PYTHON CRAWLER

From Beginner To Intermediate

2018 © 資訊工業策進會

Self Introduction

Cheng-Yi, Yu

erinus.startup@gmail.com

- **LifePlus Inc.**

  Technical Manager

- **Paganini Plus Inc.**

  Senior Software Developer

- **Freelancer**

  ~ 10 Years

2018 © 資訊工業策進會

DAY 1

2018 © 資訊工業策進會

Python

demo01.py

demo02.py

demo03.py

- **Entry**

```
if __name__ == '__main__':
    # do something
```

- **Method**

```
def main ():
    # do something
```

- **Package**

```
import [package]
import [package] as [alias]
```

- **Format**

```
'%s' % ([parameters ...])
```

Python

- **If … Else …**

```
if [condition]:
    # do something
else:
    # do something
```

- **For Loop**

```
for item in list:
    # do something
```

- **Array Slice**

```
array[start:end]
```

Python

demo04.py

- Array Creation From For Loop
  - Object

    `[item.attr for item in array]`

  - Dictionary

    `[item[key] for item in array]`

2018 © 資訊工業策進會

Python

- In
  - String

    ```
    if <str> in <str>:
    ```

  - Array

    ```
    if item in list:
    ```

  - Dictionary

    ```
    if <str:key> in <dict>:
    ```

2018 © 資訊工業策進會

Installation

- **Ubuntu Fonts**

  https://design.ubuntu.com/font/

- **Source Han Sans Fonts**

  https://github.com/adobe-fonts/source-han-sans

- **Visual Studio Code And Extensions**

  https://code.visualstudio.com/

Installation

- **Cmder**

  http://cmder.net/

- **Python 3.6**

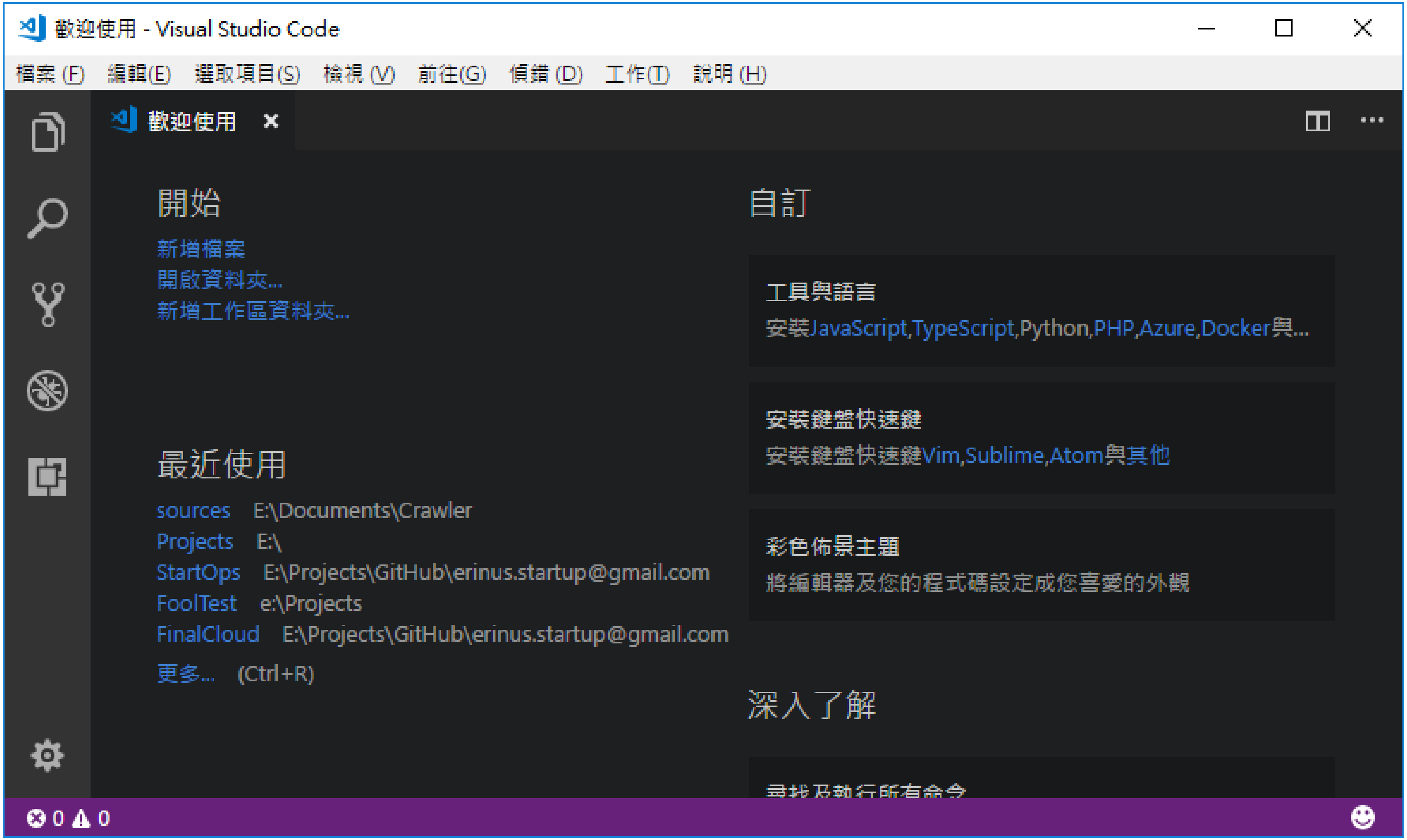
  https://www.python.org/

- **Python Packages**

  ```
  pip install requests

  pip install pyquery

  pip install beautifulsoup4

  pip install js2py

  pip install selenium
  ```

2018 © 資訊工業策進會

Visual Studio Code



歡迎使用 - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)
歡迎使用

開始

新增檔案
開啟資料夾...
新增工作區資料夾...

最近使用

sources E:\Documents\Crawler
Projects E:\
StartOps E:\Projects\GitHub\erinus.startup@gmail.com
FoolTest e:\Projects
FinalCloud E:\Projects\GitHub\erinus.startup@gmail.com
更多... (Ctrl+R)

自訂

工具與語言
安裝JavaScript,TypeScript,Python,PHP,Azure,Docker與...

安裝鍵盤快速鍵
安裝鍵盤快速鍵Vim,Sublime,Atom與其他

彩色佈景主題
將編輯器及您的程式碼設定成您喜愛的外觀

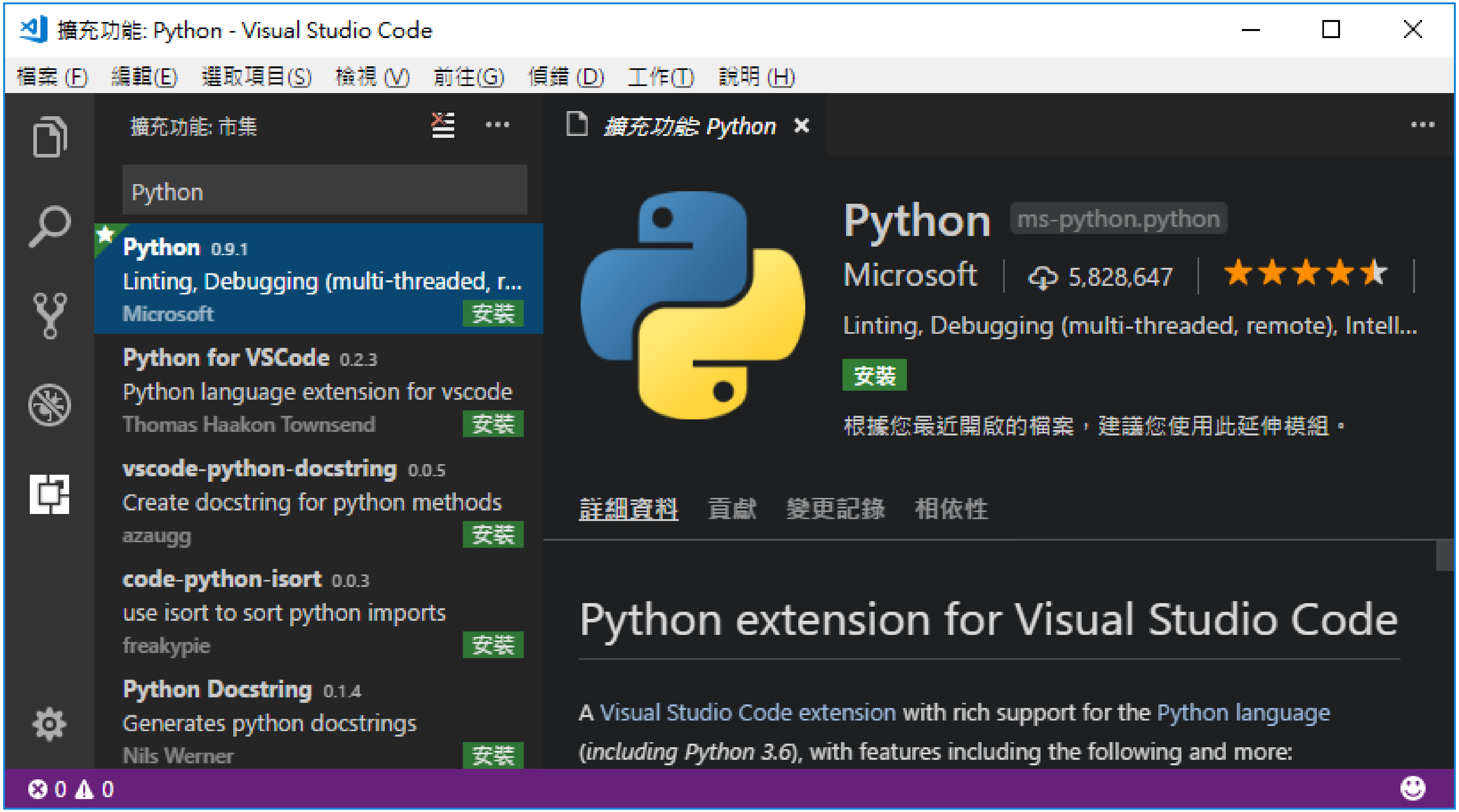
深入了解

尋找及執行所有命令

2018 © 資訊工業策進會

Visual Studio Code

- Install Python Extensions



擴充功能: Python - Visual Studio Code
檔案 (F) 編輯(E) 選取項目(S) 檢視 (V) 前往(G) 偵錯 (D) 工作(T) 說明 (H)
擴充功能: 市集
擴充功能 Python
Python

Python 0.9.1
Linting, Debugging (multi-threaded, r...
Microsoft 安裝

Python for VSCode 0.2.3
Python language extension for vscode
Thomas Haakon Townsend 安裝

vscode-python-docstring 0.0.5
Create docstring for python methods
azaugg 安裝

code-python-isort 0.0.3
use isort to sort python imports
freakypie 安裝

Python Docstring 0.1.4
Generates python docstrings
Nils Werner 安裝

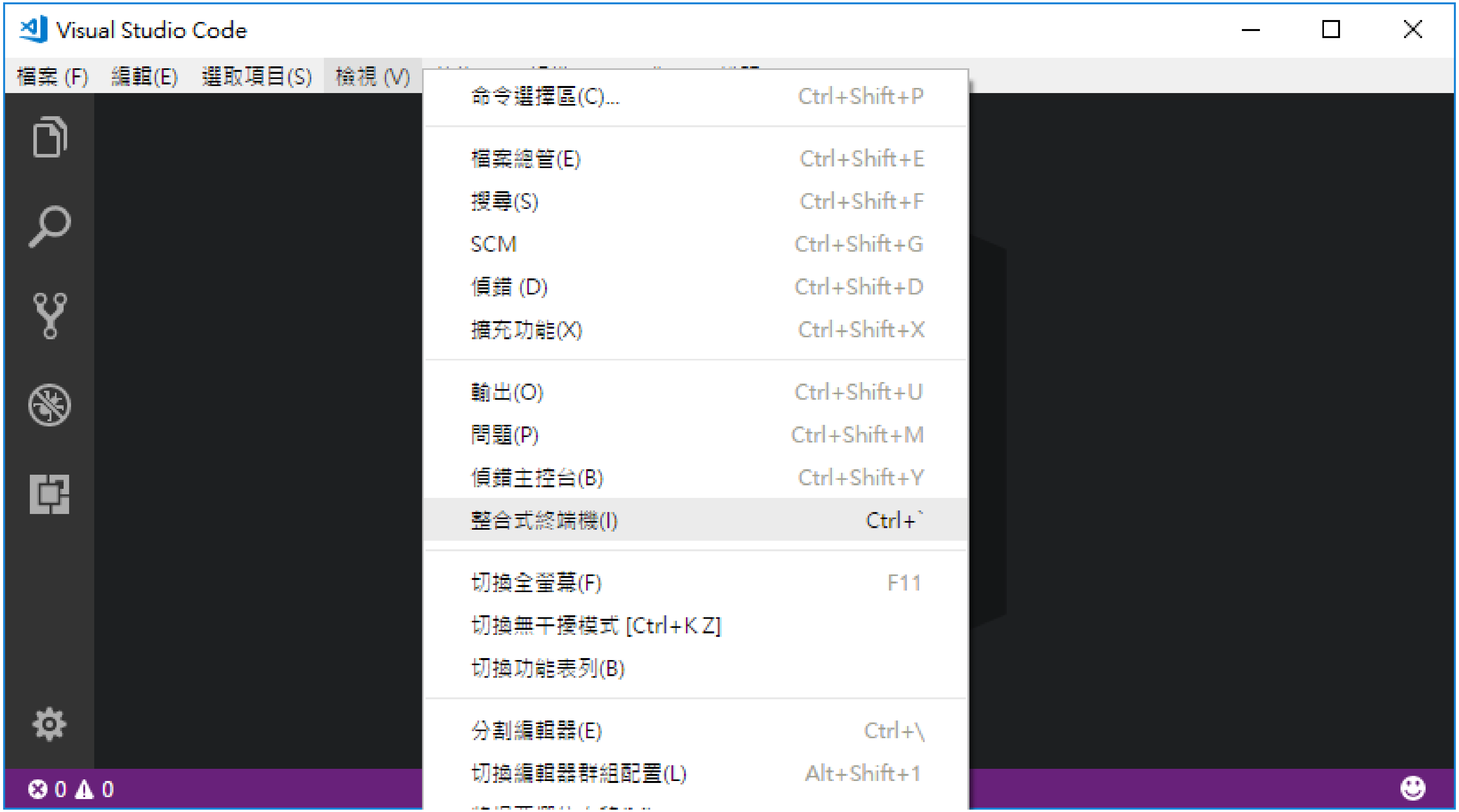
Python ms-python.python
Microsoft | 5,828,647 | ★★★★★
Linting, Debugging (multi-threaded, remote), Intell...
安裝
根據您最近開啟的檔案，建議您使用此延伸模組。

Python extension for Visual Studio Code

A Visual Studio Code extension with rich support for the Python language (*including Python 3.6*), with features including the following and more:

Visual Studio Code

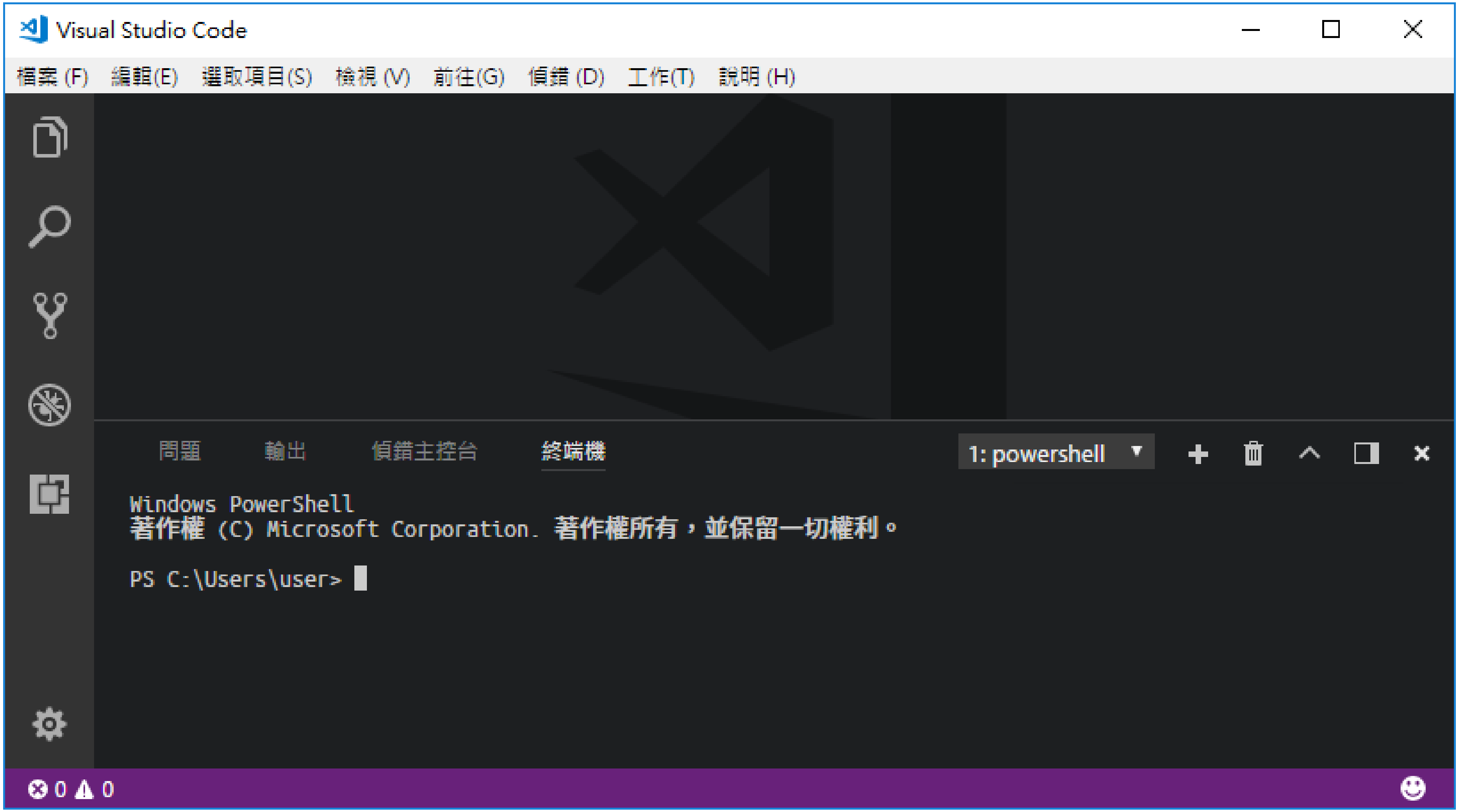
- Open Integrated Terminal



2018 © 資訊工業策進會

Visual Studio Code

- Open Integrated Terminal



```
Windows PowerShell
著作權 (C) Microsoft Corporation. 著作權所有，並保留一切權利。

PS C:\Users\user> █
```

2018 © 資訊工業策進會

Built-in

demo05.py

demo06.py

- **Json**

  import json

  - **json.loads**

    ```
    json.loads(<str>)
    ```

  - **json.dumps**

    ```
    json.dumps(<dict>)
    ```

Built-in

demo07.py

demo08.py

- **Xml**

  import xml.etree.ElementTree as ET

  - **Load From File**

    ```
    tree = ET.ElementTree(file=<str:filepath>)

    tree = ET.parse(<str:filepath>)

    root = tree.getroot()
    ```

  - **Load From String**

    ```
    root = ET.fromstring(<str>)
    ```

Built-in

demo07.py

demo08.py

- Xml
  - Child Nodes

    Only One Level

```
for node in root:
    # do something
```

  - XPath

    Multiple Levels

```
nodes = root.findall(<str:expression>)
for node in nodes:
    # do something
```

2018 © 資訊工業策進會

Built-in

- Url

  import urllib.parse as UP

  - urlparse

    result = UP.urlparse(<str:url>)

  - urlunparse

    url = UP.urlunparse(<ParseResult>)

  - quote

    str = UP.quote(<str:unquoted>)

  - unquote

    str = UP.unquote(<str:quoted>)

demo09.py

demo10.py

2018 © 資訊工業策進會

Built-in

demo11.py

- **Regular Expression**

```
import re
```

  - **re.search**

    Find First Match

```
match = re.search(<str:pattern>, <str:text>)
match.group(<int:index>)
```

  - **re.findall**

    Find All Matches

```
finds = re.findall(<str:pattern>, <str:text>)
for find in finds:
    # do something
```

Built-in

demo11.py

- Regular Expression
  - re.split

    Split By Pattern

    re.split(<str:pattern>, <str:text>)
  - re.sub

    Replace By Pattern

    re.sub(<str:pattern>, <str:replace>, <str:text>)

Built-in

demo12.py

- Regular Expression
  - Expressions

1. Range

   [Start-End]

   [0-9], [a-z], [A-Z], [a-zA-Z], [0-9a-zA-Z], ...

2018 © 資訊工業策進會

Built-in

demo12.py

- Regular Expression
  - Expressions
    1. Zero Or More Times

       *

    2. One Or More Times

       +

    3. Zero Or One Time

       ?

Built-in

demo12.py

- Regular Expression
  - Expressions
    1. Numbers

       \d = [0-9]

    2. Words

       \w = [a-zA-Z0-9] (ANSI)

       \w = [a-zA-Z0-9] + Non-ANSI Characters (UTF-8)

    3. Spaces, Tabs, …

       \s

Built-in

demo12.py

- Regular Expression
  - Expressions
    1. Start With

       ^

    2. End With

       $

2018 © 資訊工業策進會

Built-in

demo13.py

- Regular Expression
  - Expressions

1. Named Group

   (?P<name>expr)

   (?P<country>\+\d+)-(?P<phone>\d+)

2018 © 資訊工業策進會

AnalySIS

- Chrome Developer Tools
  - Elements

    See Elements In DOM

    Id, Class, Attribute, ...

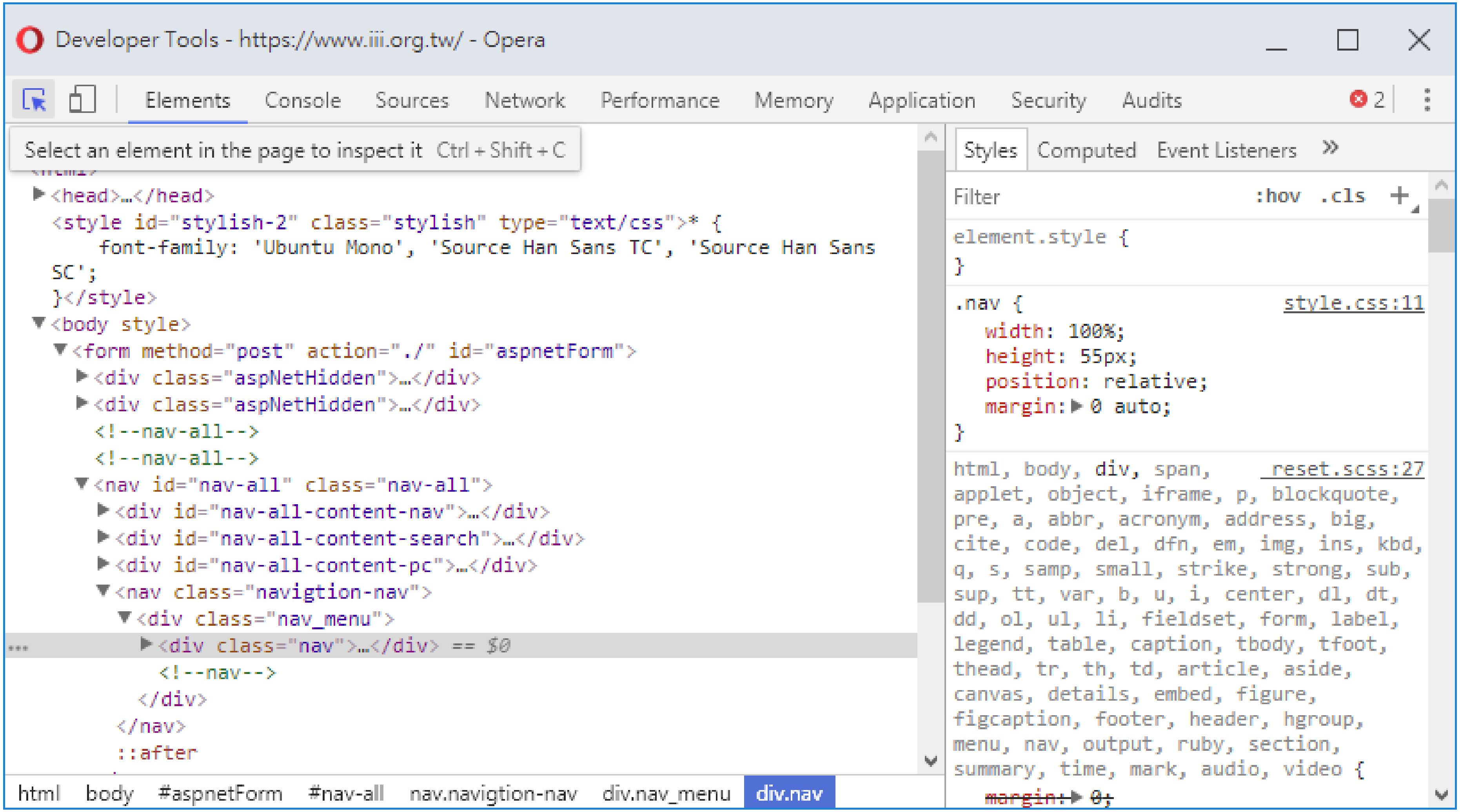
  - Network

    See Requests, Responses

    Urls, Methods, Headers, Cookies, Bodies, ...

Elements

- Find Element by Mouse Pointer



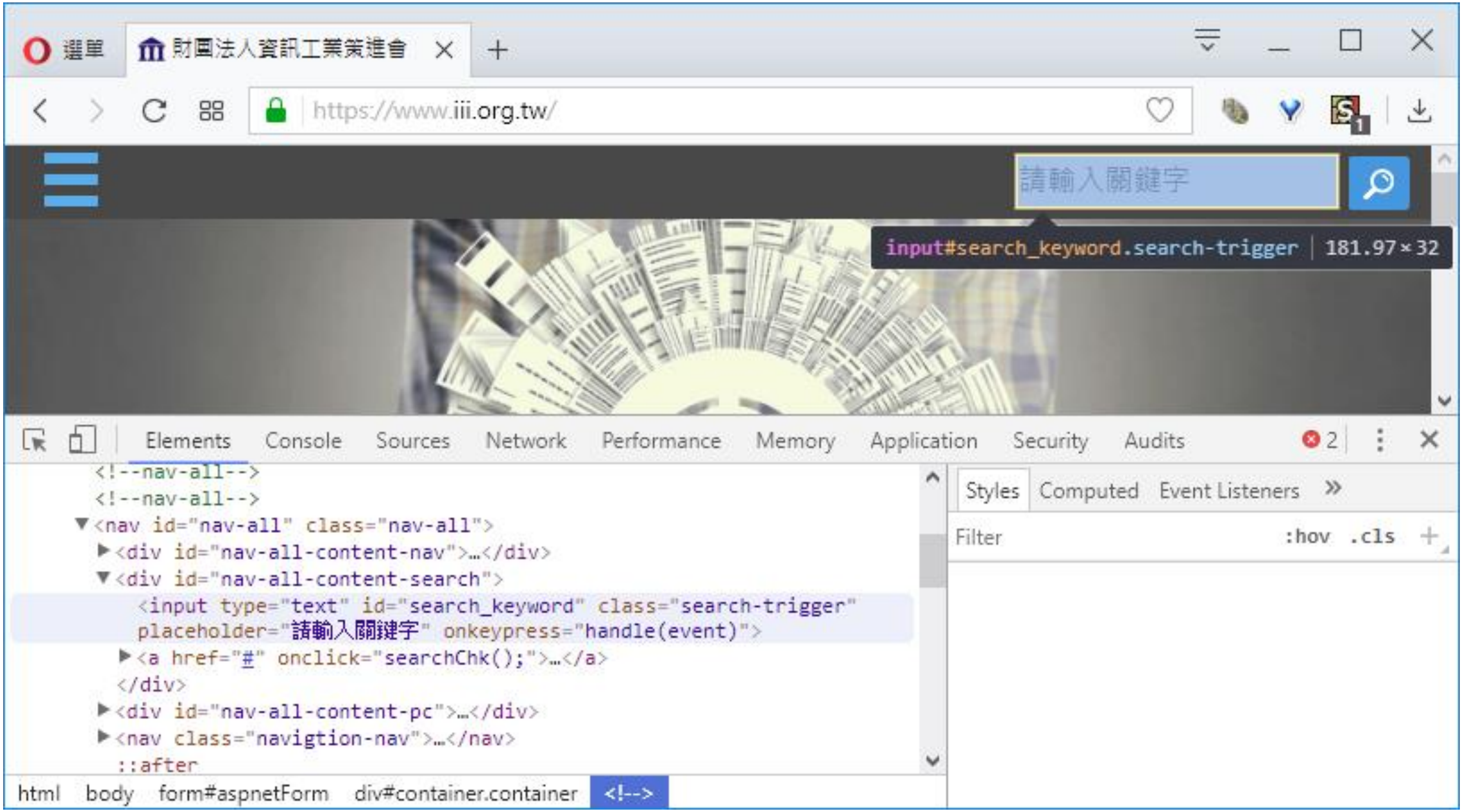
Select an element in the page to inspect it Ctrl + Shift + C

```
 ▶<head>…</head>
  <style id="stylish-2" class="stylish" type="text/css">* {
      font-family: 'Ubuntu Mono', 'Source Han Sans TC', 'Source Han Sans
  SC';
  }</style>
 ▼<body style>
   ▼<form method="post" action="./" id="aspnetForm">
     ▶<div class="aspNetHidden">…</div>
     ▶<div class="aspNetHidden">…</div>
      <!--nav-all-->
      <!--nav-all-->
     ▼<nav id="nav-all" class="nav-all">
       ▶<div id="nav-all-content-nav">…</div>
       ▶<div id="nav-all-content-search">…</div>
       ▶<div id="nav-all-content-pc">…</div>
       ▼<nav class="navigtion-nav">
         ▼<div class="nav_menu">
...        ▶<div class="nav">…</div> == $0
            <!--nav-->
          </div>
        </nav>
        ::after
```

```
element.style {
}
.nav {                                    style.css:11
   width: 100%;
   height: 55px;
   position: relative;
   margin:▶ 0 auto;
}
html, body, div, span,                  reset.scss:27
applet, object, iframe, p, blockquote,
pre, a, abbr, acronym, address, big,
cite, code, del, dfn, em, img, ins, kbd,
q, s, samp, small, strike, strong, sub,
sup, tt, var, b, u, i, center, dl, dt,
dd, ol, ul, li, fieldset, form, label,
legend, table, caption, tbody, tfoot,
thead, tr, th, td, article, aside,
canvas, details, embed, figure,
figcaption, footer, header, hgroup,
menu, nav, output, ruby, section,
summary, time, mark, audio, video {
   ~~margin:▶ 0;~~
```

Elements

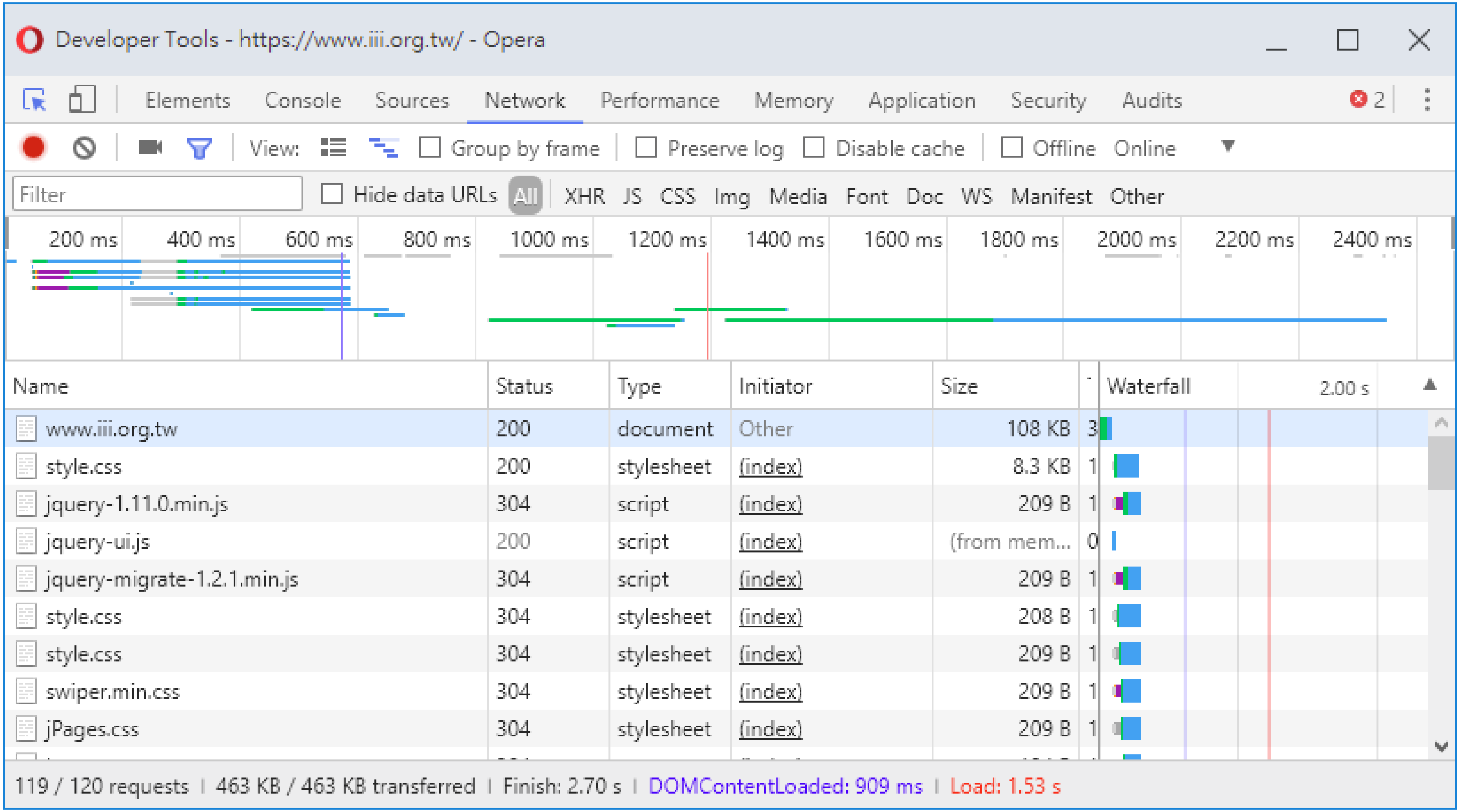
- Find Element by HTML Tag



```
    <!--nav-all-->
    <!--nav-all-->
  ▼<nav id="nav-all" class="nav-all">
    ▶<div id="nav-all-content-nav">…</div>
    ▼<div id="nav-all-content-search">
        <input type="text" id="search_keyword" class="search-trigger"
        placeholder="請輸入關鍵字" onkeypress="handle(event)">
      ▶<a href="#" onclick="searchChk();">…</a>
      </div>
    ▶<div id="nav-all-content-pc">…</div>
    ▶<nav class="navigtion-nav">…</nav>
      ::after
```

Networks

- See All Requests And Responses



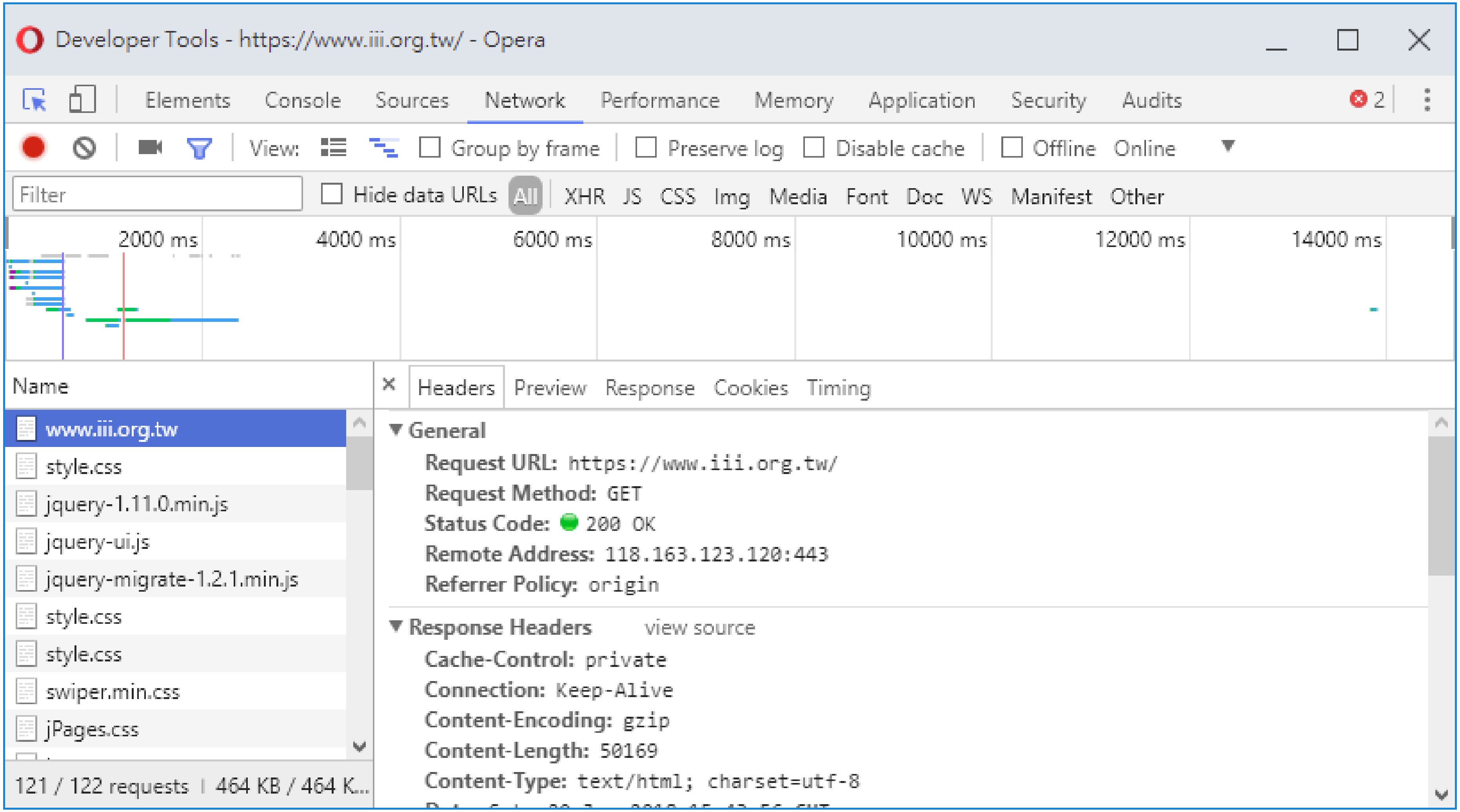
Developer Tools - https://www.iii.org.tw/ - Opera

Name	Status	Type	Initiator	Size		Waterfall
www.iii.org.tw	200	document	Other	108 KB	3	
style.css	200	stylesheet	(index)	8.3 KB	1	
jquery-1.11.0.min.js	304	script	(index)	209 B	1	
jquery-ui.js	200	script	(index)	(from mem...	0	
jquery-migrate-1.2.1.min.js	304	script	(index)	209 B	1	
style.css	304	stylesheet	(index)	208 B	1	
style.css	304	stylesheet	(index)	209 B	1	
swiper.min.css	304	stylesheet	(index)	209 B	1	
jPages.css	304	stylesheet	(index)	209 B	1	

119 / 120 requests | 463 KB / 463 KB transferred | Finish: 2.70 s | DOMContentLoaded: 909 ms | Load: 1.53 s

Networks

- See Details Of Request And Response



Name

- www.iii.org.tw
- style.css
- jquery-1.11.0.min.js
- jquery-ui.js
- jquery-migrate-1.2.1.min.js
- style.css
- style.css
- swiper.min.css
- jPages.css

121 / 122 requests | 464 KB / 464 K...

▼ General

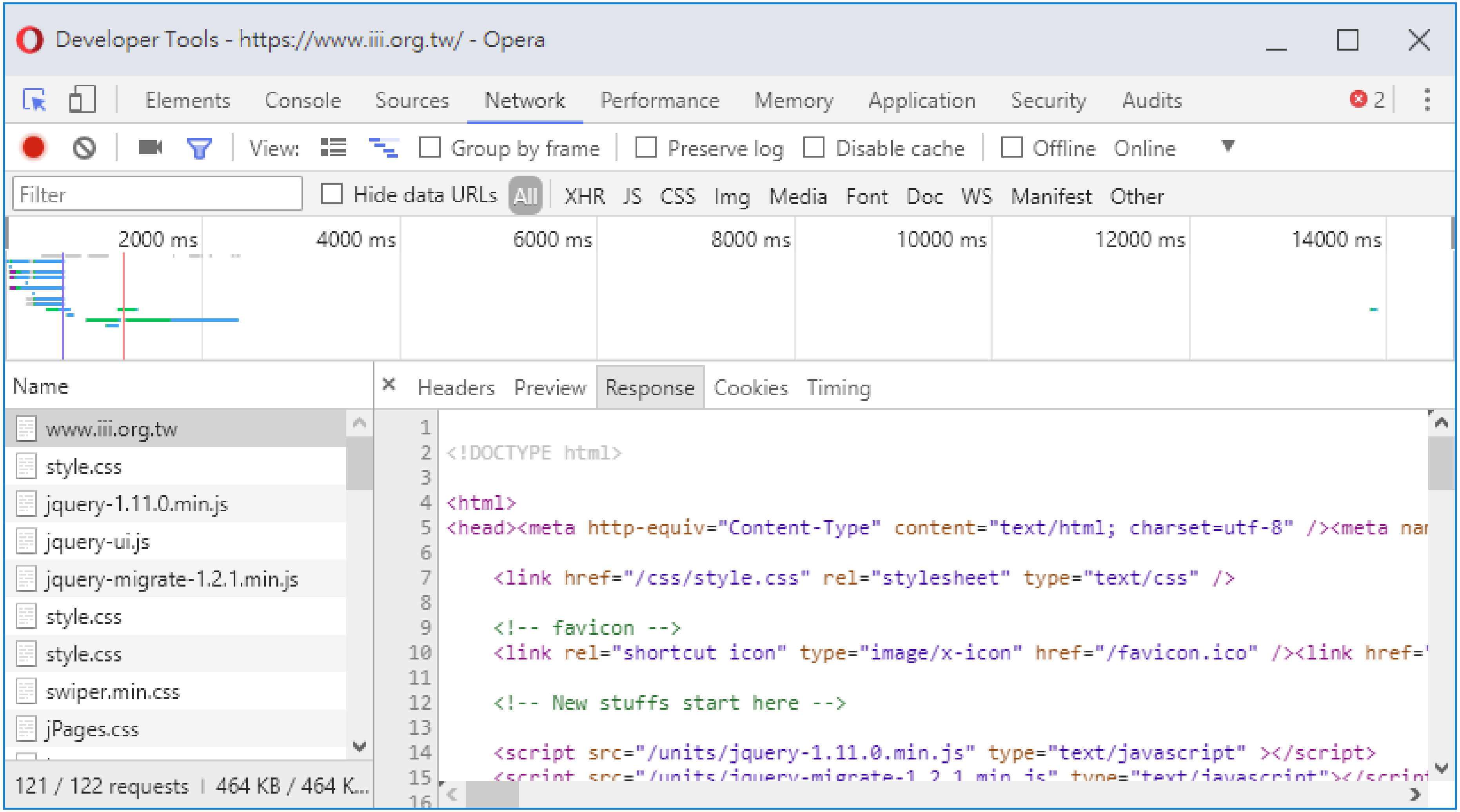
Request URL: https://www.iii.org.tw/
Request Method: GET
Status Code: ● 200 OK
Remote Address: 118.163.123.120:443
Referrer Policy: origin

▼ Response Headers

Cache-Control: private
Connection: Keep-Alive
Content-Encoding: gzip
Content-Length: 50169
Content-Type: text/html; charset=utf-8

Networks

- See Response Content



Name

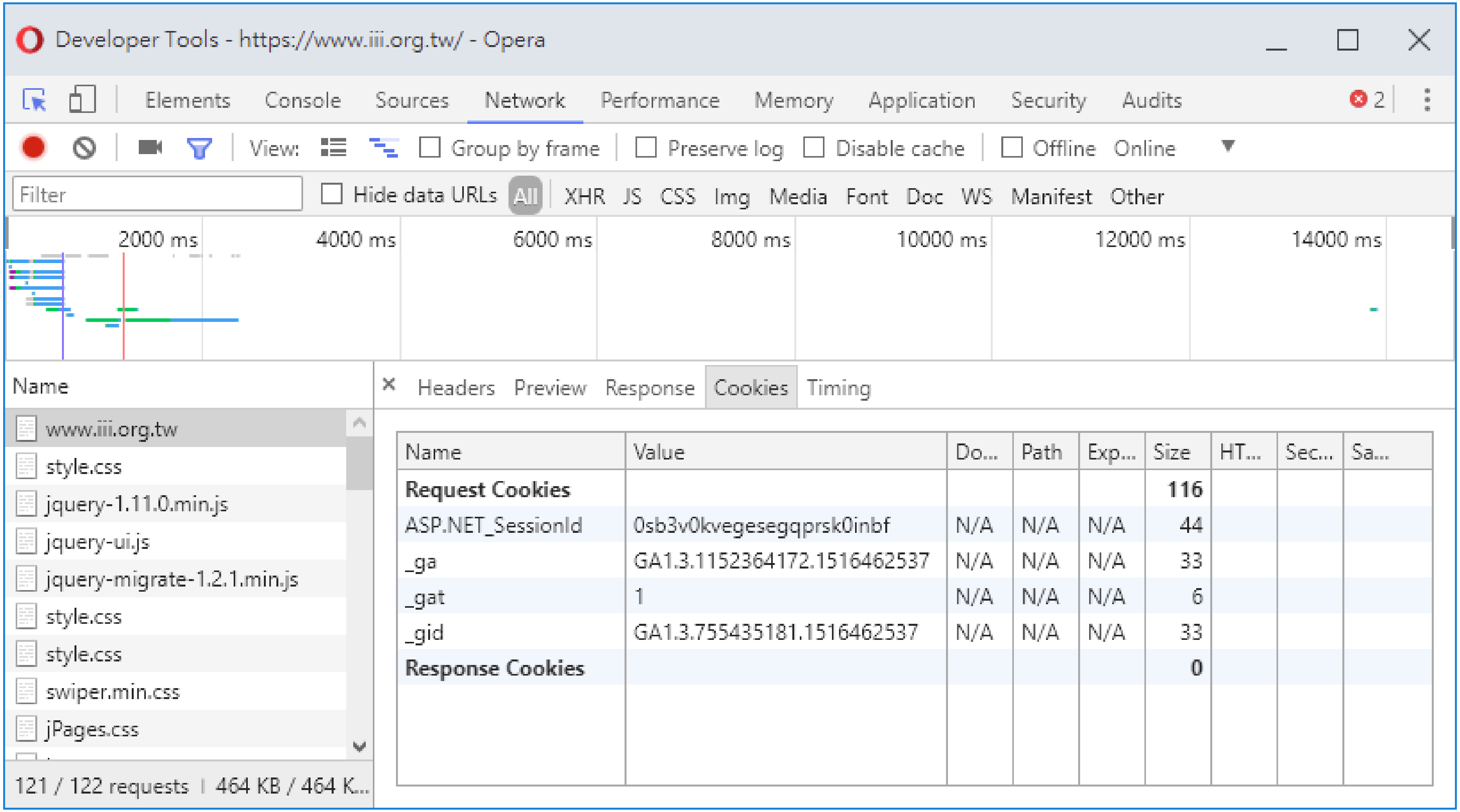
- www.iii.org.tw
- style.css
- jquery-1.11.0.min.js
- jquery-ui.js
- jquery-migrate-1.2.1.min.js
- style.css
- style.css
- swiper.min.css
- jPages.css

121 / 122 requests | 464 KB / 464 K...

```

<!DOCTYPE html>

<html>
<head><meta http-equiv="Content-Type" content="text/html; charset=utf-8" /><meta nar

    <link href="/css/style.css" rel="stylesheet" type="text/css" />

    <!-- favicon -->
    <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico" /><link href='

    <!-- New stuffs start here -->

    <script src="/units/jquery-1.11.0.min.js" type="text/javascript" ></script>
    <script src="/units/jquery-migrate-1.2.1.min.js" type="text/javascript"></scrip
```

Networks

- **See Cookies Sent And Set**



Name

- www.iii.org.tw
- style.css
- jquery-1.11.0.min.js
- jquery-ui.js
- jquery-migrate-1.2.1.min.js
- style.css
- style.css
- swiper.min.css
- jPages.css

121 / 122 requests | 464 KB / 464 K...

Name	Value	Do...	Path	Exp...	Size	HT...	Sec...	Sa...
Request Cookies					**116**			
ASP.NET_SessionId	0sb3v0kvegesegqprsk0inbf	N/A	N/A	N/A	44			
_ga	GA1.3.1152364172.1516462537	N/A	N/A	N/A	33			
_gat	1	N/A	N/A	N/A	6			
_gid	GA1.3.755435181.1516462537	N/A	N/A	N/A	33			
Response Cookies					**0**			

Documents

- **Requests**

  http://docs.python-requests.org/
- **PyQuery**

  https://pythonhosted.org/pyquery/
- **Beautiful Soup 4**

  https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- **Js2Py**

  https://github.com/PiotrDabkowski/Js2Py

Packages

demo14.py

- **Requests**

  import requests

  - **Request**

    1. **Method**

       GET, POST, …

       response = requests.get(<str:url>)

       response = requests.post(<str:url>, data=<str:body>)

       response = requests.post(<str:url>, data=<dict:body>)

    2. **Session**

       session = requests.Session()

       response = session.get(<str:url>)

Packages

- Requests

  import requests

  - Request

    - Headers

      response = requests.get(<str:url>, headers=<dict>)

    - Cookies

      response = requests.get(<str:url>, cookies=<dict>)

2018 © 資訊工業策進會

Packages

demo14.py

- Requests
  - Response

1. Status Code

   response.status_code

2. Headers

   response.headers[<str:name>]

3. Cookies

   response.cookies[<str:name>]

2018 © 資訊工業策進會

Packages

demo14.py

- Requests
  - Response
    1. Binary Content

       response.content
    2. Text Content

       response.text
    3. Json Content

       response.json()

2018 © 資訊工業策進會

Packages

demo15.py

- **PyQuery**

  import pyquery

  - **Load From String**

    d = pyquery.PyQuery(<str:html>)

  - **Load From Url**

    d = pyquery.PyQuery(url=<str:url>)

  - **Load From File**

    d = pyquery.PyQuery(filename=<str:filepath>)

2018 © 資訊工業策進會

Packages

demo15.py

- PyQuery
  - Find

    ```
    p = d(<str:expression>)
    ```

  - Element To HTML

    ```
    p.html()
    ```

  - Extract Text From Element

    ```
    p.text()
    ```

  - Get Value From Element Attribute

    ```
    val = p.attr[<str:name>]
    ```

2018 © 資訊工業策進會

Packages

demo16.py

- ## Beautiful Soup 4

import bs4

  - ### Load From String

d = bs4.BeautifulSoup(<str:html>, 'html.parser')

Packages

- **Beautiful Soup 4**
  - **Find**

demo16.py

demo17.py

```
p = d.find_all(<str:tag>, <attr-key>=<attr-val>, ...)
p = d.find_all(<regex>, <attr-key>=<attr-val>, ...)
p = d.find_all(<array>, <attr-key>=<attr-val>, ...)
p = d.find(<str:tag>, <attr-key>=<attr-val>, ...)
p = d.find(<regex>, <attr-key>=<attr-val>, ...)
p = d.find(<array>, <attr-key>=<attr-val>, ...)
p = d.select(<str:expression>)
p = d.select_one(<str:expression>)
```

Packages

demo16.py

demo17.py

- Beautiful Soup 4
  - Extract Text From Element

    ```
    p.get_text()
    ```

  - Get Value From Element Attribute

    ```
    p.get(<str:name>)
    ```

2018 © 資訊工業策進會

Packages

demo18.py

- **Js2Py**

  import js2py

  - **Eval**

    js2py.eval_js(<str:code>)

    res = js2py.eval_js('var o = <str:js>; o')

2018 © 資訊工業策進會

DAY 2

2018 © 資訊工業策進會

WORKSHOP

- Apple Daily

  https://tw.appledaily.com/

  - Realtime News

    https://tw.appledaily.com/new/realtime

WORKSHOP

- Facebook Page
  - Cookies
  - Feed

2018 © 資訊工業策進會

DAY 3

2018 © 資訊工業策進會

SELENIUM

- Download ChromeDriver

  https://sites.google.com/a/chromium.org/chromedriver/

SELENIUM

- ## Import

  import selenium.webdriver

- ## Initialize

  option = elenium.webdriver.ChromeOptions()

- ## Start

  driver = selenium.webdriver.Chrome(chrome_options=option)

- ## Browse

  driver.get(<str:url>)

- ## Close

  driver.quit()

2018 © 資訊工業策進會

SELENIUM

- Source

```
driver.page_source
```

2018 © 資訊工業策進會

SELENIUM

- Find One
  - find_element_by_id
  - find_element_by_name
  - find_element_by_tag_name
  - find_element_by_class_name
  - find_element_by_css_selector

SELENIUM

- Find Multiple
  - find_elements_by_name
  - find_elements_by_tag_name
  - find_elements_by_class_name
  - find_elements_by_css_selector

SELENIUM

- Actions
  - send_keys
  - click

2018 © 資訊工業策進會

WORKSHOP

- Facebook Page
  - Login
  - Feed

THANK YOU