

Python Crawler

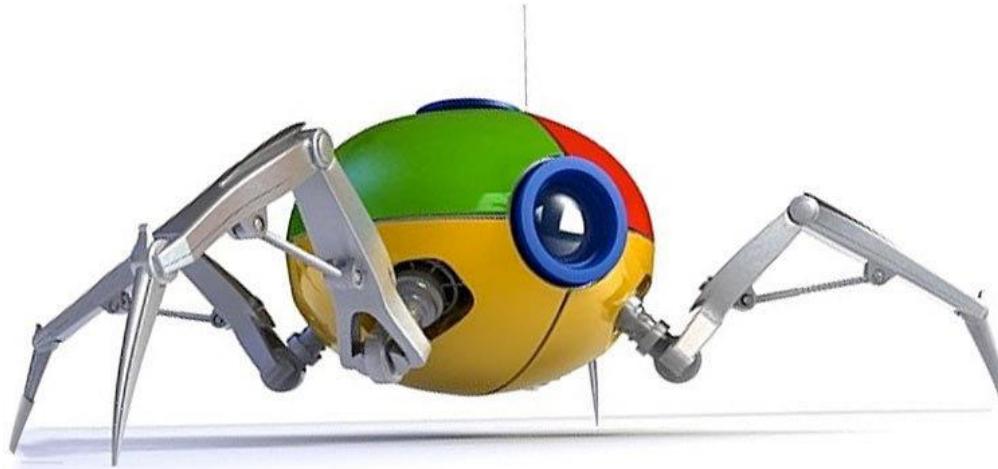
Outline

- ▶ **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

The screenshot shows the homepage of National Taiwan University. At the top, there is a dark navigation bar with links for '網站導覽' (Website Guide), '新生' (Freshman), '在校學生' (Institutional Students), '國際生' (International Students), '教職員' (Faculty and Staff), '訪客' (Guest), '校友' (Alumni), '聯絡我們' (Contact Us), '站內搜尋' (Search Site) with a search input field, and 'English'. Below the navigation bar is a yellow decorative banner featuring the university's logo and the text '國立臺灣大學' (National Taiwan University) and 'National Taiwan University'. The main content area has a large banner image of a scenic campus path lined with trees and a lake. Below the banner, the page title '學術單位' (Academic Units) is displayed in a red header. To the left, a sidebar lists '學術單位一覽表' (List of Academic Units) with links to '文學院', '理學院', '社會科學院', and '醫學院'. On the right, the '學術單位' section contains a sub-header '學術單位' (Academic Units) and a paragraph of text about the university's academic programs. It also includes a thumbnail image of a building and contact information: '文學院' (College of Liberal Arts), '設院時間: 1945年' (Established: 1945), and '電話: (02)33663988' (Phone: (02)33663988). A circular 'TOP' button is located in the bottom right corner.

What human see on a webpage

The screenshot shows a web browser displaying the National Taiwan University Academic Units page (www.ntu.edu.tw/academics/academics.html). The page features a banner with the university's name and logo, followed by a navigation bar with links like '網站導覽' and 'English'. Below the banner is a large photograph of a campus scene with a lake and people walking. On the left, a sidebar lists academic units: 文學院, 理學院, 社會科學院, 醫學院, 工學院, 生物資源暨農學院, 管理學院, 公共衛生學院, and 電機資訊學院. The main content area displays information for the Liberal Arts College (文學院) and the College of Science (理學院), each with a photo, founding year, phone number, fax number, website, and email address. A context menu is open in the top right corner, with the '檢視頁面資訊' (View page info) option highlighted with a red box.

www.ntu.edu.tw/academics/academics.html

網站導覽 | 新生 | 在校學生 | 國際生 | 教職員 | 訪客 | 校友 | 聯絡我們 | 站內搜尋 | English | 搜尋

學術庫 | 圖書館 | 博物館群 | 課程 | 招生 | 推廣教育 | 行事曆 | 捐款

認識臺大 學術單位 研究發展 行政組織 常見詢問 服務資源

學術單位

首頁 > 學術單位

學術單位

臺大現有11個學院，54個學系、103個研究所、6個碩博士學位學程，橫跨自然科學與人文社會科學。每學期所開課程數達8,000班以上，無論所擁有的學術領域或開設之課程數，在全國大學中居于领先地位。就讀於臺大，無疑是進入最為豐富的知識寶庫，可獲得最多元而優質的學習機會。

文學院

設院時間：1928年
電話：(02)33663988
傳真：(02)23638818
網址：<http://liberal.ntu.edu.tw/>
電子郵件：liberal@ntu.edu.tw

理學院

設院時間：1928年
電話：(02)33664188

TOP

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/basic?cx=011987122760880416627:zsghqazzvk4&sig=__ojBF1PpI1CTk40CeWf1D1J4PyGE=" 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&sig=__ojBF1PpI1CTk40CeWf1D1J4PyGE=" 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>
        </div>
      </div>
    </div>
  </noscript>
</header>
```

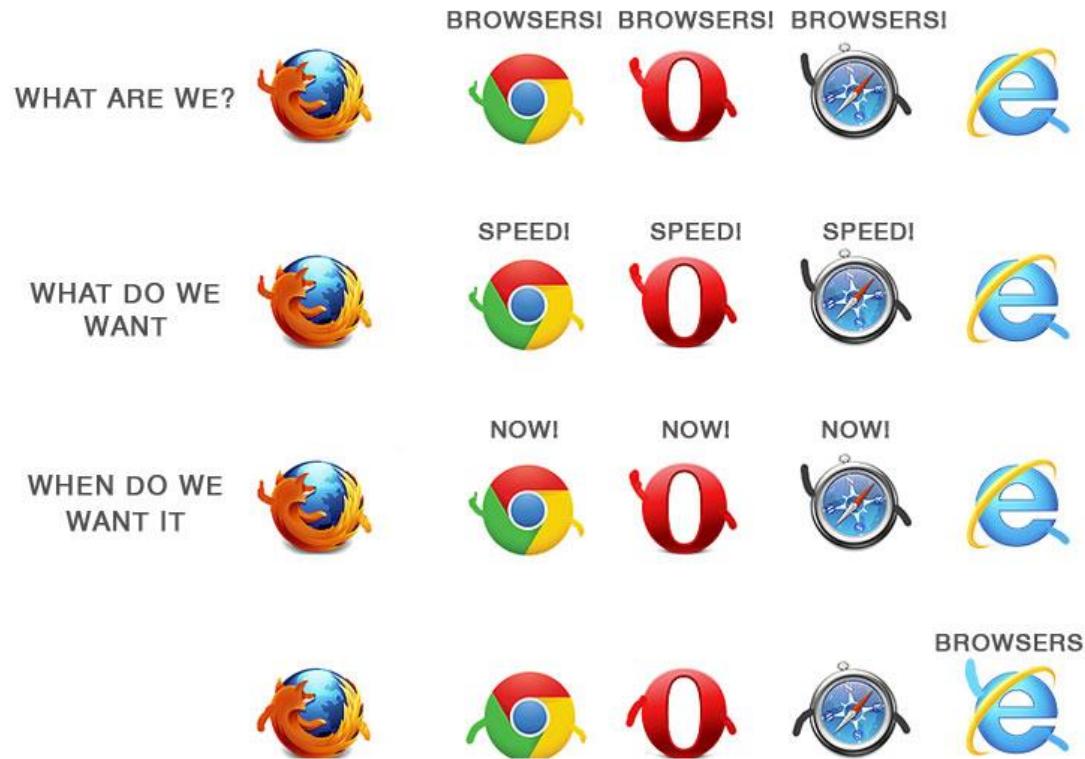


This is why we need to know HTML code.....

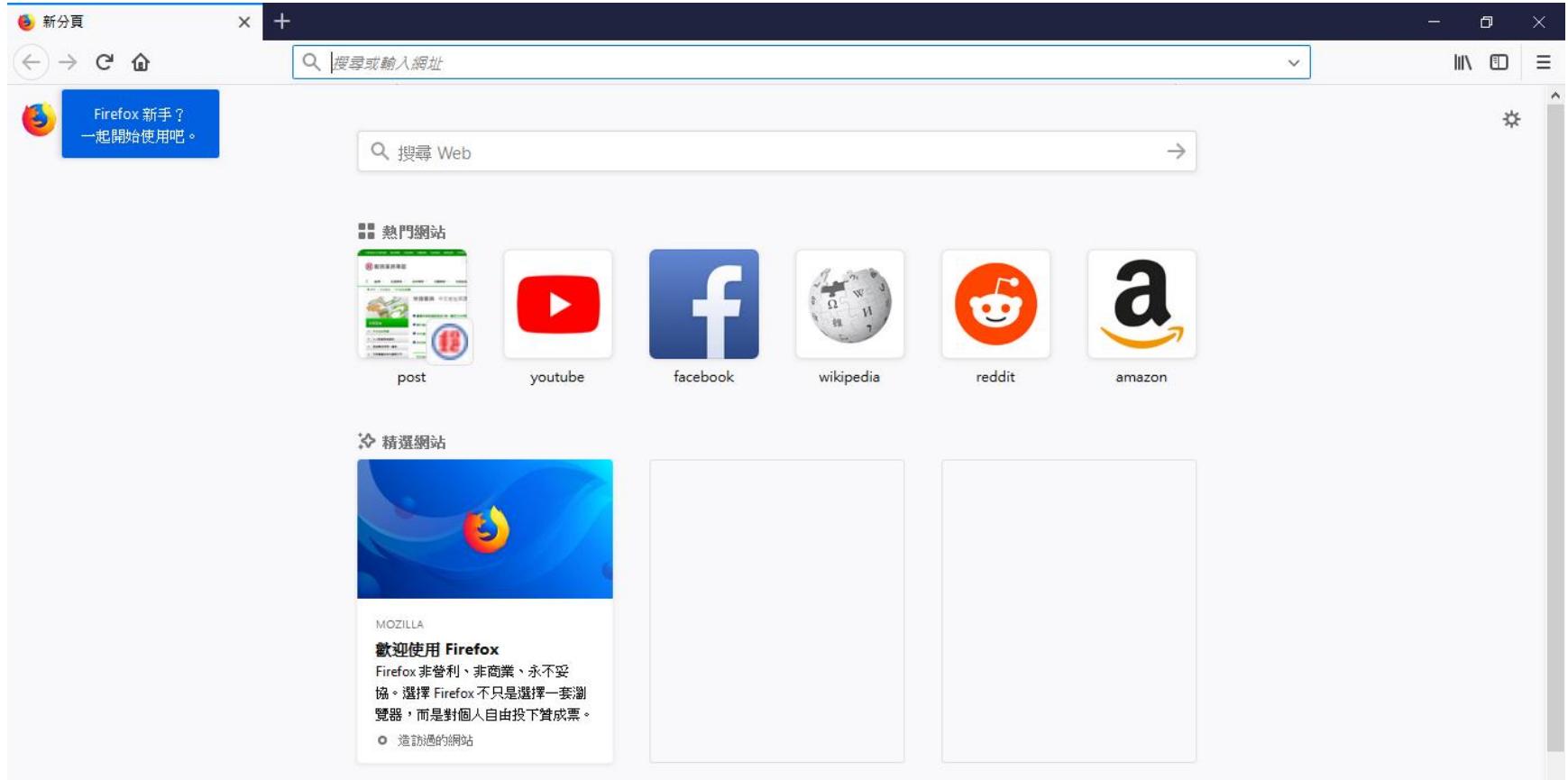


A Joke on IE

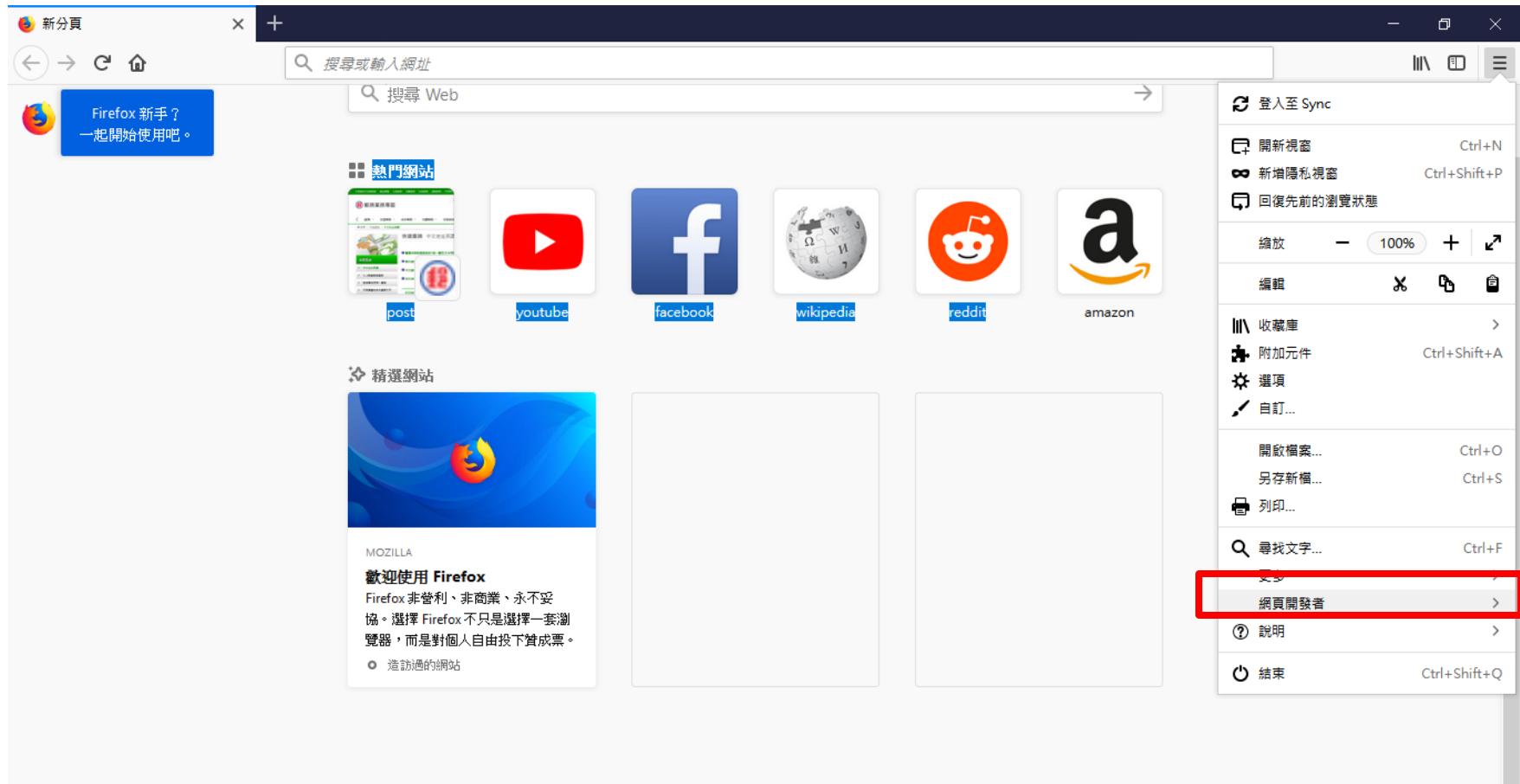
INTERNET EXPLORER'S SPEED VERSUS OTHER BROWSERS



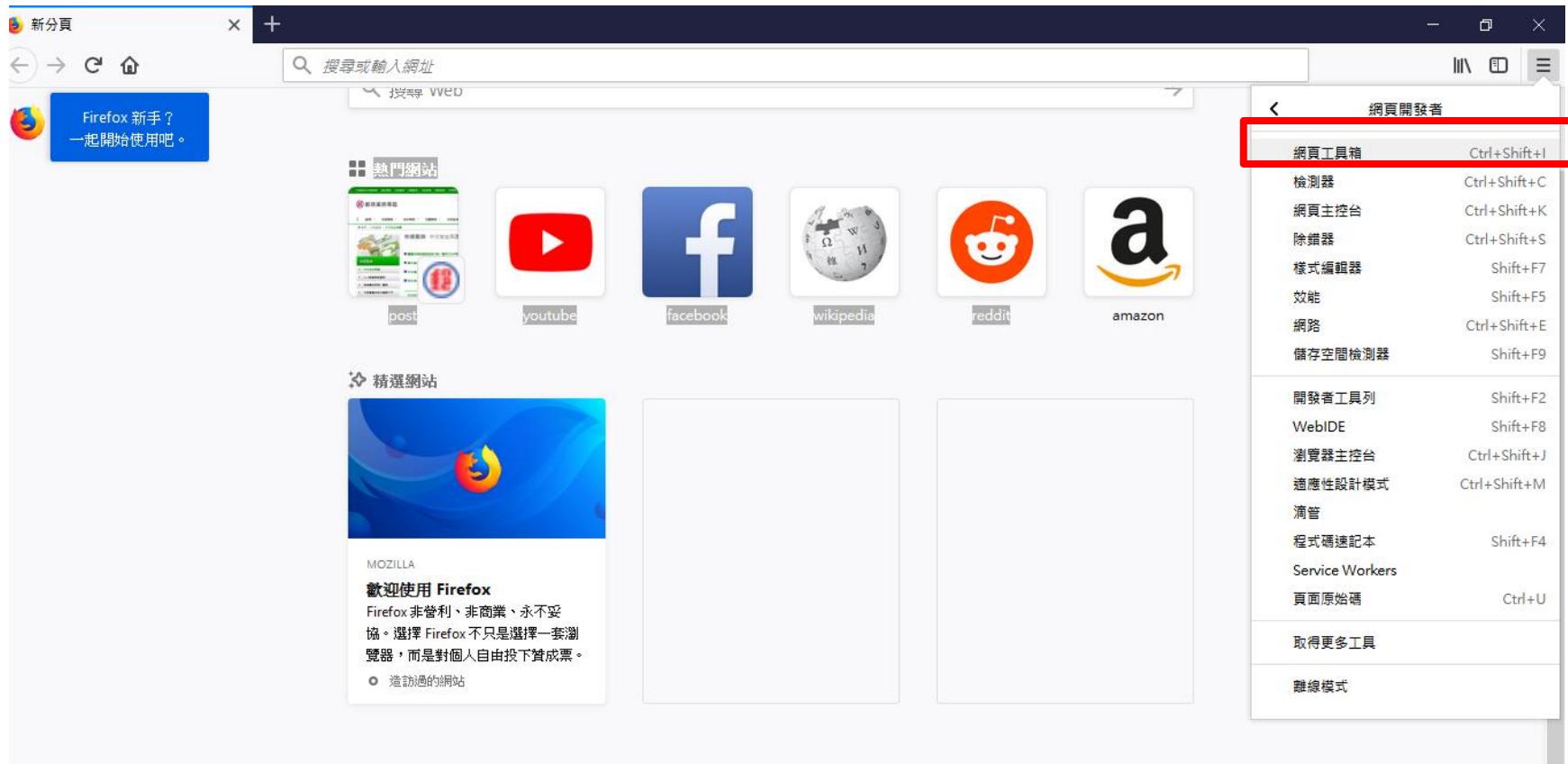
Firefox(recommend)



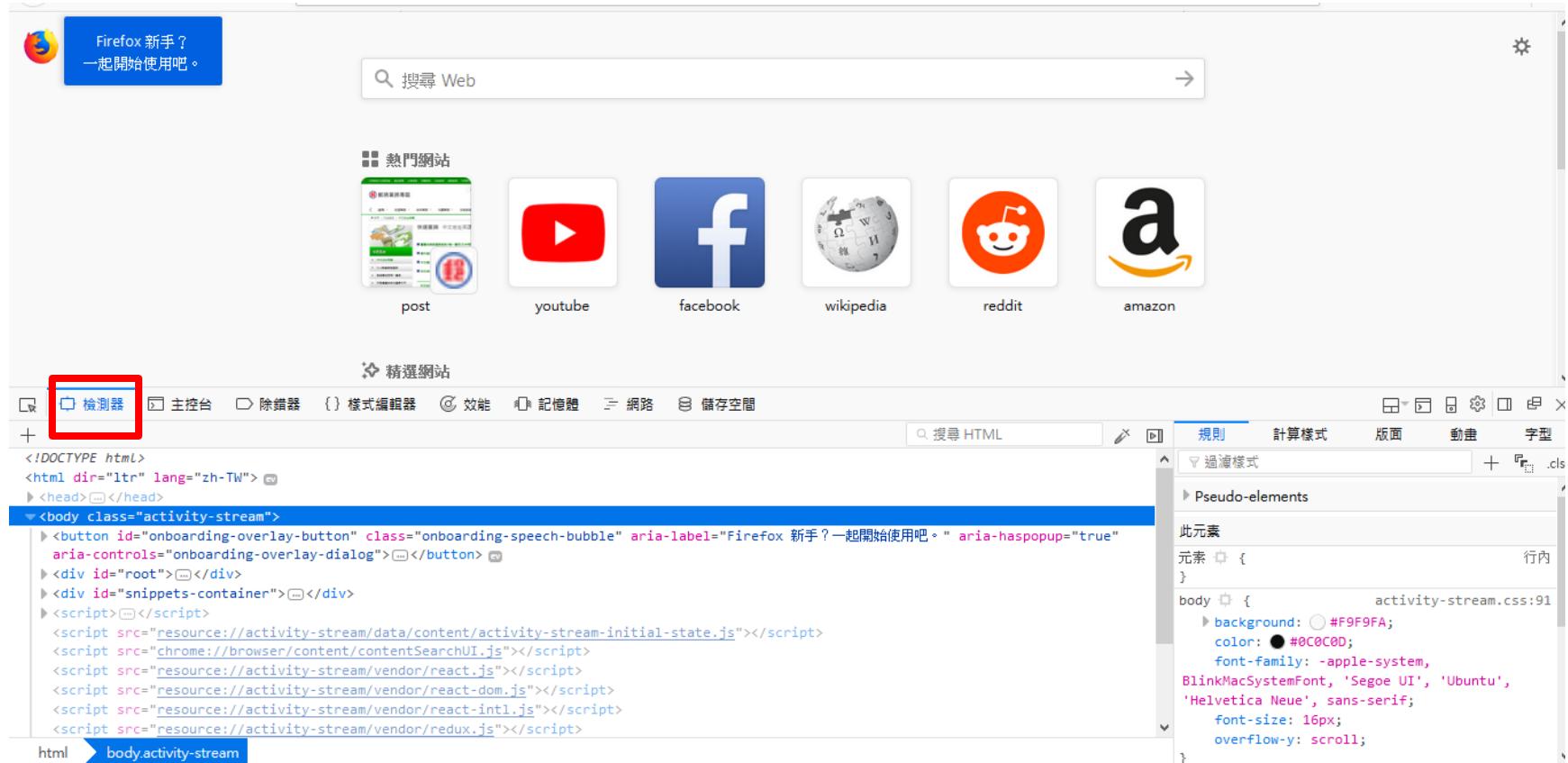
Firefox(recommend)



Firefox(recommend)



Firefox(recommend)



Use “F12” shortcut to open developer tool

Firefox(recommend)

The screenshot shows the official website of National Taiwan University (www.ntu.edu.tw). The page features a yellow header with the university's name in English and Chinese. Below the header are several navigation menus: '網站導覽' (Website Guide), '新生' (Freshman), '在校學生' (On-campus Students), '國際生' (International Students), '教職員' (Faculty and Staff), '訪客' (Visitors), '校友' (Alumni), '聯絡我們' (Contact Us), '站內搜尋' (Search Site), 'English' (English Version), and links to '學術庫' (Academic Library), '圖書館' (Library), '博物館群' (Museum Cluster), '課程' (Courses), '招生' (Admissions), '推廣教育' (Promotion Education), '行事曆' (Calendar), and '捐款' (Donations). The main content area includes sections for '最新消息' (Latest News), '活動快報' (Event Report), and '校園推廣' (Campus Promotion). The '校園推廣' section is highlighted with a dashed blue box and contains a thumbnail image of a building with palm trees. The bottom of the screen shows the Firefox developer tools interface, specifically the '檢測器' (Inspector) panel, which is used for inspecting webpage elements.

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

Firefox(recommend)

The screenshot shows the official website of National Taiwan University (www.ntu.edu.tw) displayed in the Firefox browser. The developer tools are open, specifically the Element Inspector, which highlights the HTML code for a promotional image in the 'Events' section. The image has a width of 116px and a height of 180px. The browser interface includes the address bar, navigation buttons, and various menu options.

Developer Tools Elements:

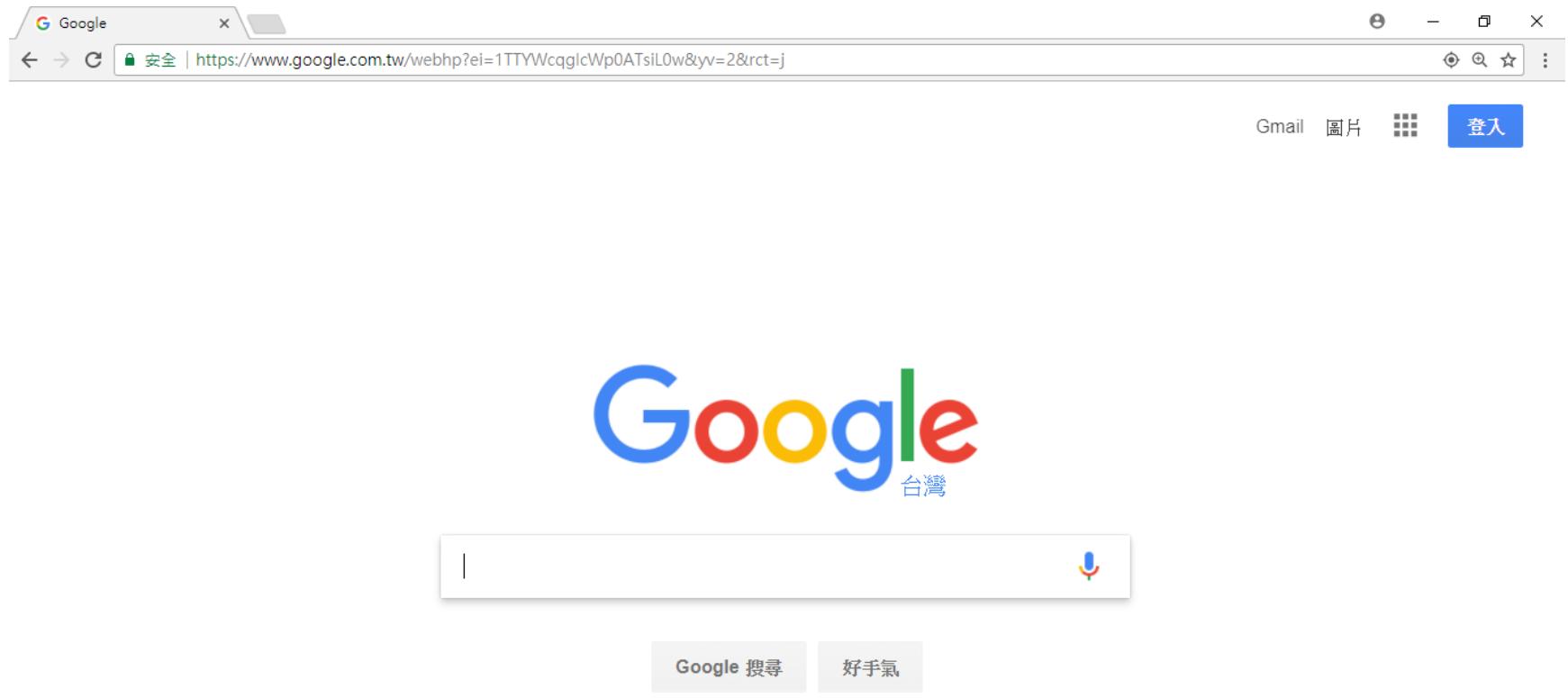
- Selected element: ``
- Element path: `html > body > div#main > div.bg > div.container > div.row > div#items > div.col2 > div#events2 > img`
- Style pane:

 - Rule: `#main responsive.css:506 @ (min-width: 600px)`
 - Element style for `.col2 #events2 img`:

 - right: 12px;
 - top: 57px;
 - position: absolute;
 - height: 180px;
 - width: 116px;

Directly click html code can also observe elements in the website

Chrome



Chrome



Chrome

The screenshot shows the Google homepage in a Chrome browser window. The address bar displays the URL <https://www.google.com.tw/webhp?ei=1TTYWcqglcWp0ATsiL0w&v=2&rct=j>. The page content features the Google logo with the word "台灣" below it. The developer tools are open at the bottom of the screen, showing the Elements tab selected. The Elements panel displays the HTML code for the page, including the Google logo and some JavaScript code. The right-hand sidebar of the developer tools shows the Styles panel with CSS rules for the body element.

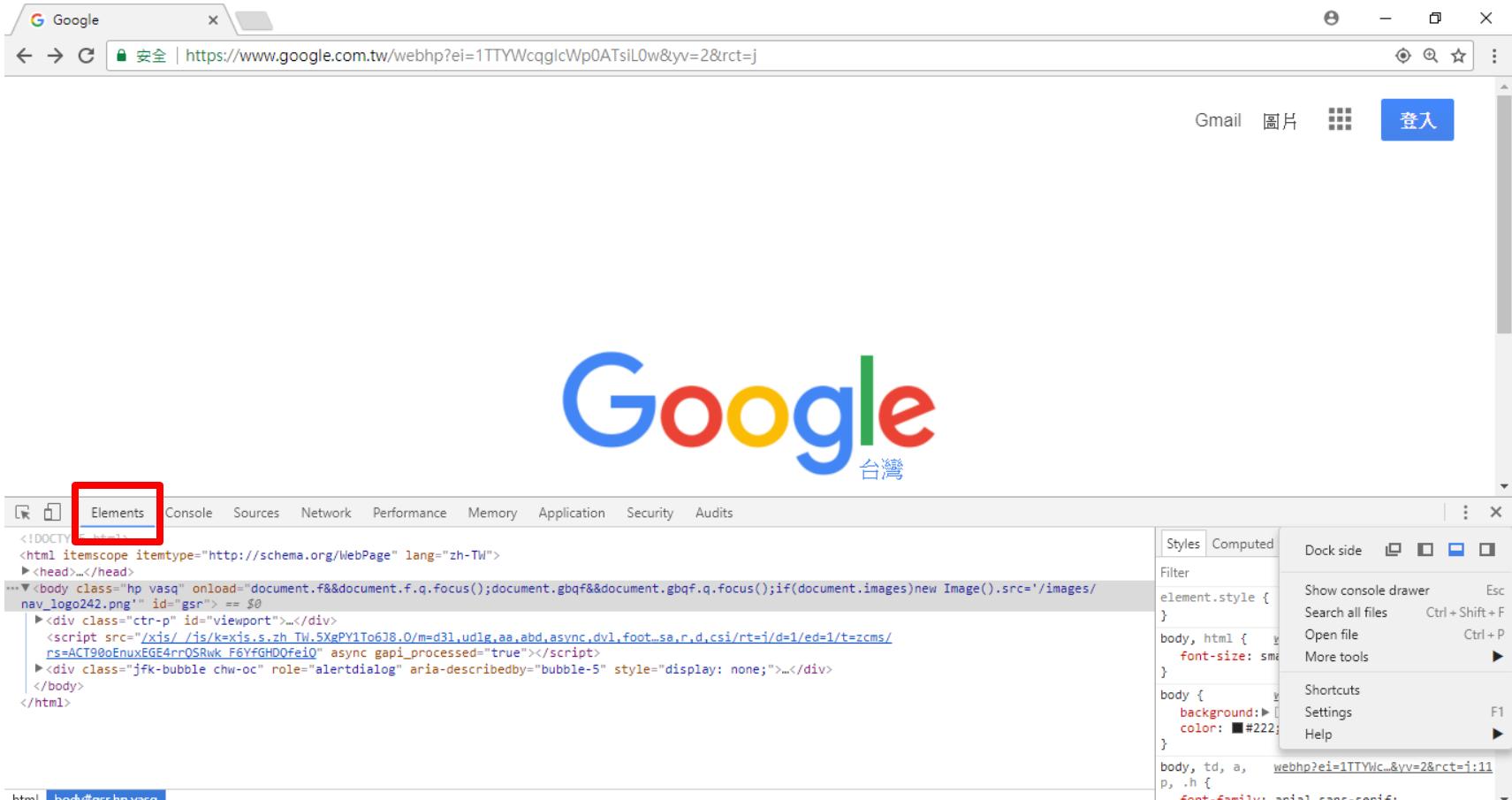
```
<!DOCTYPE html>
<html itemscope itemtype="http://schema.org/WebPage" lang="zh-TW">
  <head></head>
  <body class="hp vasq" id="gsr" style="background-color: #fff; color: #222; font-size: small; margin: 0; padding: 0; position: relative; width: 100%; z-index: 1;">
    <div class="ctr-p" id="viewport" style="background-color: #fff; height: 100%; left: 0; margin: 0; padding: 0; position: absolute; top: 0; width: 100%; z-index: 1;">
      <script src="/js/kxjs.s.js?_h_TW_5XgPY1To6J8.0/m=d31,udlg,aa,abd,async,dvl,foot_sa,r,d,csi/r=t=1/d=1/ed=1/t=zcms/rs=ACT90eEnuxEGE4rOSRwk_FGYfGHDOfeo" async=""></script>
      <div class="jfk-bubble chw-oc" role="alertdialog" aria-describedby="bubble-5" style="display: none;"></div>
    </div>
  </body>
</html>
```

Elements Console Sources Network Performance Memory Application Security Audits

Styles Computed Event Listeners

```
body, html { webhp?ei=1TTYWcqglcWp0ATsiL0w&v=2&rct=j:11; font-size: small; }
body { webhp?ei=1TTYWcqglcWp0ATsiL0w&v=2&rct=j:11; background: #fff; color: #222; }
body, td, a, webhp?ei=1TTYWcqglcWp0ATsiL0w&v=2&rct=j:11 { }
```

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/>



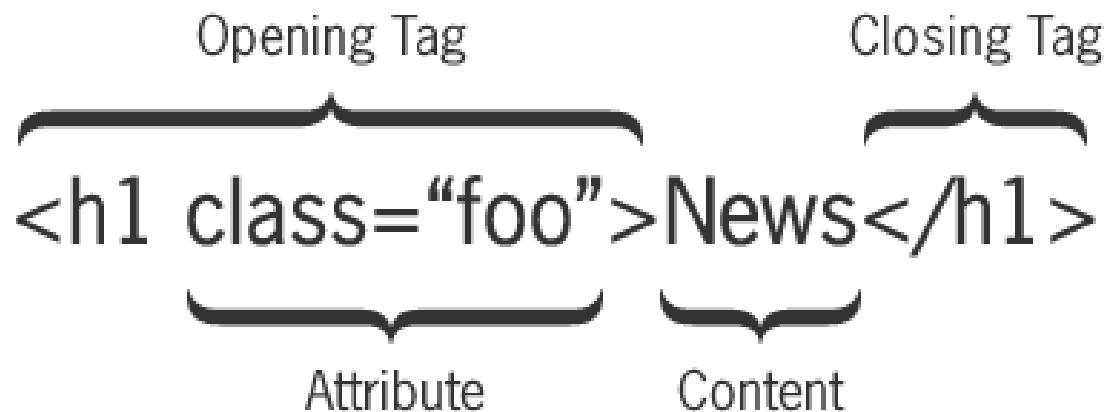
Modern web

- ▶ 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



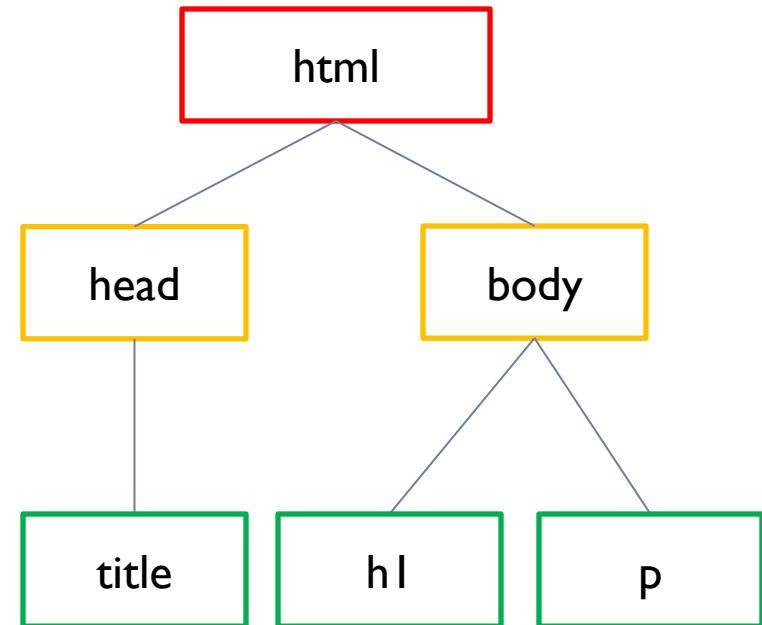
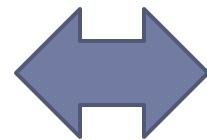
What's HTML

- ▶ **HyperText Markup Language**
 - ▶ 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 |
| | image |
| </br> | Line break(no end tag) |

•
•
•



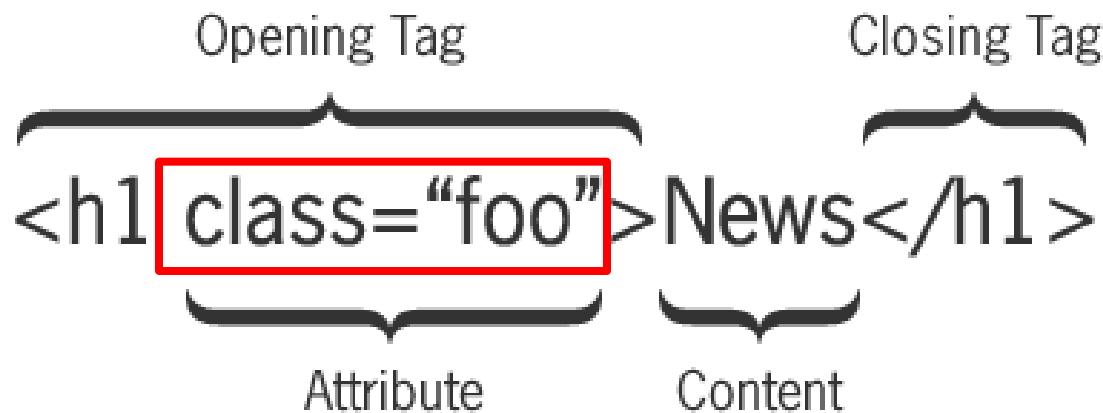
Example and Exercise (6 min)

- ▶ **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



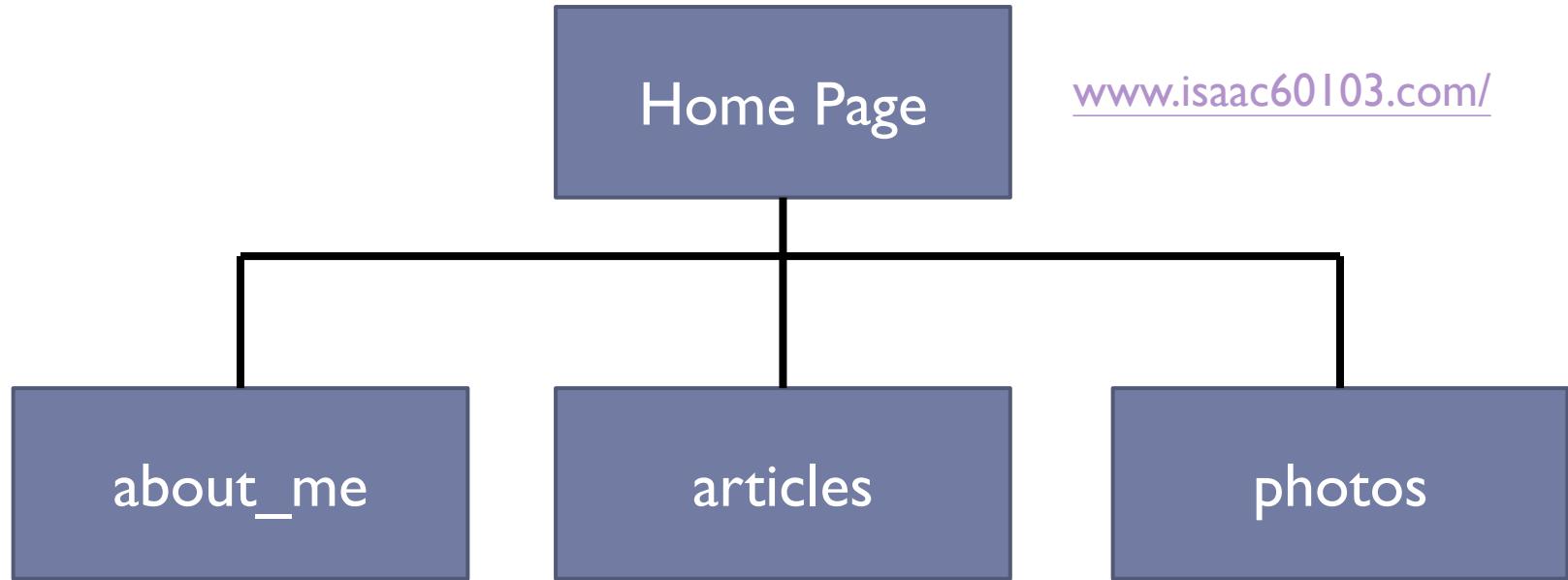
Relative url and absolute url

- ▶ 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/about_me www.isaac60103.com/articles www.isaac60103.com/photos

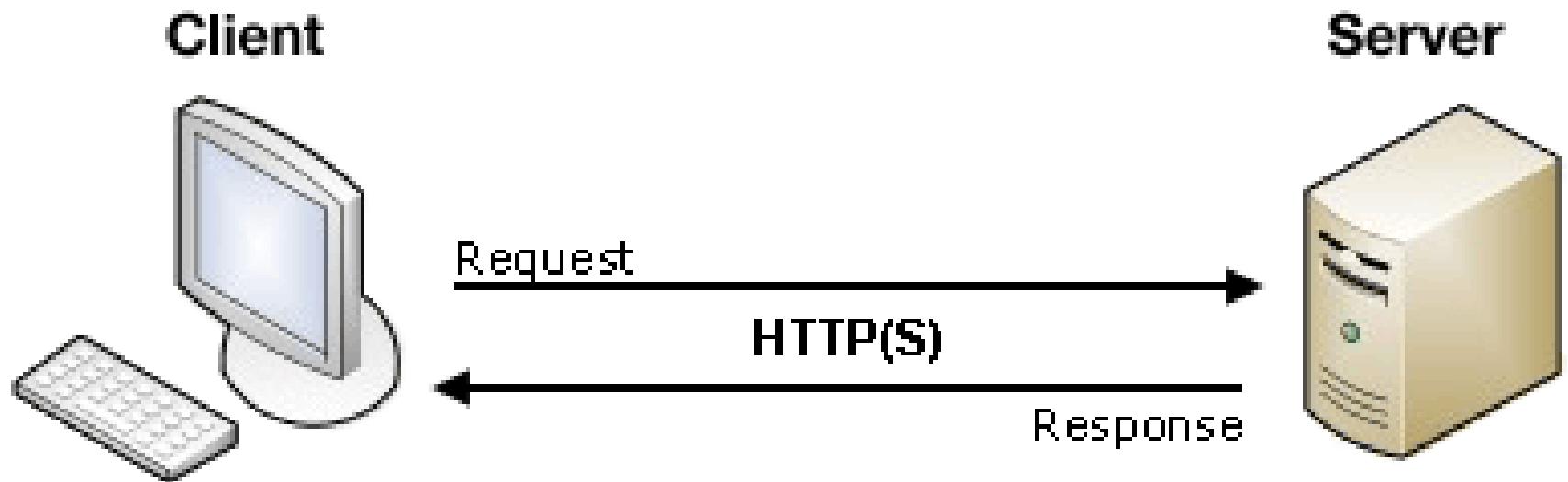


Example and Exercise

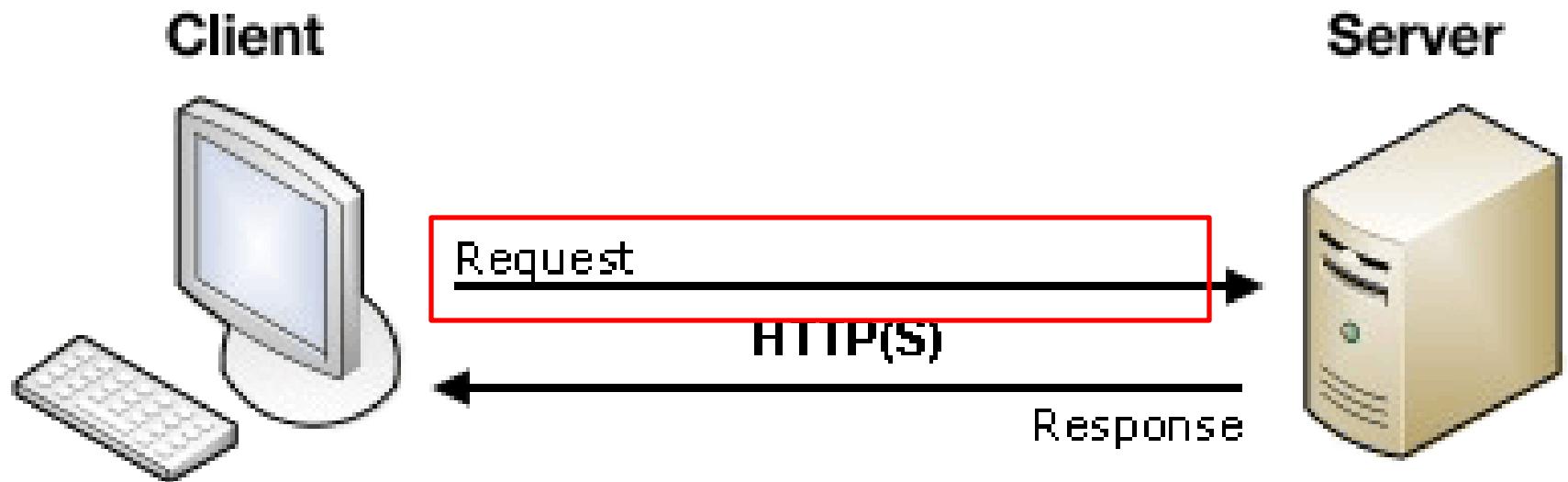
- ▶ Relative url on web page
 - ▶ <https://www.isaac60103.com/>



How to get html from server



How to get html from server

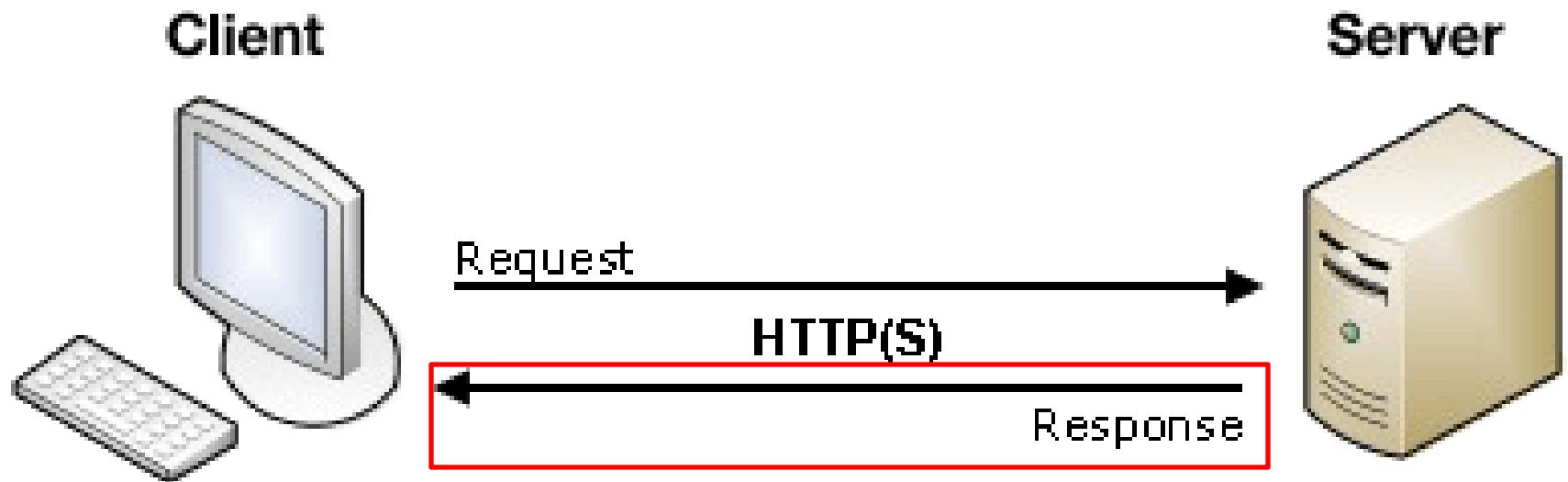


Usually request method:

- 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.ola.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>
    <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='#009999';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&sig=__ojBF1PpiI1CTk40CeWf1D1J4PyGE=" 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>
    </div>
  </div>
</header>
```

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

The screenshot shows a browser window with a search results page for "youtube" on Google. The Network tab of the developer tools is highlighted, displaying a list of network requests. A red box highlights the "Network" tab in the developer tools header and the list of requests below.

| 狀態 | 方法 | 檔案 | 網域 | 原因 | 類型 | 已傳輸 | 大小 | 0 ms | 2.56 ms | 10.12 ms | 5.12 ms | 7.68 ms | 10.12 ms |
|-------|------|-----------------------------------------------------------------------------|-------------------|----------|------|-----------|-----------|----------|---------|----------|---------|---------|----------|
| ▲ 302 | GET | search?q=youtube&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&dcr=0&ei=b... | www.google.com | document | html | 95.77 KB | 322.76 KB | → 12 ms | | | | | |
| ● 200 | GET | search?q=youtube&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&dcr=0&ei=b... | www.google.com.tw | document | html | 96.13 KB | 322.76 KB | → 413 ms | | | | | |
| ● 200 | GET | googlelogo_color_120x44dp.png | www.google.com.tw | img | png | 5.46 KB | 4.97 KB | → 20 ms | | | | | |
| ● 200 | GET | it_1967ca6a.png | ssl.gstatic.com | img | png | 7.67 KB | 7.15 KB | → 429 ms | | | | | |
| ● 200 | GET | nav_logo242.png | www.google.com.tw | img | png | 16.89 KB | 16.39 KB | → 56 ms | | | | | |
| ● 204 | POST | gen_204?&s=weabft&atyp=csi&ei=b10hWqL... | www.google.com.tw | beacon | html | 369 B | 0 B | → 18 ms | | | | | |
| ● 200 | GET | rs=ACT90oEfrTwc2QIJMrE45jnAvSfe7kvKkg... | www.google.com.tw | script | js | 137.19 KB | 402.14 KB | → 92 ms | | | | | |
| ● 200 | GET | tia.png | www.google.com | img | png | 774 B | 258 B | → 15 ms | | | | | |
| ● 204 | GET | client_204?&atyp=i&biw=1366&bih=327&e... | www.google.com.tw | img | html | 514 B | 0 B | → 39 ms | | | | | |
| ● 200 | GET | rs=ACT90oEfrTwc2QIJMrE45jnAvSfe7kvKkg... | www.google.com.tw | script | js | 62.23 KB | 186.63 KB | → 35 ms | | | | | |
| ● 200 | GET | rs=AA2YrTs_ngofTsE0VbwztD6RCCGxyiQRg | www.gstatic.com | script | js | 47.56 KB | 137.09 KB | → 591 ms | | | | | |
| ● 200 | GET | cb=gapi.loaded_0 | apis.google.com | script | js | 47.51 KB | 135.47 KB | → 461 ms | | | | | |
| ● 204 | POST | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

下方状态栏显示：15 筆請求 | 已傳輸 1.50 MB / 519.45 KB | 完成: 9.00 秒 | DOMContentLoaded: 834 ms | load: 2.09 秒

How to observe GET

The screenshot shows a Google search results page for "youtube" and the Network tab of the developer tools for the YouTube website.

Google Search Results:

- Search term: youtube
- Results: 約有 7,400,000,000 項結果 (搜尋時間: 0.33 秒)
- Top result: YouTube (<https://www.youtube.com/?gl=TW&hl=zh-tw>)

YouTube Network Tab:

- Selected Request: 200 GET [search?q=youtube&hl=zh-tw](#) (Duration: 1.28 秒)
- Request Headers (highlighted with a red box):
 - q: youtube
 - ie: utf-8
 - oe: utf-8
 - client: firefox-b
 - gfe_rd: cr
 - dcr: 0
 - ei: YWkhWtH2N7D48AfUsY_IDg
- Request Body:

```
q=youtube&hl=zh-tw
```

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

The screenshot shows a Google search results page for "youtube" and a NetworkMiner tool capturing traffic between the browser and YouTube.

Google Search Results:

- Search term: youtube
- Results: 約有 7,400,000,000 項結果 (搜尋時間: 0.38 秒)
- Top result: YouTube (https://www.youtube.com/?gl=TW&hl=zh-tw)
- Description: 在YouTube上盡情享受您喜愛的影片和音樂、上傳原創內容，並與親朋好友和全世界觀眾分享您的影片。

NetworkMiner Analysis:

- Network Tab:** Shows a list of 37 requests. The first request, a 200 OK response from YouTube, is highlighted with a red box.
- Response Tab:** A red box highlights this tab, indicating the focus of the analysis. The response content is partially visible:

```
<!doctype html><html itemscope="" itemtype="http://schema.org/Search<b>Result</b>"><head><meta http-equiv="Content-Type" content="text/html; charset=UTF-8" /><title>YouTube - 搜尋結果</title><script>var _CONFIG=[[[0,"www.gstatic.com"],[1,"www.google.com"]]]</script><script>var fa=ca[ca.length-1],ha=ba[fa],ia=ha?ha:function(a,c){if(null==t||c==t){_ja=_ja||{};_m=this;_n=function(a){return void 0!==a};_p=functi</script>
```

How to observe GET

The screenshot shows a Google search results page for "youtube" and the NetworkMiner tool interface. The NetworkMiner tool is capturing traffic for the search query. A red box highlights the first request in the list:

| 狀態 | 方法 | 標題 | 網域 | 原因 | 類型 | 已... 96.47 KB | 大小 324.02 ... | 耗時 540 ms | 往返時間 1.28 ms | 總耗時 1.91 ms |
|-----|------|------------------|----------------|---------|------|------------------|------------------|--------------|-----------------|----------------|
| 200 | GET | search?q=yout... | www.go...js | docu... | html | 96.47 KB | 324.02 ... | → 85 ms | | |
| 204 | POST | gen_204?atyp... | www.go...js | beacon | html | 260 B | 0 B | → 22 ms | | |
| 304 | GET | googlelogo.co... | www.go...js | img | png | 快取 | 4.97 KB | → 15 ms | | |
| 304 | GET | i1_1967ca6a.png | ssl.gstatic... | img | png | 快取 | 7.15 KB | → 0 ms | | |
| 304 | GET | nav_logo242.p... | www.go...js | img | png | 快取 | 16.39 KB | → 39 ms | | |
| 204 | POST | gen_204?s=we... | www.go...js | beacon | html | 369 B | 0 B | → 15 ms | | |
| 304 | GET | rs=ACT90oEfrT... | www.go...js | script | js | 快取 | 402.14 ... | → 16 ms | | |
| 304 | GET | tia.png | www.go...js | img | png | 快取 | 258 B | → 0 ms | | |
| 304 | GET | rs=ACT90oG6f... | www.go...js | script | js | 快取 | 186.63 ... | → 16 ms | | |
| 304 | GET | rs=AA2YrTs_n... | www.g...js | script | js | 快取 | 137.09 ... | → 0 ms | | |
| 304 | GET | cb=gapi.load... | apis.go...js | script | js | 快取 | 135.47 ... | → 47 ms | | |
| 204 | POST | gen_204?atyp... | www.go...js | beacon | html | 369 B | 0 B | → 15 ms | | |
| 204 | GET | ui | adservic...js | img | html | 734 B | 0 B | → 15 ms | | |
| 204 | POST | gen_204?atyp... | www.go...js | beacon | html | 369 B | 0 B | → 16 ms | | |

Request details for the highlighted row:

- 請求 URL: <https://www.google.com.tw/search?q=youtube&ie=utf-8&oe=utf-8&client>
- 請求方法: GET
- 遠端地址: 172.217.160.99:443
- 狀態代碼: 200 OK
- 版本: HTTP/2.0
- 回應標頭 (557 B):
 - alt-svc: hq=":443"; ma=2592000; quic=51...a=2592000; v="41,39,38,37,35"
 - cache-control: private, max-age=0
 - content-encoding: br
 - content-type: text/html; charset=UTF-8
 - date: Fri, 01 Dec 2017 23:44:56 GMT
 - expires: -1
 - server: gws
 - set-cookie: 1P_JAR=2017-12-01-23; expires=...path=/; domain=.google.com.tw
 - strict-transport-security: max-age=3600
 - X-Firefox-Spdy: h2
 - x-frame-options: SAMEORIGIN

NetworkMiner status bar: 14 筆請求 | 已傳輸 1.19 MB / 422.15 KB | 完成: 1.91 秒 | DOMContentLoaded: 523 ms | load: 1.09 秒

Note that parameters in “get” is append to url

How to observe POST

中文地址英譯

中文譯音拼音查詢

地址英譯寫法

中文地址英譯

● 請選擇縣市

臺北市

● 請選擇鄉鎮市區

中山區

● 道路或街名或村里名稱 [使用說明](#)

民生東路 2 段

巷 弄 號之 檢之 室

● *驗證碼

2458 如無法辨識請點此讀取

[重新產生驗證碼](#)

查詢 清除

台北市中山區民生東路二段141號號 6F

How to observe POST

漢語拼音英譯地址如下，並請參照下表詳細填寫五碼郵遞區號，俾利加速郵遞時效
No.141-6, Sec. 2, Minsheng E. Rd., Zhongshan Dist., Taipei City 104, Taiwan (R.O.C.)

其他英譯方式查詢

3+2郵遞區號對照表

| 郵遞區號 | 區域 | 路名 | 段號 | 投遞段範圍 |
|-------|-----|------|-----|----------|
| 10469 | 中山區 | 民生東路 | 2 段 | 單 71 號以下 |

檢測器 主控台 除錯器 樣式編輯器 効能 記憶體 網路 儲存空間 全部 HTML CSS JS XHR 字型 圖片 燥體 Flash WS 其他 保留紀錄 停用快取

| 狀態 | 方法 | 檔案 | 網域 | 原因 | 類型 | 已... | 大小 | 0 ms | 20.48 秒 | 40.96 秒 |
|-----|------|-------------------------------------------------------------------------------|----------|------------|------|-----------|-----------|----------|---------|---------|
| 200 | POST | index.jsp?ID=... | www.p... | docu... | html | 122.90 KB | 122.72... | → 717 ms | | |
| 200 | GET | PrintJs/rx=15... | www.p... | script | js | 5.02 KB | 5.30 KB | → 60 ms | | |
| 304 | GET | SetFont.js | www.p... | script | js | 快取 | 2.20 KB | → 52 ms | | |
| 304 | GET | sys_acckey.css | www.p... | stylesheet | css | 快取 | 630 B | → 35 ms | | |
| 304 | GET | external_style... | www.p... | stylesheet | css | 快取 | 69.91 KB | → 77 ms | | |
| 304 | GET | reset.css | www.p... | stylesheet | css | 快取 | 1.02 KB | → 73 ms | | |
| 304 | GET | jquery-migrat... | www.p... | script | js | 快取 | 9.82 KB | → 78 ms | | |
| 304 | GET | function_inne... | www.p... | script | js | 快取 | 322 B | → 80 ms | | |
| 304 | GET | hoverMenus.js | www.p... | script | js | 快取 | 1.23 KB | → 55 ms | | |
| 304 | GET | alIMG_hover.js | www.p... | script | js | 快取 | 229 B | → 55 ms | | |
| 304 | GET | Tabs_general.... | www.p... | stylesheet | css | 快取 | 5.23 KB | → 56 ms | | |
| 304 | GET | fg.menu.css | www.p... | stylesheet | css | 快取 | 4.12 KB | → 55 ms | | |
| 304 | GET | ui.theme.css | www.p... | stylesheet | css | 快取 | 16.64 KB | → 56 ms | | |
| 304 | GET | fg.drop.menu... | www.p... | stylesheet | css | 快取 | 1.01 KB | → 51 ms | | |
| 304 | GET | global.css | www.p... | stylesheet | css | 快取 | 1.57 KB | → 52 ms | | |
| 124 | 筆請求 | 已傳輸 1.40 MB / 1.39 MB 完成: 44.11 秒 DOMContentLoaded: 2.98 秒 load: 7.82 秒 | | | | | | | | |

過渡請求參數

do_s_1: 1
vKey: b7ae3bec-cec6-4b96-808d-4f84b04f5f36
showMode: 1
city: 臺北市
change_city: 2
cityarea: 中山區
street: 民生東路 2 段
lane:
alley:
num: 141
num_hyphen: 6
fl:
hyphen:
suite:
list: true
checkImage: 3406
submit: 查詢

How to observe POST

The screenshot shows a browser developer tools interface with the Network tab selected. A red box highlights the first POST request in the list:

| 狀態 | 方法 | 檔案 | 網址 | 原因 | 類型 | 已... 完成 | 大小 | 耗時 | 時間 | 追蹤堆疊 | 安全性 |
|-----|------|-------------------|----------|------------|------|------------|-----------|----------|----|------|-----|
| 200 | POST | index.jsp?ID=... | www.p... | docu... | html | 122.90 KB | 122.72... | → 717 ms | | | |
| 200 | GET | Printgold_15... | www.p... | comp... | js | 5.62 KB | 5.36 KB | → 28 ms | | | |
| 304 | GET | SetFont.js | www.p... | script | js | 快取 | 2.20 KB | → 52 ms | | | |
| 304 | GET | sys_acckey.css | www.p... | stylesheet | css | 快取 | 630 B | → 35 ms | | | |
| 304 | GET | external_style... | www.p... | stylesheet | css | 快取 | 69.91 KB | → 77 ms | | | |
| 304 | GET | reset.css | www.p... | stylesheet | css | 快取 | 1.02 KB | → 73 ms | | | |
| 304 | GET | jquery-migrat... | www.p... | script | js | 快取 | 9.82 KB | → 78 ms | | | |
| 304 | GET | function_inne... | www.p... | script | js | 快取 | 322 B | → 80 ms | | | |
| 304 | GET | hoverMenu.js | www.p... | script | js | 快取 | 1.23 KB | → 55 ms | | | |
| 304 | GET | allIMG_hover.js | www.p... | script | js | 快取 | 229 B | → 55 ms | | | |
| 304 | GET | Tabs_general.... | www.p... | stylesheet | css | 快取 | 5.23 KB | → 56 ms | | | |
| 304 | GET | fg.menu.css | www.p... | stylesheet | css | 快取 | 4.12 KB | → 55 ms | | | |
| 304 | GET | ui.theme.css | www.p... | stylesheet | css | 快取 | 16.64 KB | → 56 ms | | | |
| 304 | GET | fg.drop.menu... | www.p... | stylesheet | css | 快取 | 1.01 KB | → 51 ms | | | |
| 304 | GET | global.css | www.p... | stylesheet | css | 快取 | 1.57 KB | → 52 ms | | | |

The right panel displays the detailed response for the highlighted POST request:

- 請求 URL: <https://www.post.gov.tw/post/internet/Postal/index.jsp?ID=207#result>
- 請求方法: POST
- 遠端地址: 210.200.27.167:443
- 狀態代碼: 200 OK
- 版本: HTTP/1.1
- 回應檔頭 (179 B):
 - Content-Type: text/html; charset=UTF-8
 - Date: Fri, 01 Dec 2017 23:38:15 GMT
 - Server: Apache-Coyote/1.1
 - Transfer-Encoding: chunked
 - X-FRAME-OPTIONS: SAMEORIGIN
- 請求檔頭 (721 B):
 - Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
 - Accept-Encoding: gzip, deflate, br
 - Accept-Language: zh-TW,zh;q=0.8,en-US;q=0.5,en;q=0.3
 - Cache-Control: max-age=0

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 method in requests module

```
import requests

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

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

print(response.text)
```

Post method in requests module



Example and Exercise

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

```
<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"> [This line is highlighted]
<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]><![endif]-->
  <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

網站導覽 | 新生 | 在校學生 | 國際生 | 教職員 | 訪客 | 校友 | 聯絡我們 | 站內搜尋 | English | 學術庫 | 圖書館 | 博物館群 | 課程 | 招生 | 推廣教育 | 行事曆 | 揭款 | 認識臺大 | 學術單位 | 研究發展 | 行政組織 | 常見詢問 | 服務資源 | TOP

會期刊

流行性感冒病毒（以下簡稱流感病毒）是公共衛生的重要議題，每年在全球的肆虐奪走約50萬條寶貴的生命，以及至少300萬名重症病患，由於病毒本身具有高變異性，醫藥界在



狀態: 檢測器 主控台 除錯器 樣式編輯器 効能 記憶體 網路 儲存空間 全部 HTML CSS JS XHR 字型 圖片 媒體 Flash WS 其他 保留紀錄 停用快取 檔頭 Cookie 參數 回應 時間 追蹤堆疊

| 狀態 | 方法 | 檔案 | 網址 | 已 | 大小 | 耗時 | 操作 |
|-----|-----|--------------------|----------------------------|------|----------|------------|---------|
| 200 | GET | / | www.ntu.edu.tw/docu... | html | 32.60 KB | 32.33 KB | → 38 ms |
| 304 | GET | jquery-1.10.2... | www.ntu.edu.tw/script | js | 快取 | 90.93 KB | → 3 ms |
| 304 | GET | normalize.css | www.ntu.edu.tw/stylesheets | css | 快取 | 11.00 KB | → 2 ms |
| 304 | GET | ie8.css | www.ntu.edu.tw/stylesheets | css | 快取 | 24.74 KB | → 2 ms |
| 304 | GET | main.css | www.ntu.edu.tw/stylesheets | css | 快取 | 19.64 KB | → 1 ms |
| 304 | GET | flexslider.css | www.ntu.edu.tw/stylesheets | css | 快取 | 4.42 KB | → 1 ms |
| 304 | GET | responsivesli... | www.ntu.edu.tw/stylesheets | css | 快取 | 1.91 KB | → 23 ms |
| 304 | GET | responsive.css | www.ntu.edu.tw/stylesheets | css | 快取 | 16.70 KB | → 23 ms |
| 304 | GET | responsivesli... | www.ntu.edu.tw/script | js | 快取 | 3.36 KB | → 15 ms |
| 304 | GET | main.js | www.ntu.edu.tw/script | js | 快取 | 1.93 KB | → 25 ms |
| 304 | GET | jquery.flexslid... | www.ntu.edu.tw/script | js | 快取 | 53.31 KB | → 15 ms |
| 304 | GET | 1276_201711... | www.ntu.edu.tw/img | jpeg | 快取 | 925.87 ... | → 6 ms |

請求 URL: http://www.ntu.edu.tw/
請求方法: GET
遠端地址: 140.112.8.116:80
狀態代碼: 200 OK
版本: HTTP/1.1
回應標頭 (282 B)
Accept-Ranges: bytes
Connection: Keep-Alive
Content-Type: text/html
Date: Fri, 01 Dec 2017 14:30:08 GMT
ETag: "b26242-bf910d33f38d"
Keep-Alive: timeout=5, max=86
Server: Apache/2.2.9 (FreeBSD) mod_ssl/2.2.9 OpenSSL/0.9.8e DAV/2
Transfer-Encoding: chunked

64 等請求 已傳輸 8.75 MB / 8.74 MB | 完成: 5.99 秒 | DOMContentLoaded: 323 ms | load: 5.57 秒

Example and Exercise

▶ Post example

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

The screenshot shows a search results page for train tickets from Taipei to New竹 on December 6, 2017. It lists two train types: '區間快' (Zone Express) and '區間車' (Zone Train). The first train (1543) departs at 05:30 and arrives at 06:51, costing \$114. The second train (1107) departs at 05:39 and arrives at 07:11, also costing \$114. Below the table, there are icons for '每天行駛' (Daily Operation), '加班車' (Night Train), '跨日車' (Cross-day Train), '設身障旅客專用座位車' (Disabled Passenger Special Seats), '設有哺(集)乳室車廂' (Breastfeeding Room Carriage), and '人車同行班次' (Bicycle Transport Services). The page footer includes the address '臺北市北平西路三號' and the copyright notice '交通部臺灣鐵路管理局 版權所有 ©2012 All Rights Reserved'.

Below the search results, a browser's developer tools Network tab is displayed. A red box highlights the 'Request URL' entry: `http://twtraffic.tra.gov.tw/twtrain/TW_SearchResult.aspx`. The request details show a POST method with a response time of 222 ms. The response header section is visible, showing standard HTTP headers like Cache-Control, Content-Encoding, and Content-Length.

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

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

區間快 1543 山 南港→二水 05:30 06:51 01時21分

票價 \$ 114

●區間(快)車及普快車屬非對號車種,不提供網路訂票,請至車站購票。
●太魯閣及普悠瑪列車為自強號票價,不發售無座票。

每天行駛 加班車 跨日車 舟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)

地址:臺北市北平西路三號 電話:02-23815226(總機) | 服務監督 | (建議使用IE 10以上瀏覽以取得最佳瀏覽效果)
交通部臺灣鐵路管理局 版權所有 ©2012 All Rights Reserved

| 狀態 | 方法 | 檔案 | 網域 | 原因 | 類型 | 已... | 大小 | 0 ms | 640 ms | 1.28 秒 | 標頭 | Cookie | 參數 | 回應 | 時間 |
|-----|------|-------------------|------------------------|------|------|----------|----------|----------|--------|--------|----|--------|----|----|----|
| 200 | GET | collect?v=1&... | www.g... | img | gif | 556 B | 35 B | → 9 ms | | | | | | | |
| 200 | POST | TW_SearchRe... | twtraffi..._document | html | html | 16.35 KB | 89.17 KB | → 222 ms | | | | | | | |
| 200 | GET | kickstart.css | twtraffi..._stylesheet | css | 快取 | 10.95 KB | | | | | | | | | |
| 200 | GET | ks-style.css | twtraffi..._stylesheet | css | 快取 | 2.50 KB | | | | | | | | | |
| 200 | GET | site.css | twtraffi..._stylesheet | css | 快取 | 12.88 KB | | | | | | | | | |
| 200 | GET | pikaday.css | twtraffi..._stylesheet | css | 快取 | 4.12 KB | | | | | | | | | |
| 200 | GET | pikaday.js | twtraffi..._script | js | 快取 | 33.56 KB | | | | | | | | | |
| 200 | GET | commonlib.js | twtraffi..._script | js | 快取 | 22.23 KB | | | | | | | | | |
| 200 | GET | kickstart.js | twtraffi..._script | js | 快取 | 64.83 KB | | | | | | | | | |
| 200 | GET | trsearchresult.js | twtraffi..._script | js | 快取 | 7.81 KB | | | | | | | | | |

| 狀態 | 方法 | 檔案 | 網域 | 原因 | 類型 | 已... | 大小 | 0 ms | 1.37 分 | 2.73 分 | 4.10 | 標頭 | Cookie | 參數 | 回應 | 時間 |
|-----|------|-----------------|------------------------|------|------|----------|----------|----------|--------|--------|------|----|--------|----|----|----|
| 200 | GET | collect?v=1&... | www.g... | img | gif | 556 B | 35 B | → 9 ms | | | | | | | | |
| 200 | POST | TW_SearchRe... | twtraffi..._document | html | html | 16.35 KB | 89.17 KB | → 222 ms | | | | | | | | |
| 200 | GET | kickstart.css | twtraffi..._stylesheet | css | 快取 | 10.95 KB | | | | | | | | | | |
| 200 | GET | ks-style.css | twtraffi..._stylesheet | css | 快取 | 2.50 KB | | | | | | | | | | |
| 200 | GET | site.css | twtraffi..._stylesheet | css | 快取 | 12.88 KB | | | | | | | | | | |
| 200 | GET | pikaday.css | twtraffi..._stylesheet | css | 快取 | 4.12 KB | | | | | | | | | | |
| 200 | GET | pikaday.js | twtraffi..._script | js | 快取 | 33.56 KB | | | | | | | | | | |
| 200 | GET | commonlib.js | twtraffi..._script | js | 快取 | 22.23 KB | | | | | | | | | | |

FromStationName: 1008
ToStationName: 1025
TrainClass: 2
searchdate: 2017-12-06
FromTimeSelect: 0000
ToTimeSelect: 2359
Timetype: 1

```
367 </div>
368
369 <script>var QueryData='1008;1025;2;2017-12-06;0000;2359;1';</script>
370 {
371   "Train_Code": "1543",
372   "Class_Code": "1132",
373   "Begin_Code": "1006",
374   "Begin_Name": "南港",
375   "Begin_EName": "Nangang",
376   "End_Code": "1207",
377   "End_Name": "二水",
378   "End_EName": "Ershui",
379 }
```

Example and Exercise

▶ Some beautiful example

- ▶ https://isaac60103.github.io/crawl_course/example/beautifulsoup_example.html
- ▶ [example\crawl\example_code.pdf\(I-3\)](#)

▶ Handle one line html tags

- ▶ [example\crawl\example_code.pdf\(I-4\)](#)

```
<script src="//cdn.optimizely.com/js/128727546.js"></script><script>try {window.performance.mark("optimizelyEnd");}  
catch (err) {}</script><script>try {window.performance.mark("headEnd");} catch (err) {}</script></head><body class="pg  
pg-hidden pg-intl_homepage pg-section international t-light" data-eq-ptx="xsmall: 0, medium: 460, large: 780, full16x9:  
1100"><div class="ad ad--epic ad--all"><div id="ad_bnr_atf_02" class="ad-ad_bnr_atf_02 ad-refresh-adbody"></div></div>  
<div class="user-msg"><div class="user-msg--container"><div class="user-msg--header"><div class="user-msg--header-text"  
js-user-msg--header-text"></div><div class="user-msg--close js-user-msg--close"></div></div><div class="user-msg--body"  
<div class="user-msg--body-text js-user-msg--body-text"></div></div></div><div id="nav_plain-header" class="nav-  
plain-header"><div id="breaking-news" class="breaking-news__background"><div class="l-container"><div class="breaking-  
news"><div class="breaking-news__close-btn"></div><div class="breaking-news__title"><span class="breaking-news__title-  
text">Breaking News</span></div><div class="breaking-news__msg"></div></div></div><div id="nav" class="nav nav-  
index-0"><div class="nav_color-strip"></div><div class="nav_container"><a href="/" id="logo" class="nav__logo"></a>  
<div class="nav-section"><div class="nav-section__name js-nav-section-name" data-analytics-header="main-  
menu_intl_homepage"><a href="/">Home</a><span class="nav-section__expand-icon">+</span></div><div id="nav-section-  
submenu" class="nav-section__submenu" data-analytics-header="main-menu_intl_homepage"></div><div id="js-nav-section-  
article-title" class="js-nav-section-article-title nav-section__article-title"></div></div><div class="nav-menu-links">  
<a class="nav-menu-links__link" href="/regions" data-analytics-header="main-menu_intl_regions">Regions</a><a class="nav-  
menu-links__link" href="/politics" data-analytics-header="main-menu_politics">U.S. Politics</a><a class="nav-menu-  
links__link" href="http://money.cnn.com/INTERNATIONAL/" data-analytics-header="main-menu_money">Money</a><a class="nav-  
menu-links__link" href="/entertainment" data-analytics-header="main-menu_entertainment">Entertainment</a><a class="nav-
```



Example and Exercise (15 min)

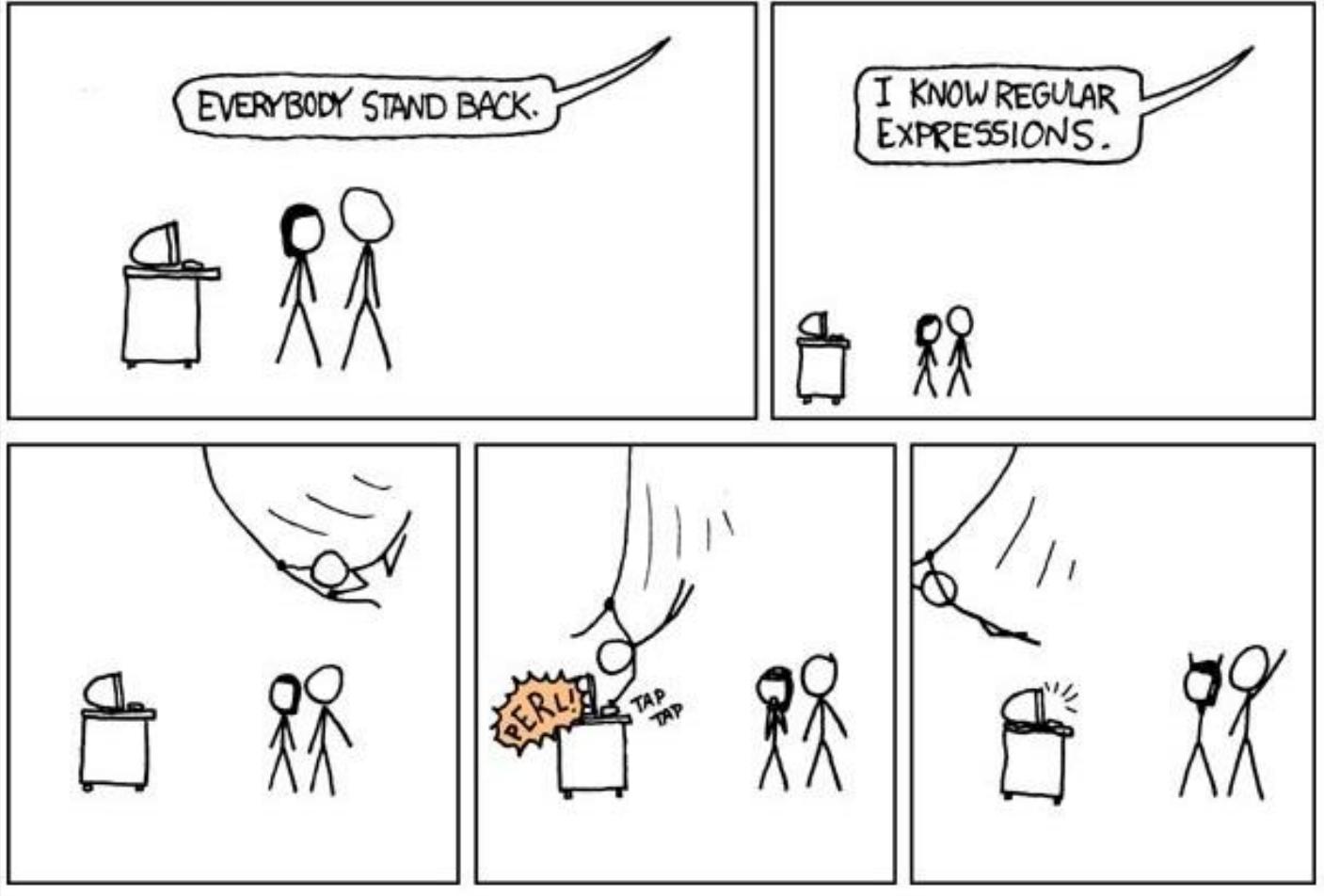
- ▶ Download requests module
 - ▶ pip install requests/ conda install requests
- ▶ Download beautiful soup module
 - ▶ pip install beautifulsoup4/ conda install beautifulsoup4
- ▶ Please find all h1 tag in the following page
 - ▶ <http://www.nccu.edu.tw/>
 - ▶ practice\crawl\practice_code.pdf(I-1)
- ▶ 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(I-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





Common Regular Expression

| symbol | meaning |
|--------|-----------------------------------|
| * | 0 or more |
| + | 1 or more |
| ? | 0 or 1 |
| {3} | Exactly 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)

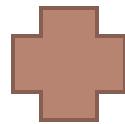
- ▶ Hello regular express
 - ▶ example\crawl\example_code.pdf(I-5)
- ▶ Grab email on web
 - ▶ https://isaac60103.github.io/crawl_course/practice/email_list.html
 - ▶ practice\crawl\practice_code.pdf(I-3)



Beautifulsoup with regular express

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

Beautifulsoup



Regular express



Example and Exercise (5 min)

- ▶ BeautifulSoup with regular express
 - ▶ example\crawl\example_code.pdf(I-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(I-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



The screenshot shows the Taiwan High-Speed Rail website. A red box highlights the orange '24h 網路訂位' (24-hour online booking) button in the top navigation bar. Another red box highlights the CSS inspector in the developer tools, focusing on the HTML code for a 24-hour booking button. The code includes a link to a 24-hour booking image.

```
<li class=">
    <a href="/tw/Article/ArticleContent/383d9d45-698b-4e17-a84f-c7cb90b5adc5" title="搭高鐵・遊台灣">搭高鐵・遊台灣</a>
<div class="menu_block"></div>
</li>
<li class=""></li>
</ul>
<div class="revision02_24hrs">
    <a href="https://tts.tsrc.com.tw/IMINI?locale=tw" alt="24小時網路訂票[將另開新視窗]" title="24小時網路訂票[將另開新視窗]" target="_blank">
        
    </a>
</div>
</div>
</div>
...</pre>
```

Please note src = XXX, which point to image location

Crawl file

臺灣證券交易所 首頁 > 交易資訊 > 盤後資訊 > 每日收盤行情

日期： 民國 106 年 12月 06日 (三) 分類： 大盤統計資訊 搜尋

※ 本資訊自民國93年2月11日起提供。

列印 / HTML CSV 下載 單位：元、股

106年12月06日 大盤統計資訊

| 指數 | 收盤指數 | 漲跌(+/-) | 漲跌點數 | 漲跌百分比(%) |
|--------|-----------|---------|--------|----------|
| 寶島股價指數 | 11,967.27 | - | 200.87 | -1.65 |

檢測器 主控台 除錯器 樣式編輯器 功能 記憶體 網路 儲存空間 搜尋 HTML 規則 計算樣式 版面 動畫 字型

```
<a class="csv" href="/exchangeReport/MI_INDEX?response=csv&date=20171206&type=MS" data-et="每日收盤行情">CSV</a>
<div class="title"><h2>106年12月06日 大盤統計資訊</h2></div>
<div id="subtitle1" class="subtitle" style="display: block;">106年12月06日 大盤統計資訊</div>
<div class="data-table" style="display: block;">
  <div id="report-table1_wrapper" class="dataTables_wrapper no-footer">
```

Please note href= XXX, which point to file location

Example and Exercise (10 min)

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

The screenshot shows a browser window with a sidebar on the left containing links like '首頁', '發燒影片', and '觀看紀錄'. Below this is a 'YOUTUBE 精選' section with categories like '音樂' and '運動'. The main content area displays a '發燒影片' section with four video thumbnails. The bottom part of the screenshot shows the Network tab of the developer tools, listing various requests made by the browser. A red box highlights the 'User-Agent' entry in the Headers section of the Network tab, which shows 'User-Agent: Mozilla/5.0 (Windows NT 6.1; W...) Gecko/20100101 Firefox/57.0'.

| 狀態 | 方法 | 檔案 | 網域 | 原因 | 類型 | 已... | 大小 | 0 ms | 10.24 秒 | 20.48 秒 |
|-----|------|-------------------|-----------------|------------|------|-----------|------------|----------|---------|---------|
| 200 | GET | ?gl=TW&hl=zh-TW | www.youtube.com | documen... | html | 140.30 KB | 819.84 ... | → 73 ms | | |
| 204 | POST | csi_204?v=2&... | www.youtube.com | beacon | html | 531 B | 0 B | → 204 ms | | |
| 200 | POST | log_event?alt... | www.youtube.com | xhr | json | 531 B | 28 B | → 58 ms | | |
| 200 | POST | log_interactio... | www.youtube.com | xhr | json | 1.05 KB | 2.11 KB | → 80 ms | | |
| 304 | GET | networkjs | www.youtube.com | script | js | 快取 | 11.59 KB | → 11 ms | | |
| 304 | GET | webcomponent... | www.youtube.com | script | js | 快取 | 39.60 KB | → 11 ms | | |
| 304 | GET | web-animatio... | www.youtube.com | script | js | 快取 | 47.43 KB | → 10 ms | | |
| 304 | GET | www-onepick... | www.youtube.com | stylesheet | css | 快取 | 841 B | → 10 ms | | |

Headers:

- 請求 URL: https://www.youtube.com/?gl=TW&hl=zh-tw
- 請求方法: GET
- 遠端地址: 172.217.160.78:443
- 狀態代碼: 200 OK
- 編輯並重新傳送
- 原始標頭
- 版本: HTTP/2.0
- 過濾標頭
- Cookie: s_gi=ca02d225c00601da00e00470...y223tvo1r, FRCT=11-3000000
- Host: www.youtube.com
- Referer: https://www.google.com.tw/
- Upgrade-Insecure-Requests: 1
- User-Agent: Mozilla/5.0 (Windows NT 6.1; W...) Gecko/20100101 Firefox/57.0

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)`

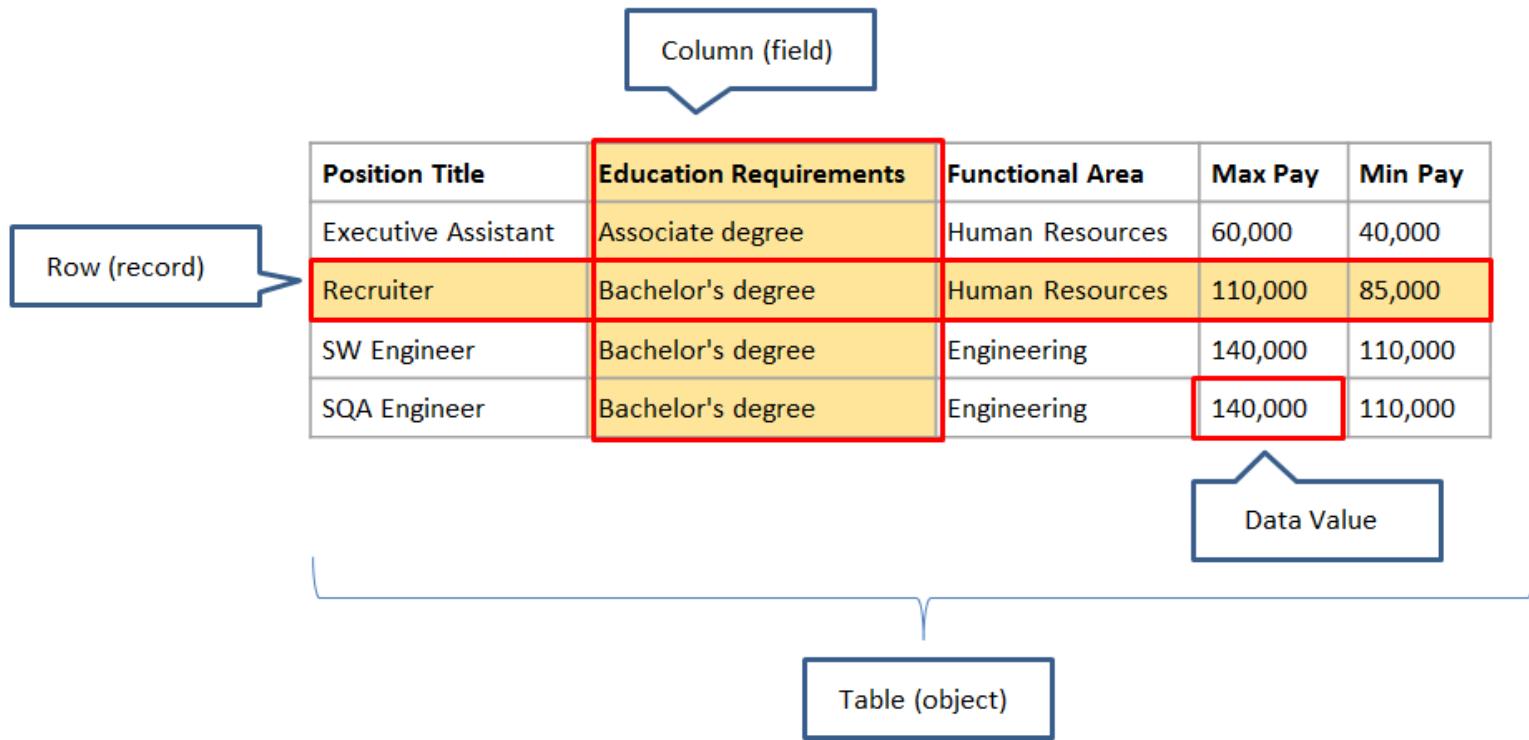


Automatically crawl

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



Database terminology



SQLite

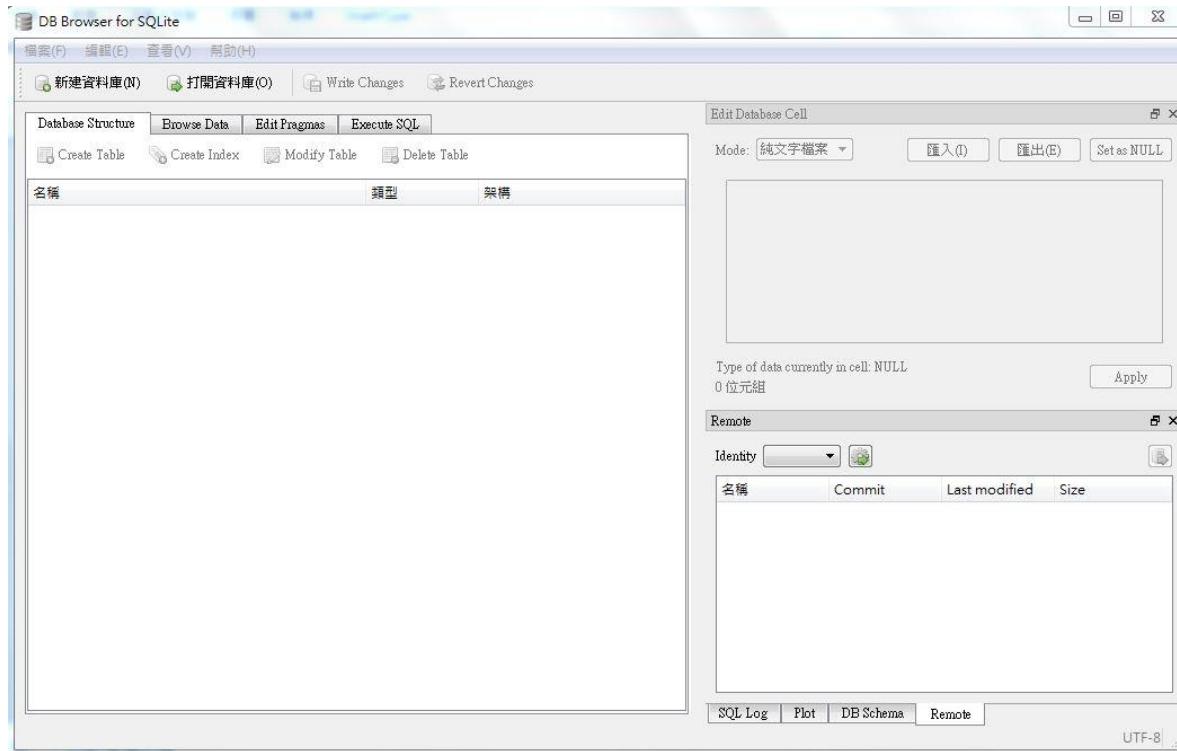
- ▶ Relational database management system
- ▶ Embedded database software for local/client storage
- ▶ Use SQL to manipulate



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 Delete Table

| 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 AUTOINCREMENT |
| 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) 汇出(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 |
|---|----|--------|--------|
| 1 | 0 | 過濾 | 過濾 |
| 2 | 1 | 20 | 10 |
| 3 | 2 | 20 | 10 |
| 4 | 3 | 20 | 10 |
| 5 | 4 | 20 | 10 |

過濾 過濾 過濾

轉到: 1

Edit Database Cell

Mode: 純文字檔案匯入(I) 汇出(E) Set as NULL

Type of data currently in cell: NULL
0 位元組

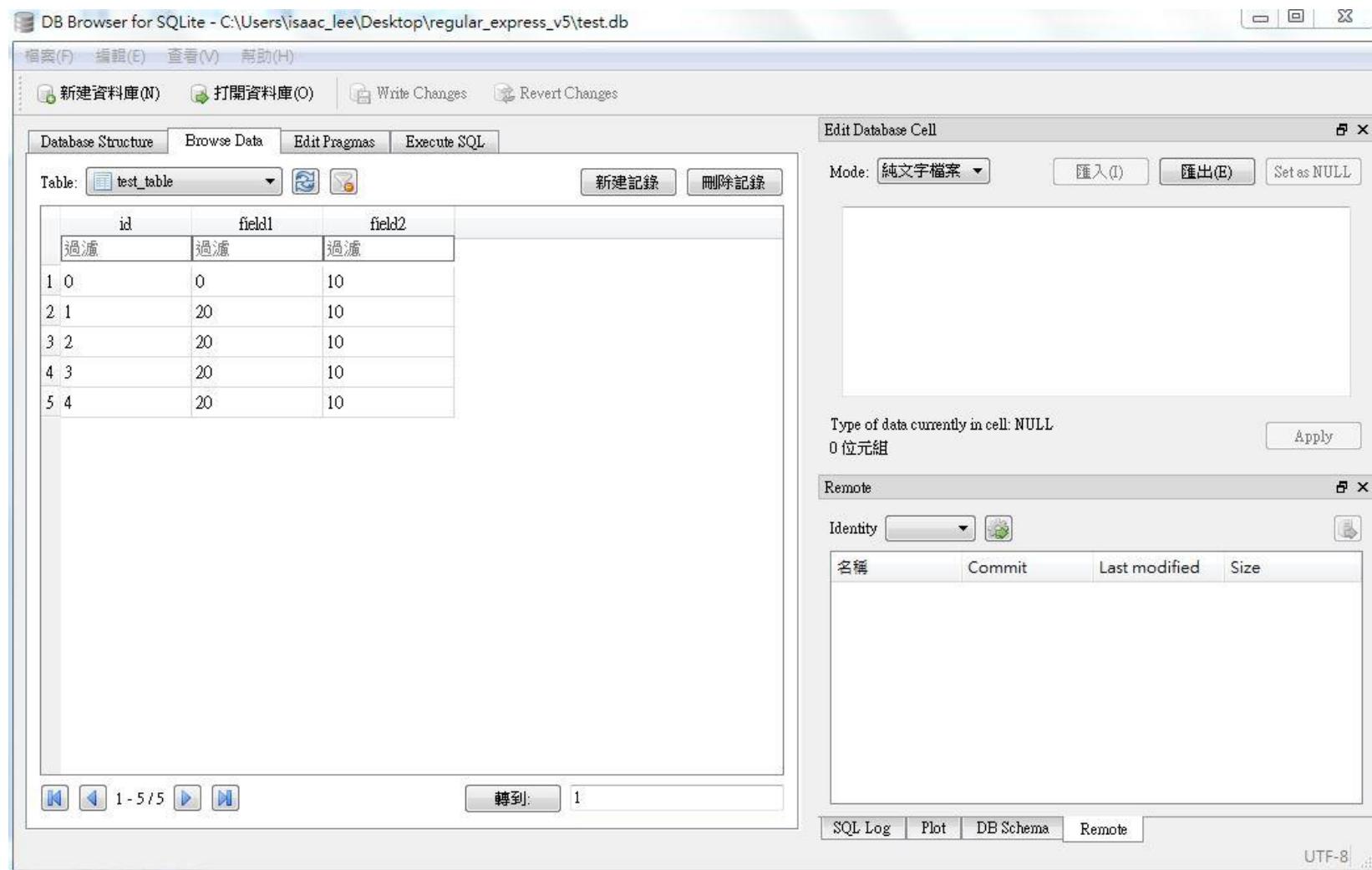
Remote

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.....



Example and Exercise (10 min)

- ▶ Use python to insert and query
 - ▶ example\crawl\example_code.pdf(I-12)
- ▶ Use panda to insert and query
 - ▶ <http://www.taipeibo.com/year/2017>
 - ▶ example\crawl\example_code.pdf(I-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(I-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

