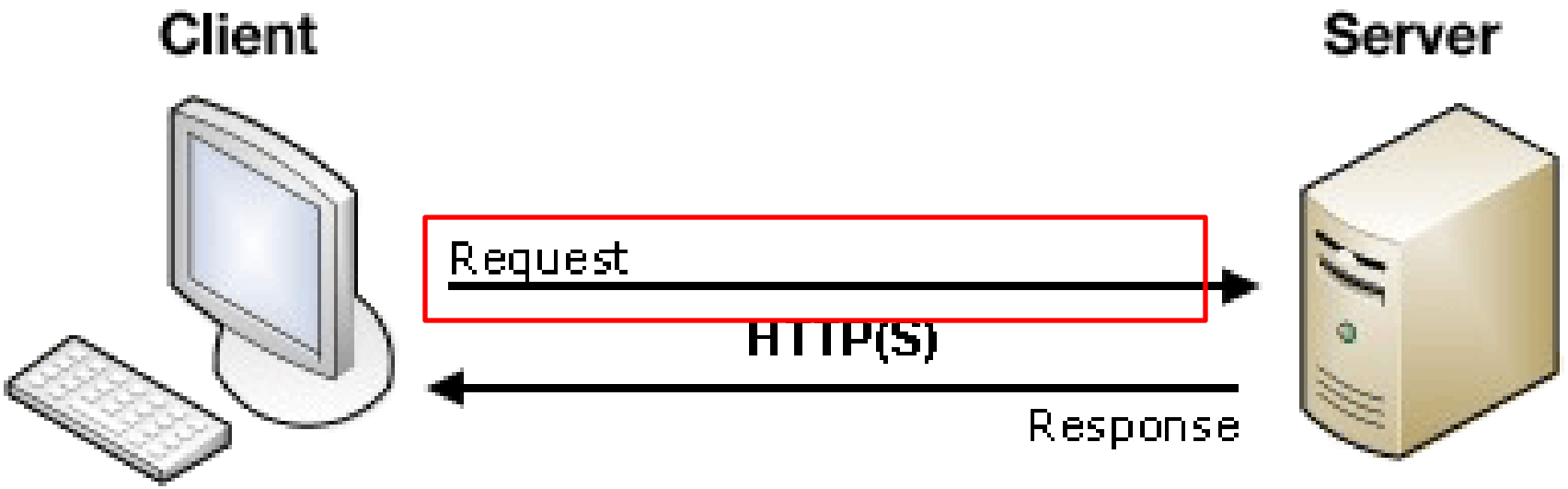
Example and Exercise

- Relative url on web page
  - https://www.isaac60103.com/

How to get html from server



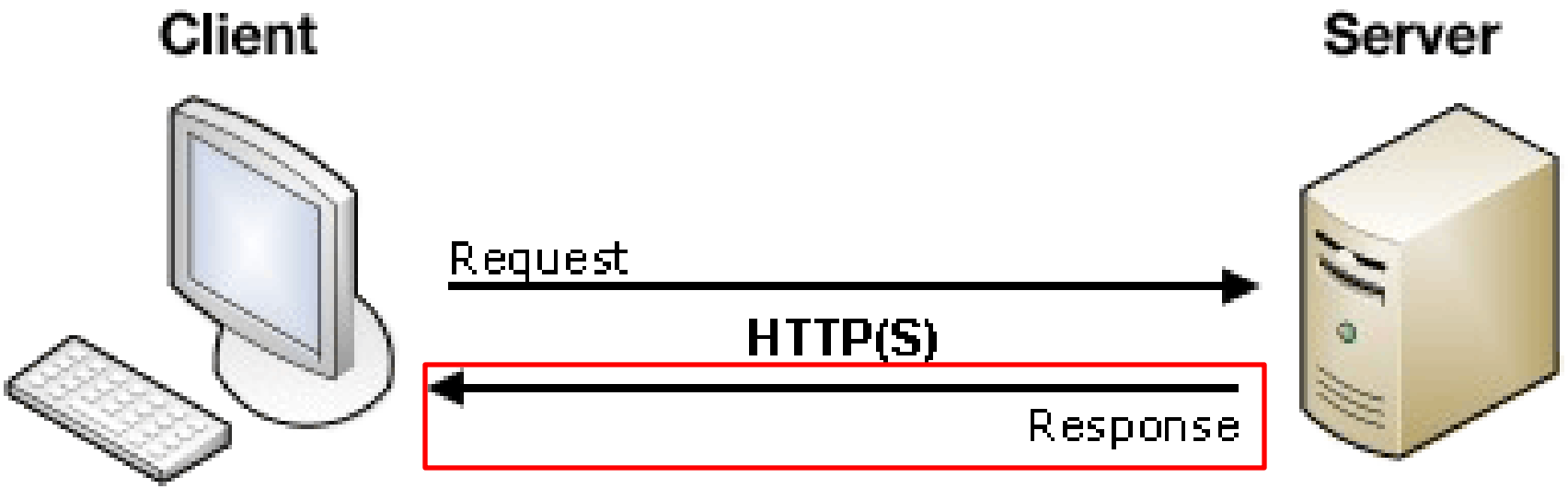
How to get html from server



Usually request method:

- Get
- Post

How to get html from server



Response:

- Status code
- Html file

Status code

Status code	Meaning
200	OK
403	Forbidden
404	Not found
500	Internal Server Error
503	Service Unavailable
504	Gateway Timeout

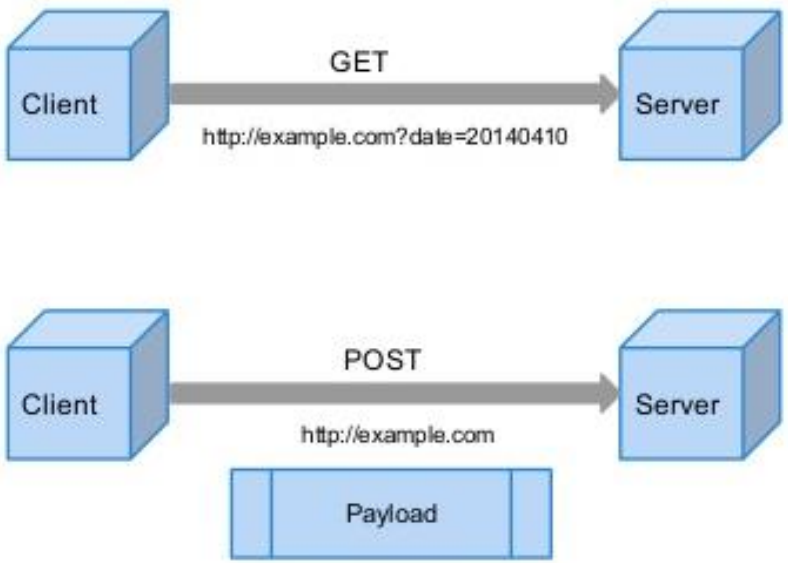
Html file

- http://www.ntu.edu.tw/

```
<title>學術單位 - 國立臺灣大學</title>
</head>
<body>
<header data-collapse="off" id="Mheader">
  <noscript>
  <div id="topbar">
    <div class="container">
      <div id="student">
        <ul>
        <div class="accesskey"><a accesskey="U" title="上方導覽連結區" href="#">:::</a></div>
          <li><a href="../sitemap.html" title="網站導覽">網站導覽</a></li>
          <li><a href="http://reg.aca.ntu.edu.tw/newstu" target="_blank" title="新生(另開視窗)">新生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=5,6" target="_blank" title="在校學生(另開視窗)">在校學生</a></li>
          <li><a href="http://www.oia.ntu.edu.tw" target="_blank" title="國際生(另開視窗)">國際生</a></li>
          <li><a href="https://my.ntu.edu.tw/?block=1,2,4" target="_blank" title="教職員(另開視窗)">教職員</a></li>
          <li><a href="http://visitorcenter.cloud.ntu.edu.tw/" target="_blank" title="訪客(另開視窗)">訪客</a></li>
          <li><a href="http://homepage.ntu.edu.tw/~ntualumni/" target="_blank" title="校友(另開視窗)">校友</a></li>
        </ul>
      </div>
      <div id="search">
        <ul>
          <li><a href="http://www.ntu.edu.tw/oldchinese/" target="_blank" title="回舊站(另開視窗)">回舊站</a></li>
          <li><a href="../contact.html" title="聯絡我們">聯絡我們</a></li>
          <!--<li>站內搜尋</li>-->
          <li>
            <form action="http://www.google.com/cse" id="cse-search-box" style="width: 200px;">
              <label for="keyword">站內搜尋：</label>
              <input type="hidden" name="cx" value="011987122760880416627:zsghqazzvk4">
              <input type="hidden" name="ie" value="utf-8">
              <input id="searchInput" type="text" value="請輸入關鍵字" onkeypress="if(this.value=='請輸入關鍵字'){this.style.color='';this.value='';}" onclick="if(this.value=='請輸入關鍵字')
{this.style.color='';this.value='';}" onblur="if(!this.value.length){this.style.color='#999999';this.value='請輸入關鍵字';}" maxlength="2048" name="q">
              <input type="image" alt="開始搜尋" title="開始搜尋" src="../images/search.png" name="go" class="searchBtn">
              <input type="hidden" value="www.google.com/cse/home?cx=011987122760880416627:zsghqazzvk4" name="siteurl">
              <input type="hidden" value="www.google.com/cse/panel/basics?cx=011987122760880416627:zsghqazzvk4&amp;sig=__ojBF1PpI1CTk4OCeWf1DlJ4PyGE=" name="ref">
            </form>
          </li>
        </ul>
      </div>
      <div id="webType">
        <ul>
          <li class="mobileWebType2"><a id="EngLink" href="../english/academics/academics.html" title="English">English</a></li>
        </ul>
```

GET V.S. POST

- GET
  - URL would change when browsing different content
- POST
  - URL would not change but content in web page would change under different user's requests
  - Submit private data to server (FB login ……) usually use post



Common web scraping libraries

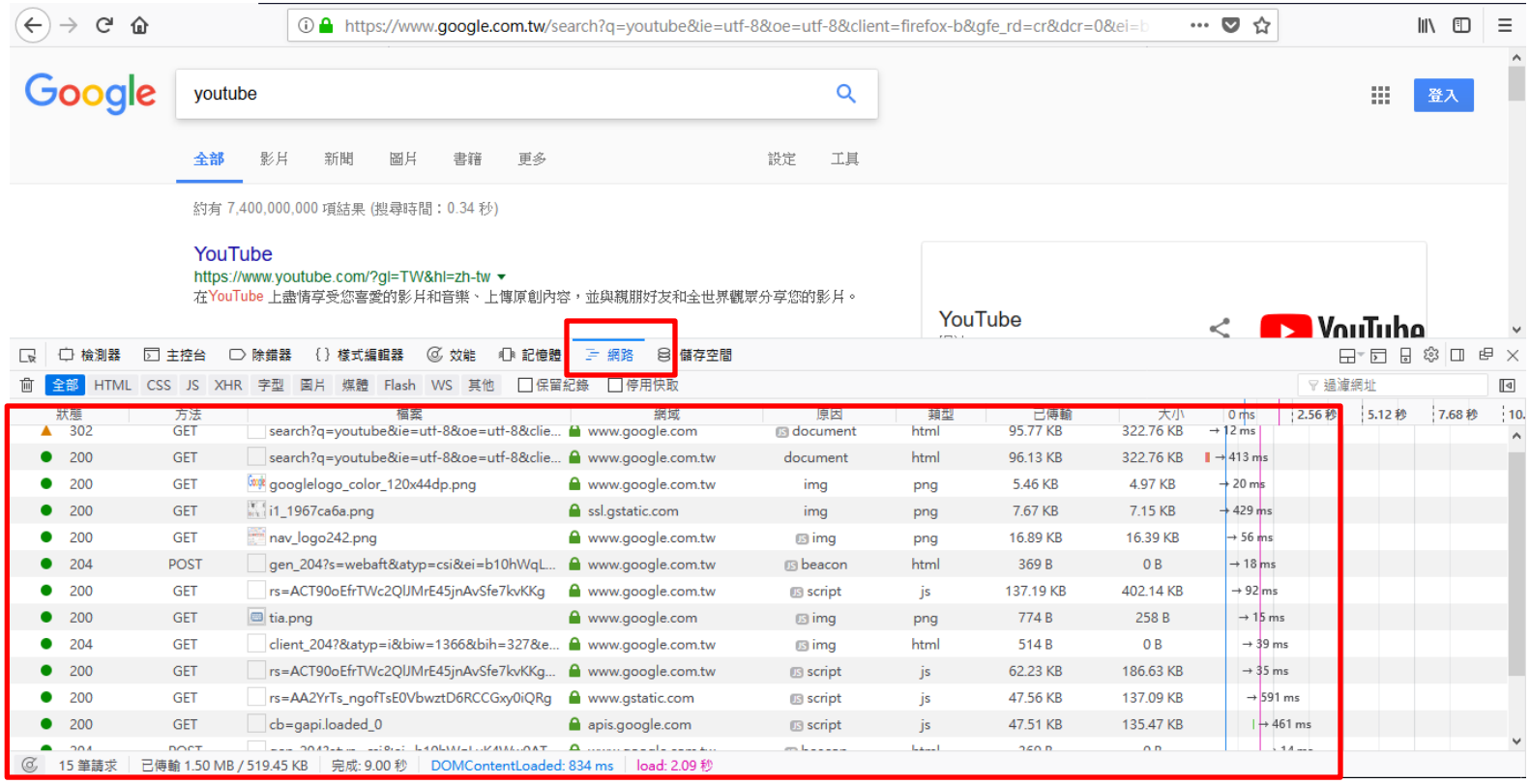
requests

selenium

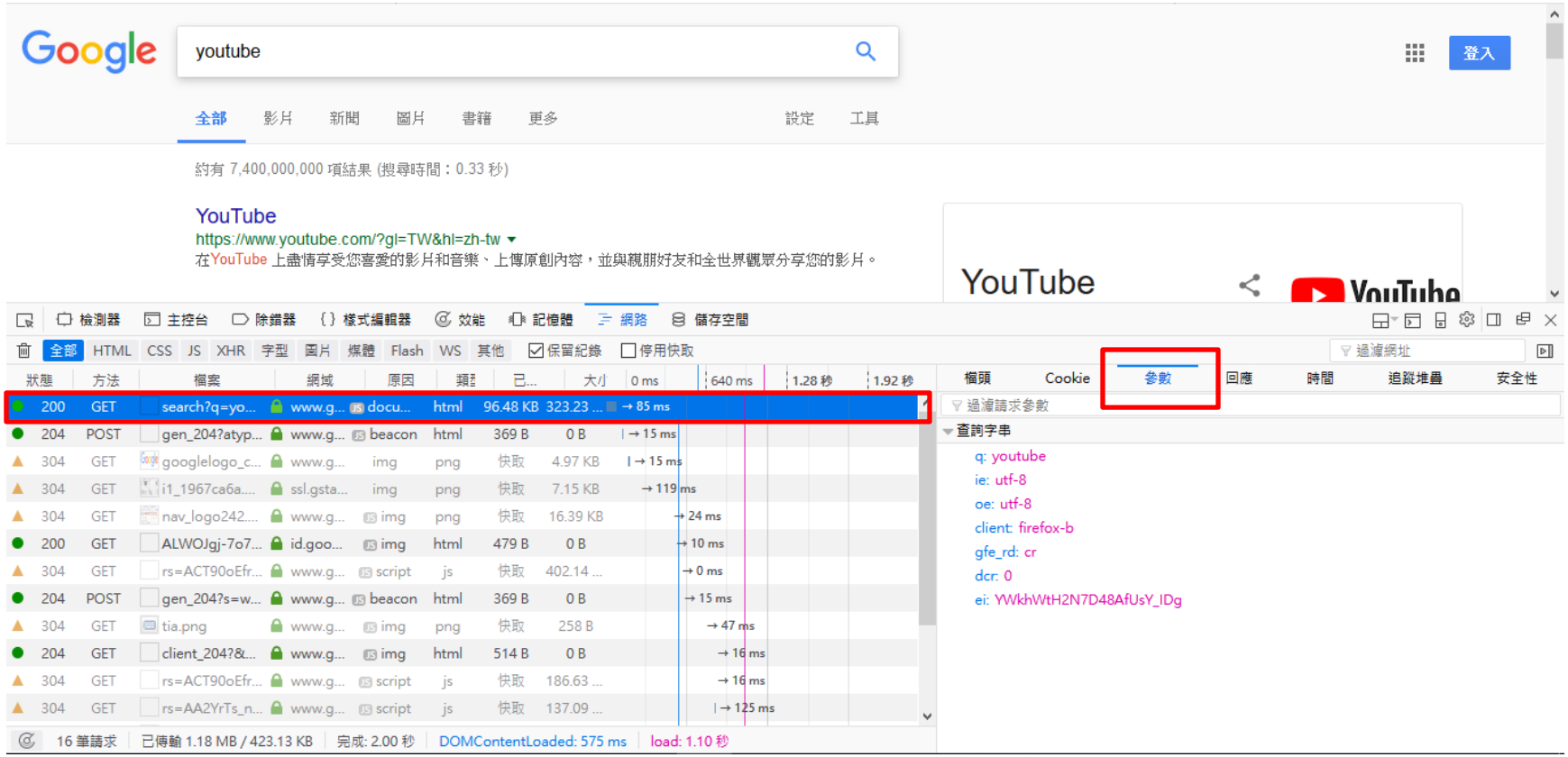
scrapy

What we will focus in this course!

How to observe GET

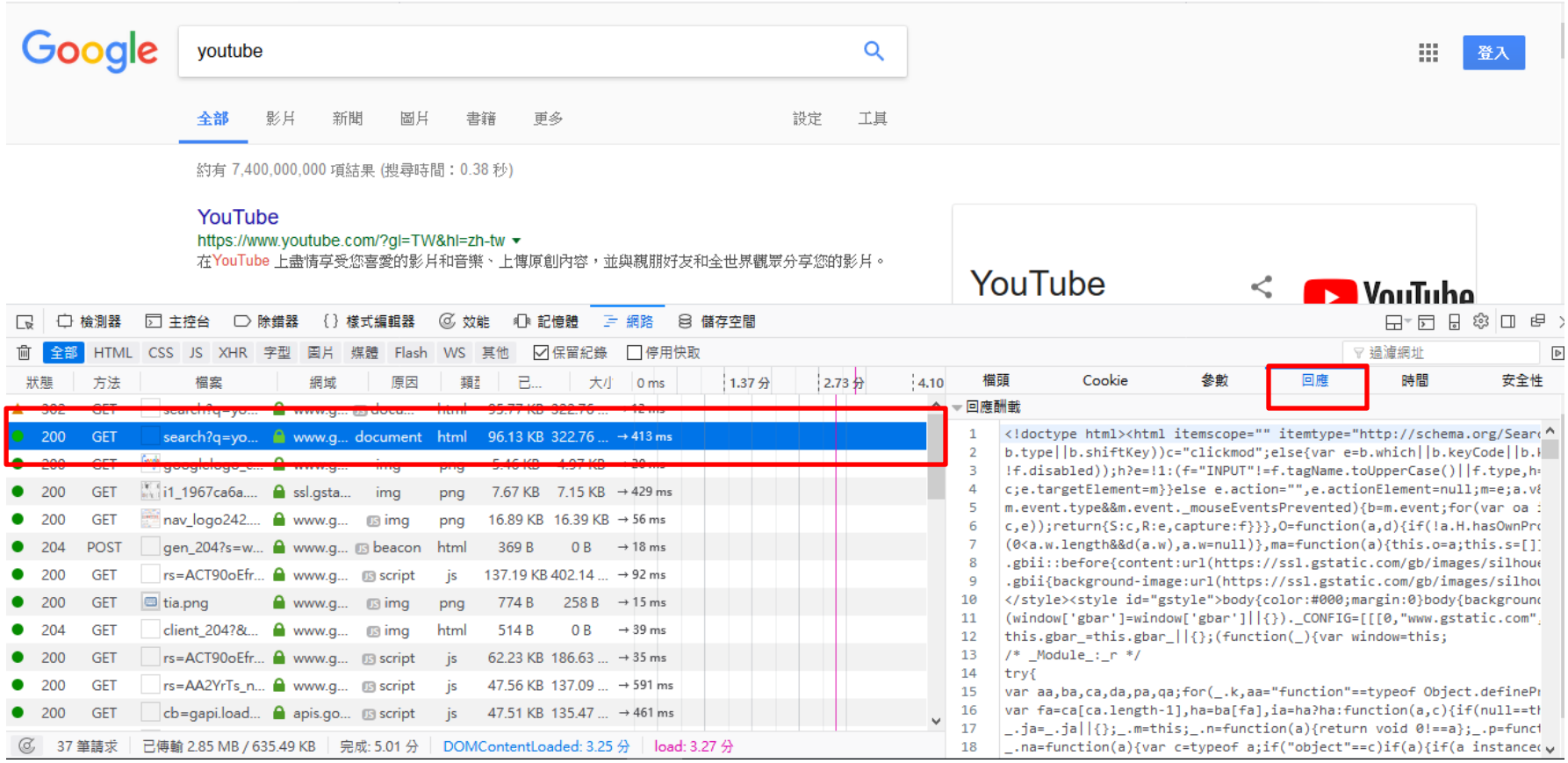


How to observe GET

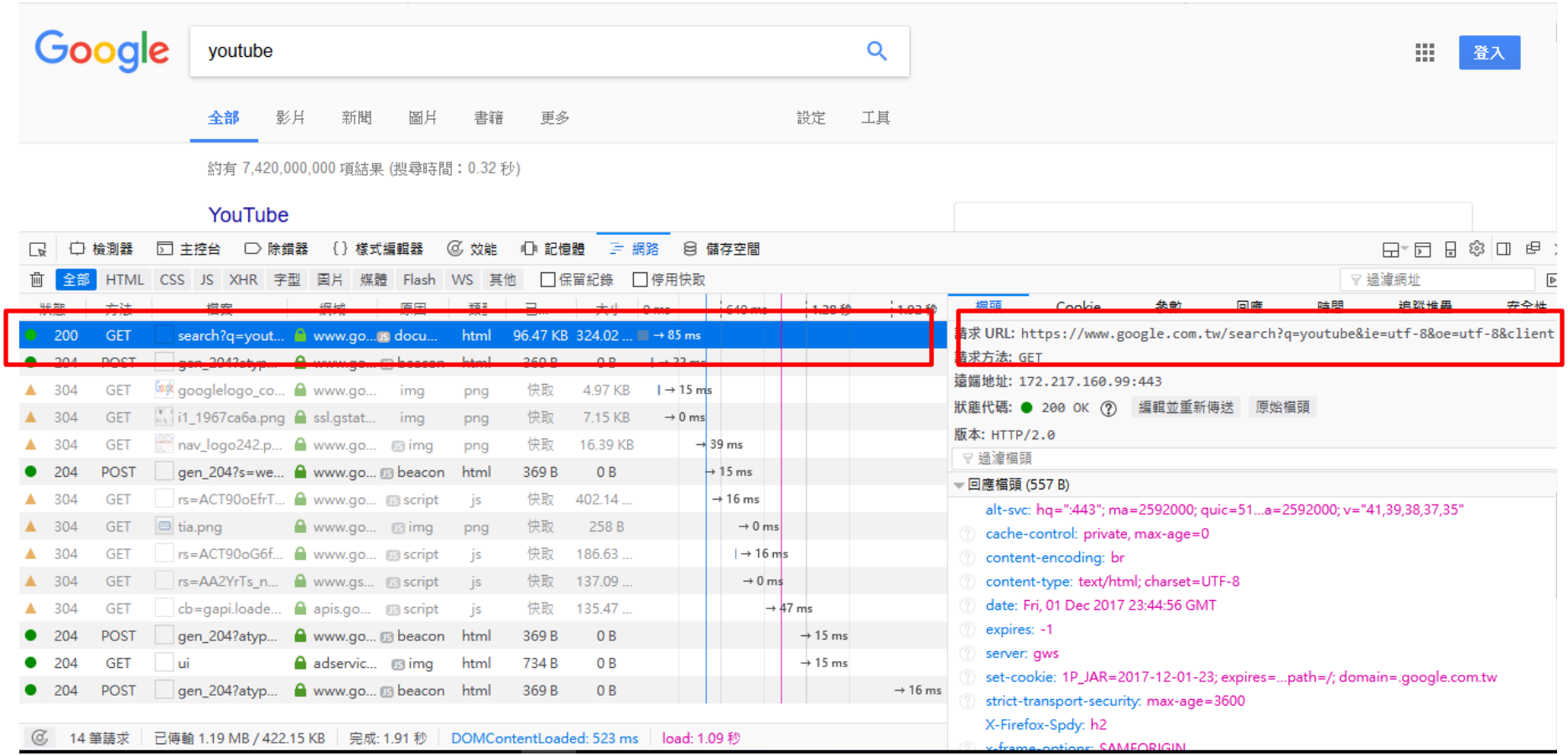


https://www.google.com.tw/search?q=youtube&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&dcr=0&ei=YWkhWtH2N7D48AfUsY_IDg

How to observe GET



How to observe GET

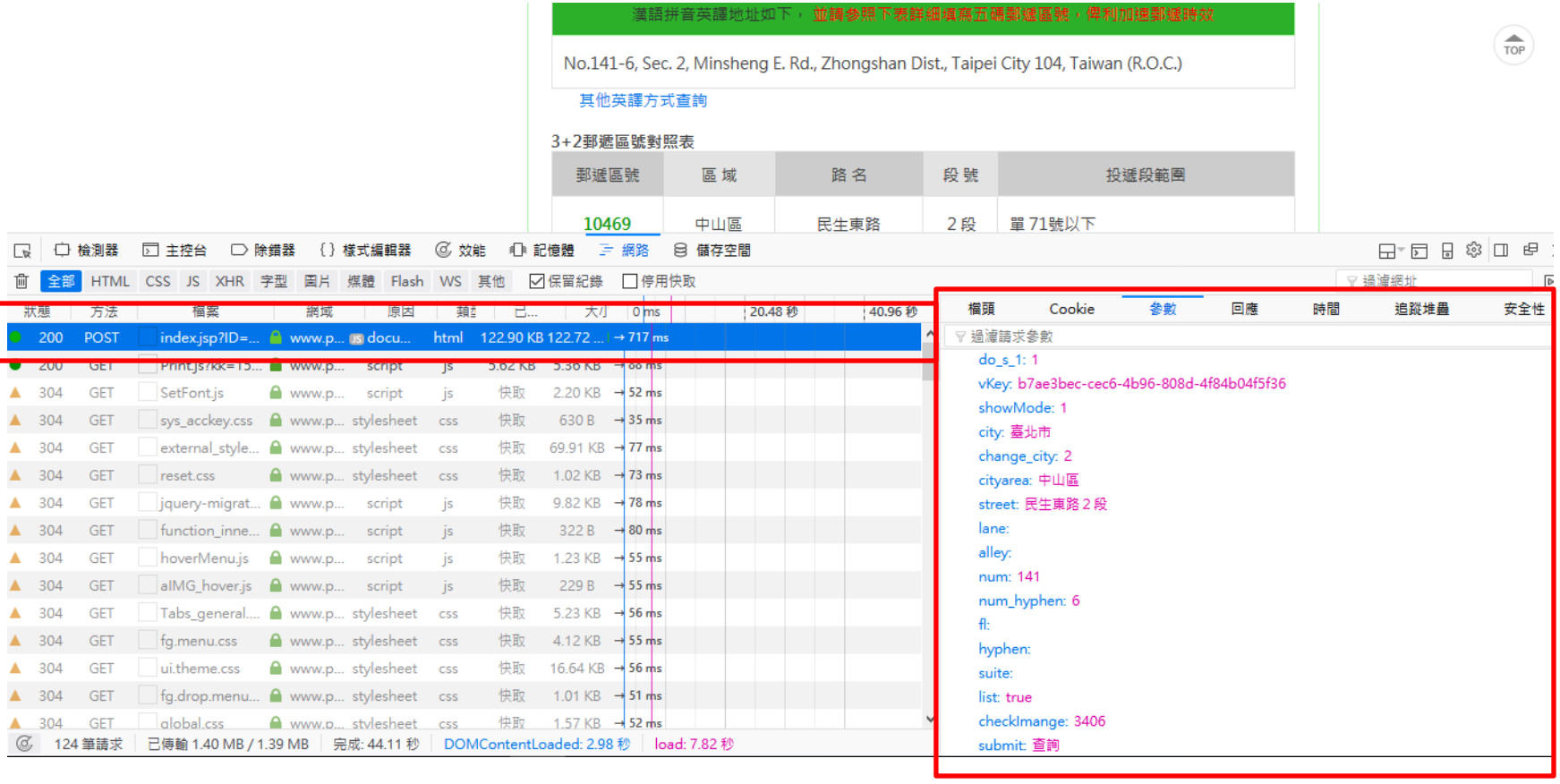


Note that parameters in "get" is append to url

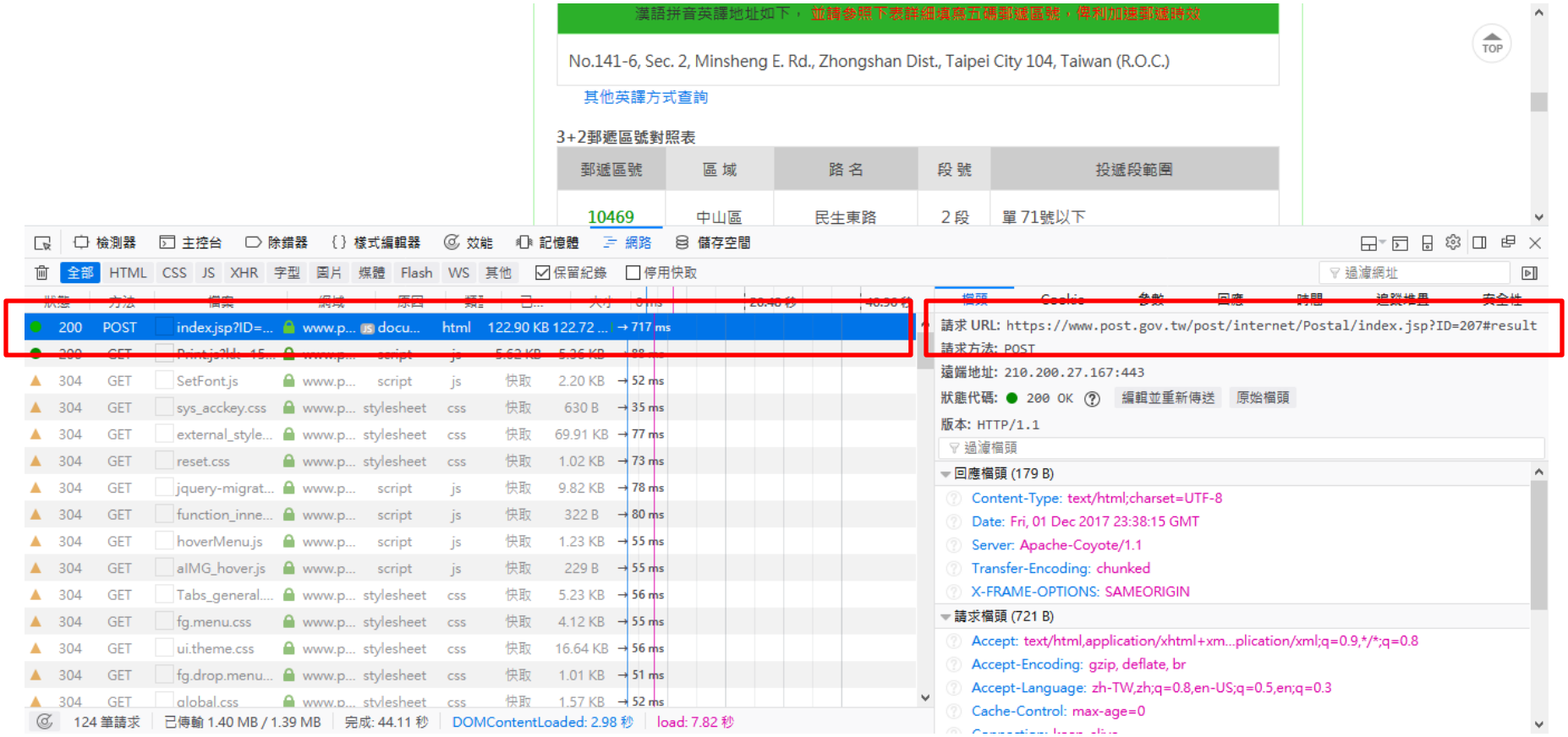
How to observe POST

台北市中山區民生東路二段141號號 6F

How to observe POST



How to observe POST



Example and Exercise

- Observe POST/GET on the following links (5 min)
  - https://www.post.gov.tw/post/internet/Postal/index.jsp?ID=207
  - https://statementdog.com/

Some reminder before crawling

Don’t request too frequently in a specific web server

Common module in python crawl

- Requests module
  - use requests module to send requests to server
- Beautifulsoup module
  - An python package for parsing html file
  - Very useful for web scraping

https://www.youtube.com/watch?v=YI62Pmk4kTs

Common beautifulsoup method

Common Beautifulsoup method	meaning
title	return page title
text	Remove html tag and return content as string
find	Find the first matched content
find_all	Find all matched content
select	Select CSS

GET V.S. POST

```
import requests
from bs4 import BeautifulSoup
url_with_parameters = ''
response = requests.get(url_with_parameters)

print(response.text)
```

Get mothod in requests module

```
import requests

url = ''
payload = {'key1': 'value1', \
           'key2': 'value2', \
           'key3': 'value3'}

response = requests.post(url, data=payload)

print(response.text)
```

Post mothod in requests module

Example and Exercise

- Hello word in crawl **(very important)**
  - example\crawl\example_code.pdf(1-1)

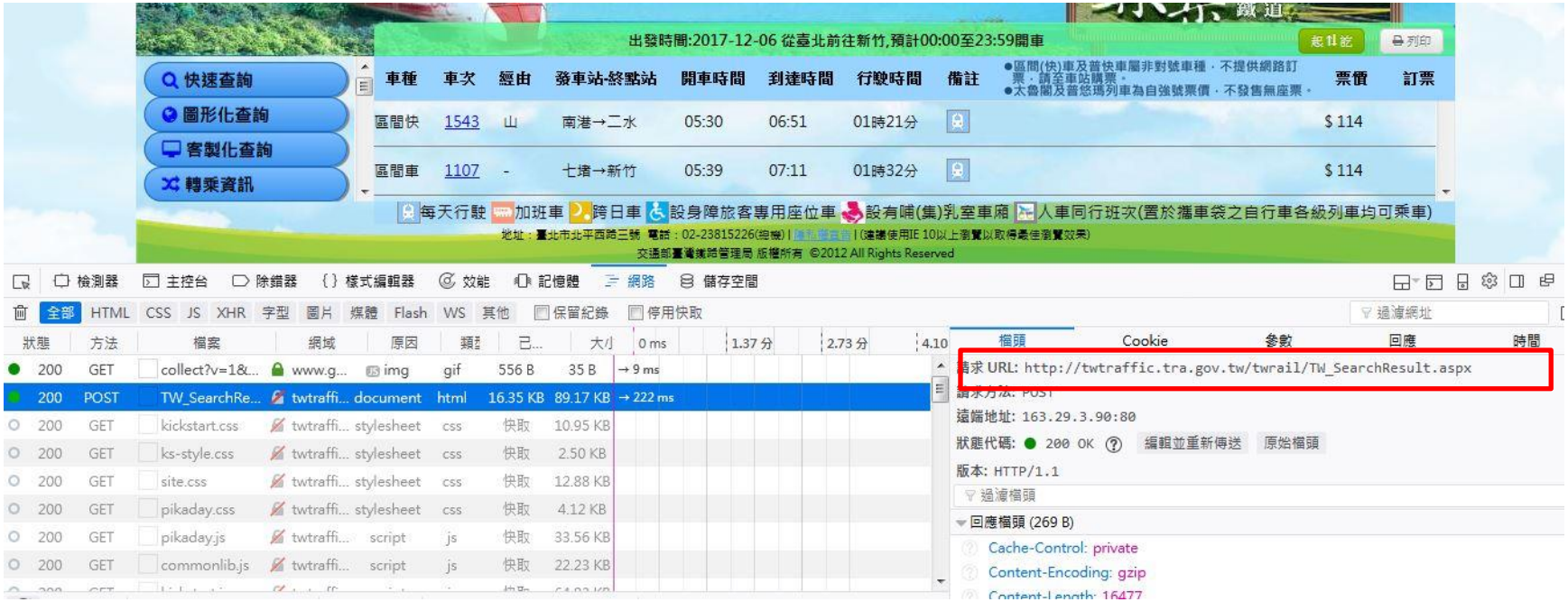
```
<html lang="zh-tw" xmlns:og="http://ogp.me/ns#">
<!--<![endif]-->
<head>

<link rel="apple-touch-icon" href="images/appletouch.gif" />
<link rel="SHORTCUT ICON" href="images/faviconntu.ico" />
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<meta name="viewport" content="width=device-width,initial-scale=1,maximum-scale=1,user-scalable:no">
<meta property="og:title" content="國立臺灣大學">
<meta property="og:site_name" content="國立臺灣大學">
<meta property="og:description" content="臺灣第一所最完整，歷史最悠久，且最具代表之綜合性高等教育學府">
<meta property="og:url" content="http://www.ntu.edu.tw/">
<meta property="og:image" content="http://www.ntu.edu.tw/images/photo/ntugate.jpg">
<meta http-equiv="Pragma" content="no_cache">
<script src="js/libs/jquery-1.10.2.min.js"></script>
<link rel="canonical" href="_"/>
<!--[if lt IE 7 ]>
<meta content="0; url=ie6.html" http-equiv="refresh" />
        <![endif]-->
<link href="css/normalize.css" rel="stylesheet">
<link href="css/ie8.css" rel="stylesheet">
<!--[if gte IE 9]><!-->
        <link href="css/main.css" rel="stylesheet">
        <link href="css/flexslider.css" rel="stylesheet">
        <link href="css/responsiveslides.css" rel="stylesheet">
        <link href="css/responsive.css" rel="stylesheet"><!--<![endif]-->
<!--[if lt IE 9]>
<script src="js/libs/html5.js"></script>
<![endif]-->
```

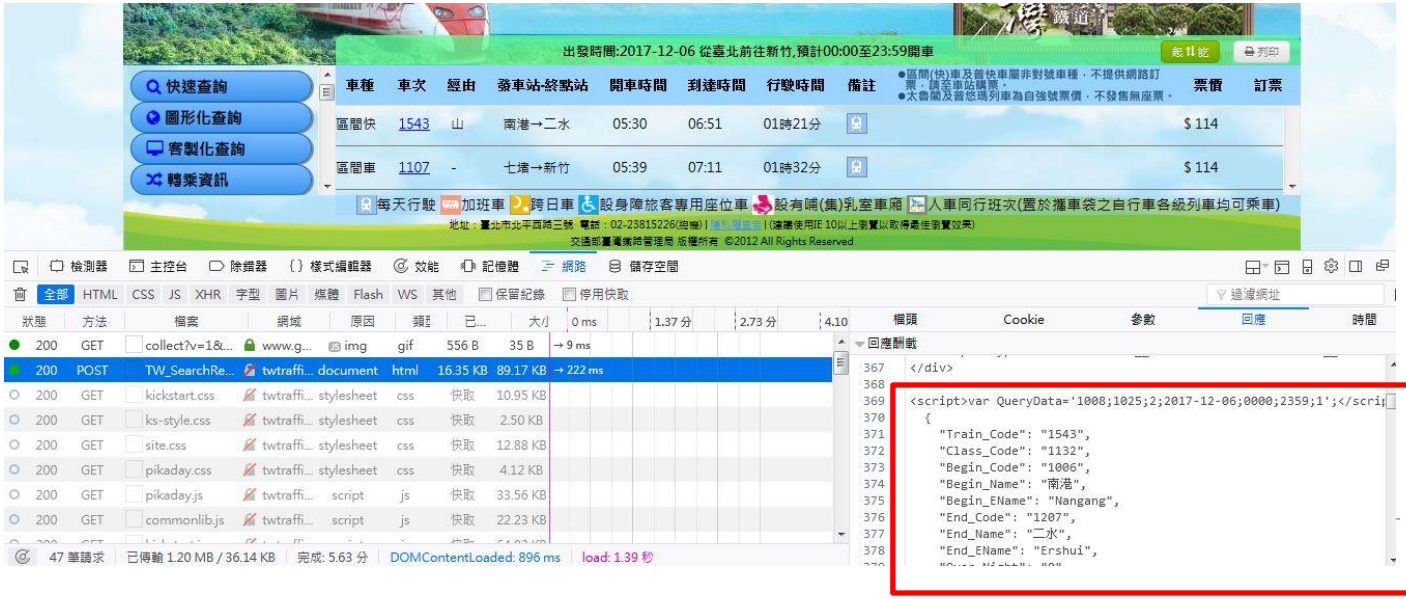
Example and Exercise

Example and Exercise

- Post example
  - example\crawl\example_code.pdf(1-2)




臺灣鐵路管理局
列車時刻查詢系統
出發時間:2017-12-06 從臺北前往新竹,預計00:00至23:59開車
快速查詢
圖形化查詢
客製化查詢
車種 車次 經由 發車站-終點站 開車時間 到達時間 行駛時間 備註 票價 訂票
區間快 1543 山 南港→二水 05:30 06:51 01時21分 $ 114
每天行駛 加班車 跨日車 殷身障旅客專用座位車 設有哺(集)乳室車廂 人車同行班次(置於攜車袋之自行車各級列車均可乘車)
FromStationName: 1008
ToStationName: 1025
TrainClass: 2
searchdate: 2017-12-06
FromTimeSelect: 0000
ToTimeSelect: 2359
Timetype: 1


出發時間:2017-12-06 從臺北前往新竹,預計00:00至23:59開車
快速查詢
圖形化查詢
客製化查詢
轉乘資訊
車種 車次 經由 發車站-終點站 開車時間 到達時間 行駛時間 備註 票價 訂票
區間快 1543 山 南港→二水 05:30 06:51 01時21分 $ 114
區間車 1107 - 七堵→新竹 05:39 07:11 01時32分 $ 114
每天行駛 加班車 跨日車 殷身障旅客專用座位車 設有哺(集)乳室車廂 人車同行班次(置於攜車袋之自行車各級列車均可乘車)
<script>var QueryData='1008;1025;2;2017-12-06;0000;2359;1';</scri
{
"Train_Code": "1543",
"Class_Code": "1132",
"Begin_Code": "1006",
"Begin_Name": "南港",
"Begin_EName": "Nangang",
"End_Code": "1207",
"End_Name": "二水",
"End_EName": "Ershui",

Example and Exercise

- Some beautiful example
  - https://isaac60103.github.io/crawl_course/example/beautifulsoup_example.html
  - example\crawl\example_code.pdf(1-3)
- Handle one line html tags
  - example\crawl\example_code.pdf(1-4)

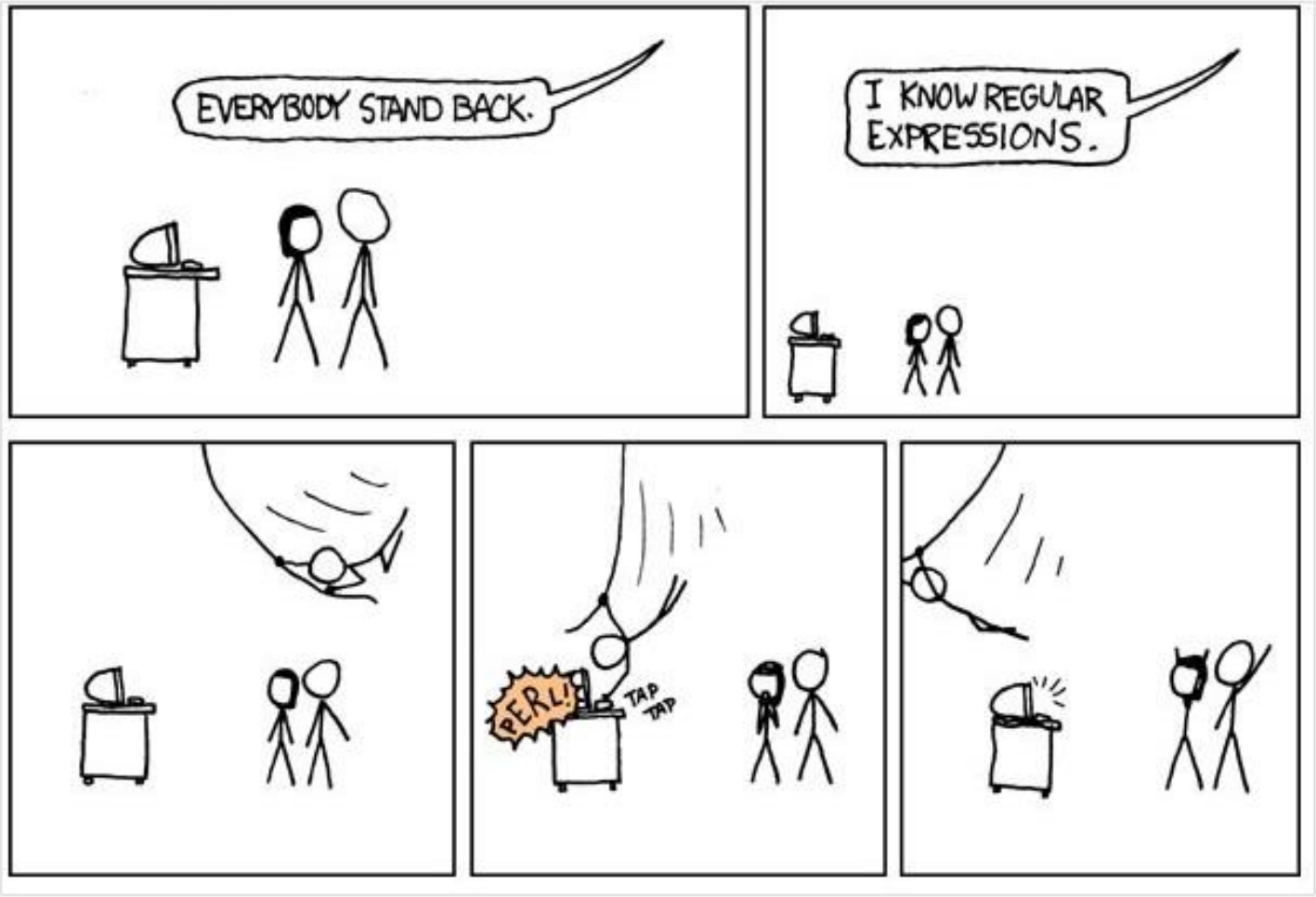
?

Example and Exercise (15 min)

- Download requests module
  - pip install requests/ conda install requests
- Download beautiful soup module
  - pip install beautifulsoup4/ conda install beautifulsoup4
- Please find all h1 tag in the following page
  - http://www.nccu.edu.tw/
  - practice\crawl\practice_code.pdf(1-1)
- Please use POST method to find that how many train from "台北" to "新竹" at 12:00 on 2017/12/11
  - https://www.thsrc.com.tw/tw/TimeTable/SearchResult
  - practice\crawl\practice_code.pdf(1-2)

Regular Expression

- Regular express
  - really smart “find" or “search"
- Regular expressions are a powerful string manipulation tool
- All modern languages have similar library packages for regular expressions


EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

Common Regular Expression

symbol	meaning
*	0 or more
+	1 or more
?	0 or 1
{3}	Exactily 3
(3,5)	3 or 4 or 5
[abc]	Range(a or b or c)
a | b	a or b
.	Any character except new line(\n)
\	Escape character

Example

- More regression expression
  - http://i26.tinypic.com/24mxgt4.png

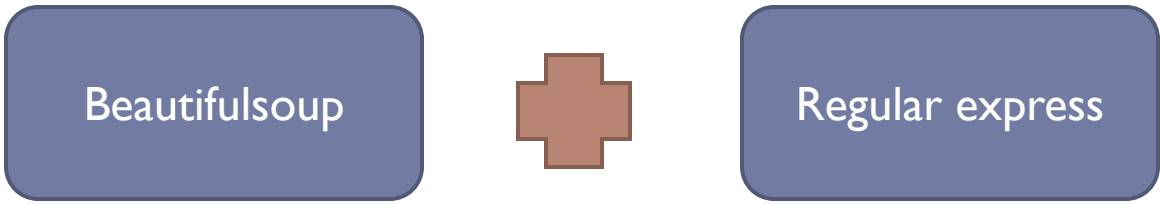
ab{2,4}c	an a followed by two, three or four b's followed by a c
ab{2,}c	an a followed by at least two b's followed by a c
ab*c	an a followed by any number (zero or more) of b's followed by a c
ab+c	an a followed by one or more b's followed by a c
ab?c	an a followed by an optional b followed by a c; that is, either abc or ac
a.c	an a followed by any single character (not newline) followed by a c
a\.c	a.c exactly
[abc]	any one of a, b and c
[Aa]bc	either of Abc and abc
[abc]+	any (nonempty) string of a's, b's and c's (such as a, abba, acbabcacaa)

Example and Exercise (5 min)

- Hello regular express
  - example\crawl\example_code.pdf(1-5)
- Grab email on web
  - https://isaac60103.github.io/crawl_course/practice/email_list.html
  - practice\crawl\practice_code.pdf(1-3)

Beautifulsoup with regular express

- Put regular expression pattern in re.compile() and use beautifulsoup to further analyze target data



Example and Exercise (5 min)

- Beautifulsoup with regular express
  - example\crawl\example_code.pdf(1-6)
- Please crawl li tag where its content contain “高雄” in the following link
  - http://yp.518.com.tw/service-life.html?ctf=10
  - practice\crawl\practice_code.pdf(1-4)

```
<li class="comp_loca">高雄市 / 湖內區 忠孝街130巷139號</li>
<li class="comp_loca">高雄市 / 楠梓區 德賢路471號</li>
```

Crawl image

Please note src = XXX, which point to image location

Crawl file

Please note href= XXX, which point to file location

Example and Exercise (10 min)

- crawl image
  - example\crawl\example_code.pdf(1-7)
- crawl and download first business card on the following link
  - https://www.mrcaca.com/
  - practice\crawl\practice_code.pdf(1-5)

Example and Exercise

- Please download any file in the following link
  - https://www.go100.com.tw/exam_download_3.php
  - practice\crawl\practice_code.pdf(1-6)

Practical issues

User agent

- Sometimes, you need to pretend you are not crawl to cheat the webpage

User-agent = XXX

Example and Exercise

- User-agent
  - example\crawl\example_code.pdf(1-8)

Third party API

- Some of services provide API
  - No need to crawl by yourself
  - https://developers.facebook.com/
  - https://www.youtube.com/yt/dev/zh-TW/api-resources.html
  - https://developers.google.com/maps/?hl=zh-tw

How to solve captcha?

- 2captcha service
  - https://2captcha.com/
  - https://github.com/Mirio/captcha2upload
  - example\crawl\example_code.pdf(1-9)

Crawl a lot of websites

```
url = ''

max_retry = 3  #set times to retry when connection got errors.
retry = 1
payload = {}

while retry <= max_retry:
    try:
        #Request the data with the payload.
        res = requests.post(url, data=payload)
        sleep(2)
        print(res.text)

        #You also can do something using BeautifulSoup if the resource is in html form.

        #Break the loop when finished the job.
        break

    except Exception as e:
        #If connection is failed.
        retry += 1
        print("I got an error: ", e)
        continue
```

Important!

add max retry and try/except when crawling very large data

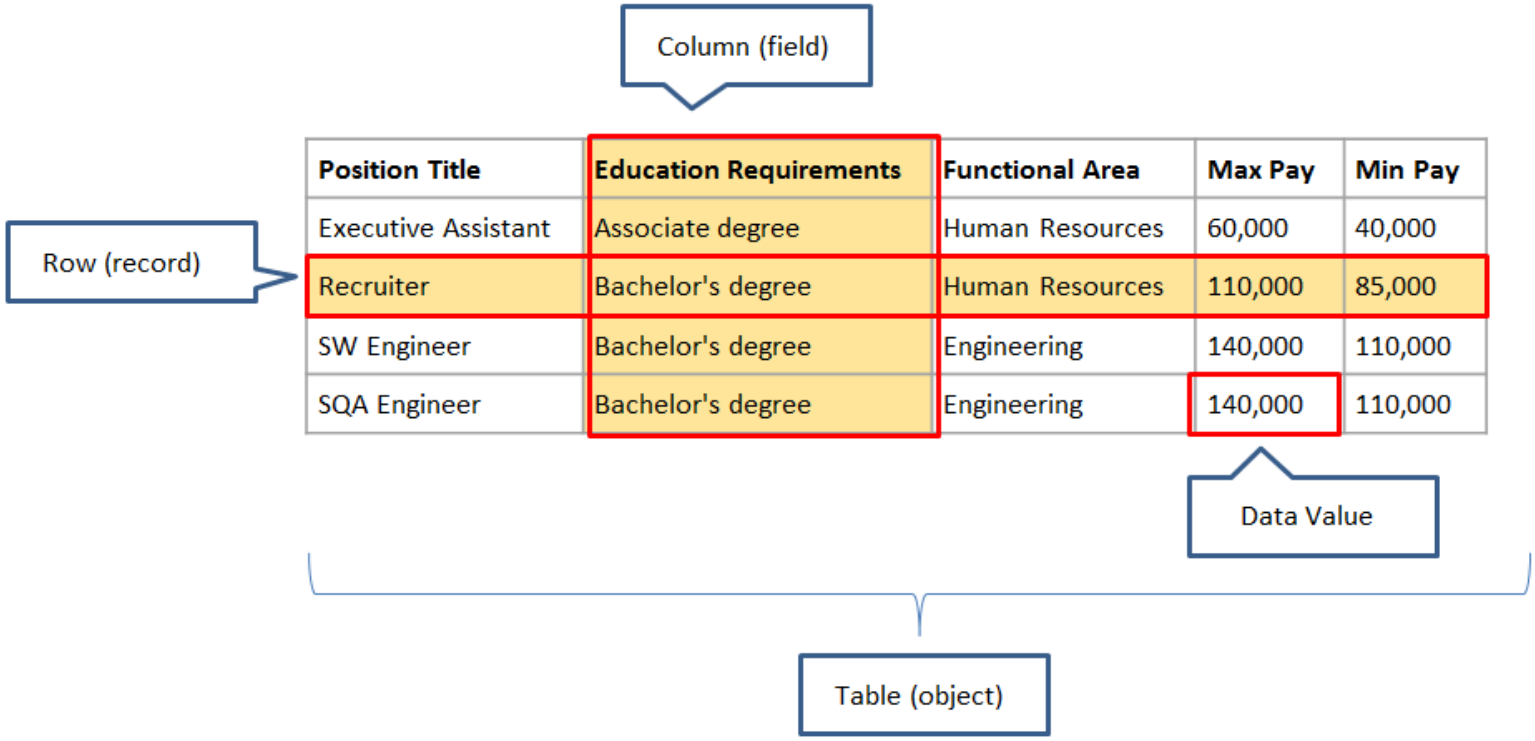
Example and Exercise

- retry and try/except crawl
  - example\crawl\example_code.pdf(1-10)

Automatically crawl

- Check if web page change or not
  - example\crawl\example_code.pdf(1-11)
- Windows automation setting
  - windows_automation_setting.pptx
- Mac automation setting
  - MAC_automation_setting.pptx

Database terminology

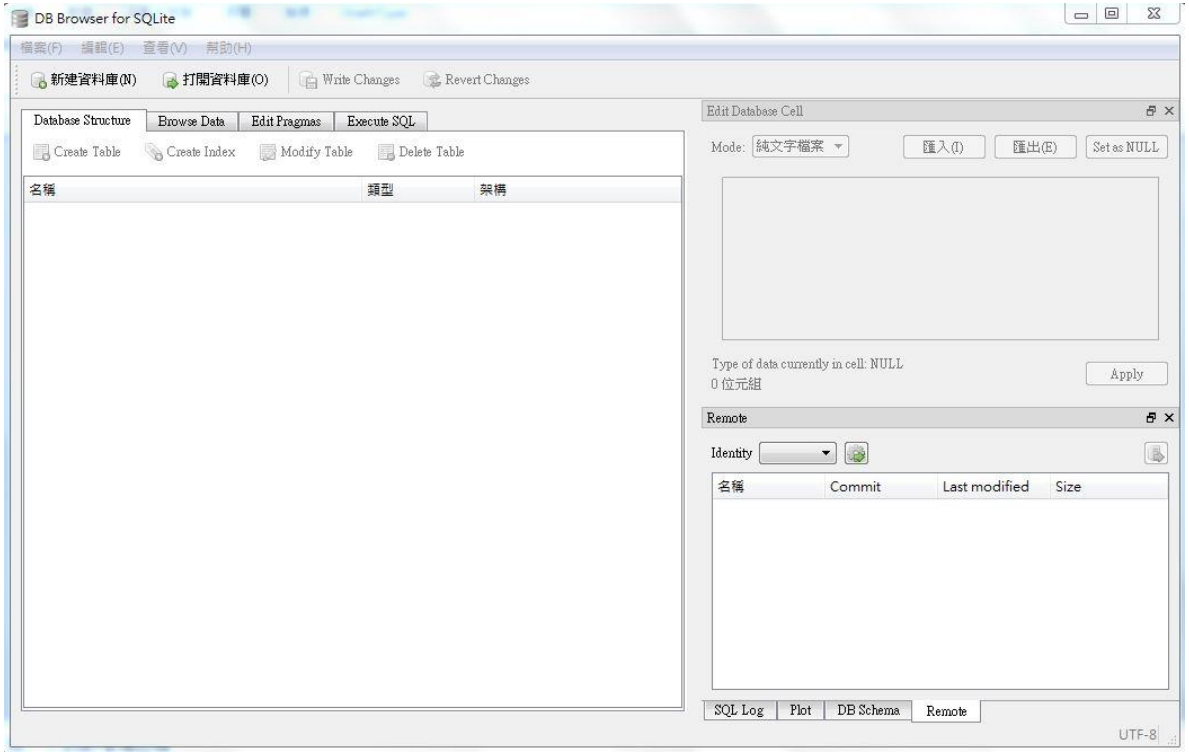


SQLite

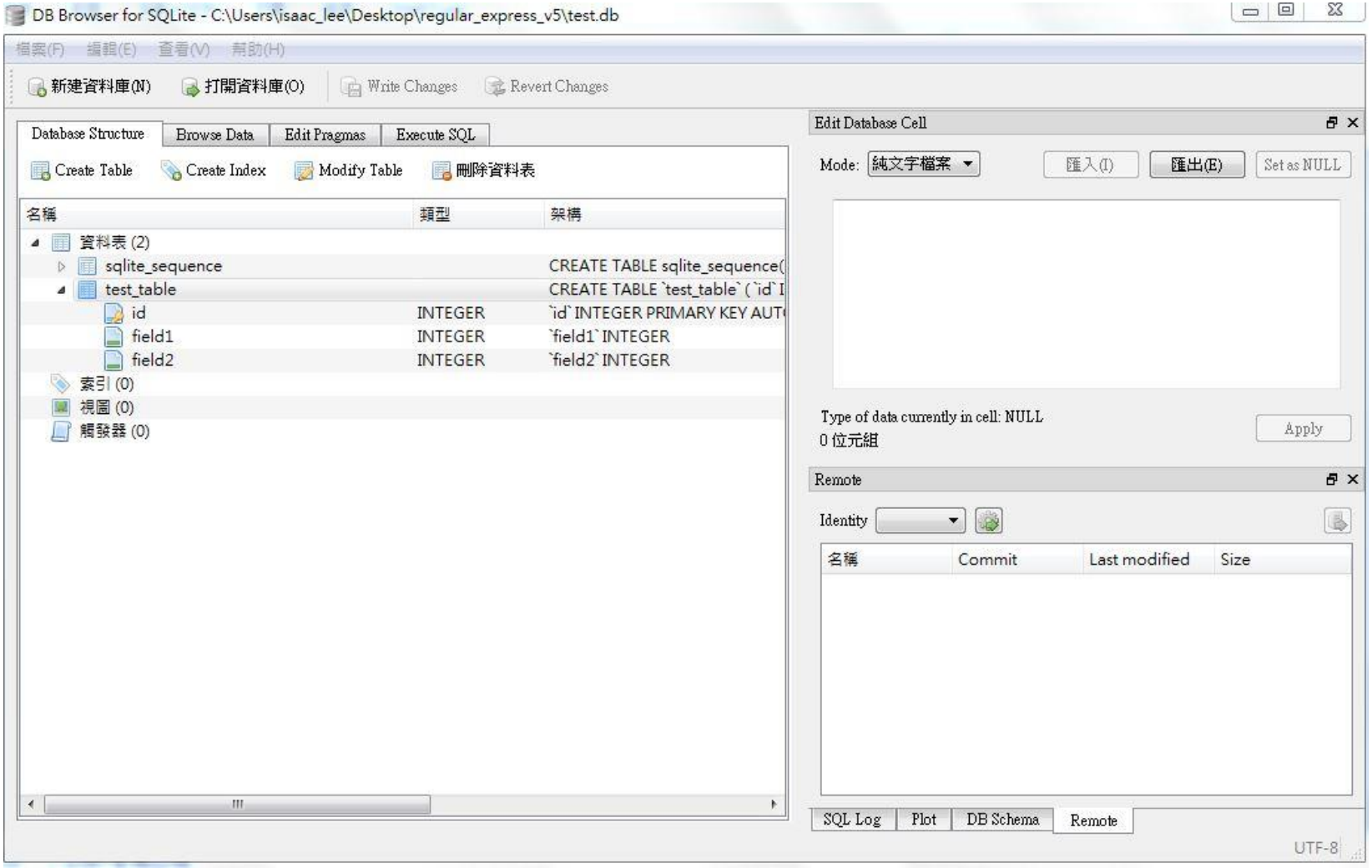
- Relational database management system
- Embedded database software for local/client storage
- Use SQL to manipulate

SQLite

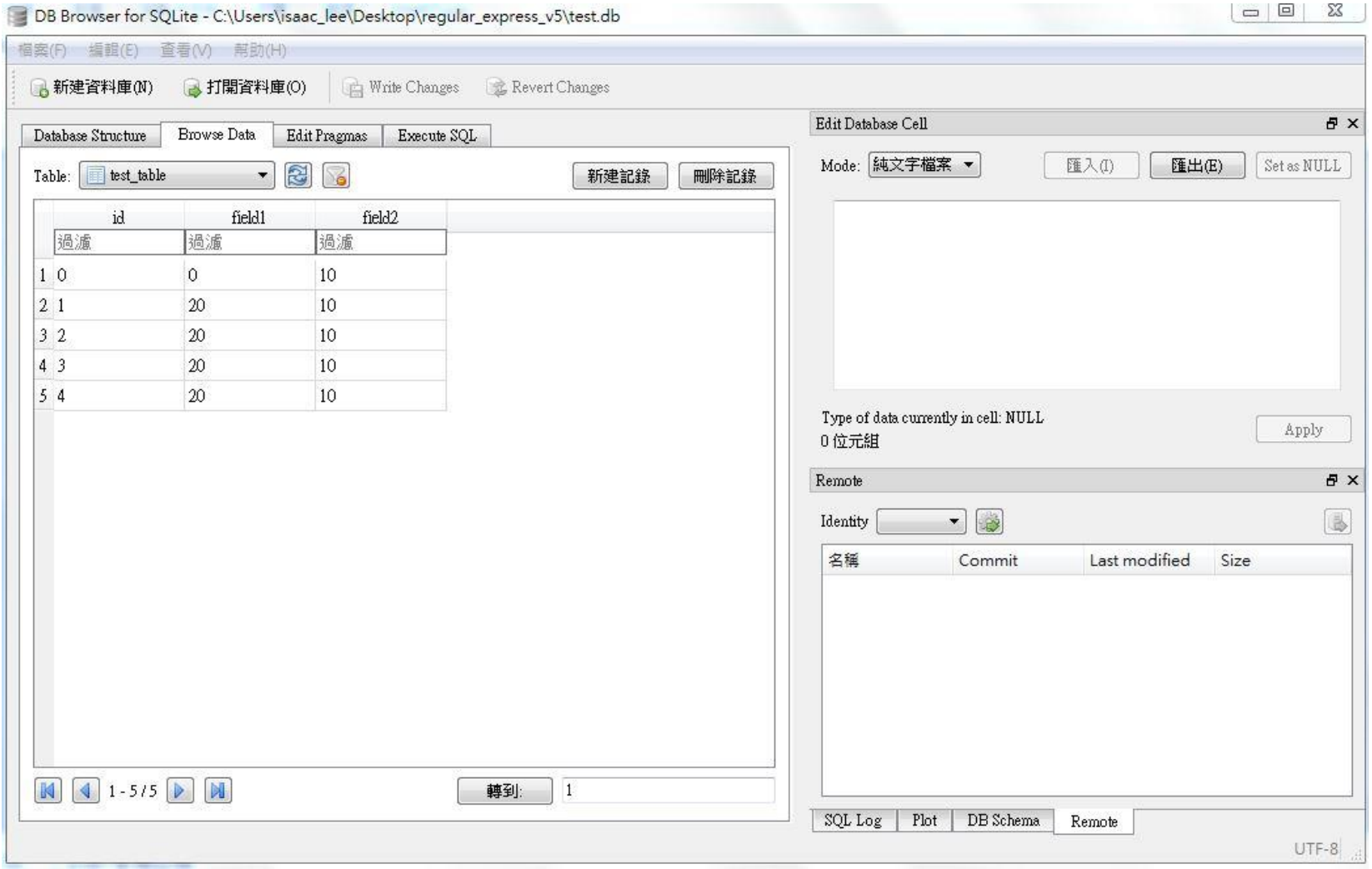
- sqlitebrowser
  - http://sqlitebrowser.org/



SQLite


DB Browser for SQLite - C:\Users\isaac_lee\Desktop\regular_express_v5\test.db
檔案(F) 編輯(E) 查看(V) 幫助(H)
新建資料庫(N)
打開資料庫(O)
Write Changes
Revert Changes
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Create Table
Create Index
Modify Table
刪除資料表
名稱
類型
架構
資料表 (2)
sqlite_sequence
CREATE TABLE sqlite_sequence(
test_table
CREATE TABLE `test_table` (`id` I
id
INTEGER
`id` INTEGER PRIMARY KEY AUT
field1
INTEGER
`field1` INTEGER
field2
INTEGER
`field2` INTEGER
索引 (0)
視圖 (0)
觸發器 (0)
Edit Database Cell
Mode:
純文字檔案
匯入(I)
匯出(E)
Set as NULL
Type of data currently in cell: NULL
0 位元組
Apply
Remote
Identity
名稱
Commit
Last modified
Size
SQL Log
Plot
DB Schema
Remote
UTF-8

SQLite


DB Browser for SQLite - C:\Users\isaac_lee\Desktop\regular_express_v5\test.db
檔案(F) 編輯(E) 查看(V) 幫助(H)
新建資料庫(N)
打開資料庫(O)
Write Changes
Revert Changes
Database Structure
Browse Data
Edit Pragmas
Execute SQL
Table: test_table
新建記錄
刪除記錄
id field1 field2
過濾 過濾 過濾
1 0 0 10
2 1 20 10
3 2 20 10
4 3 20 10
5 4 20 10
1 - 5 / 5
轉到: 1
Edit Database Cell
Mode: 純文字檔案
匯入(I)
匯出(E)
Set as NULL
Type of data currently in cell: NULL
0 位元組
Apply
Remote
Identity
名稱 Commit Last modified Size
SQL Log
Plot
DB Schema
Remote
UTF-8

SQL

- SQL is a standard language for storing, manipulating and retrieving data in databases
  - Can be used in SQLite, MySQL, SQL Server, MS Access, etc……

SQL

Example and Exercise (10 min)

- Use python to insert and query
  - example\crawl\example_code.pdf(1-12)
- Use panda to insert and query
  - http://www.taipeibo.com/year/2017
  - example\crawl\example_code.pdf(1-13)
- Please create a DB call “house_db” and a table call “house_table”
  - Insert data to DB from the following csv file
  - practice\house.csv
  - practice\crawl\practice_code.pdf(1-7)

Example and Exercise

- More reference on SQL
  - https://www.w3schools.com/sql/
- SQL cheat sheet
  - cheatsheet\sqlcheatsheet2.pdf

Quick Recap

- GET V.S. POST method
- Beautifulsoup
- Regular expression
- Beautifulsoup+ Regular expression
- Some practical issues
- Sqlite DB usage