Python & Data - Week 9

打包下載

堂上作業

一、目的：請找一個網站嘗試做 Web Scraping。運用 Week 8 提過的方法去判斷是否適合。
二、形式：二人一組完成
三、提交（展示）內容

1. 網址
2. 具有網頁頁面、網址
3. 希望截取的內容及其 HTML 片段
4. 用途 四、提交日期：未定

本期內容

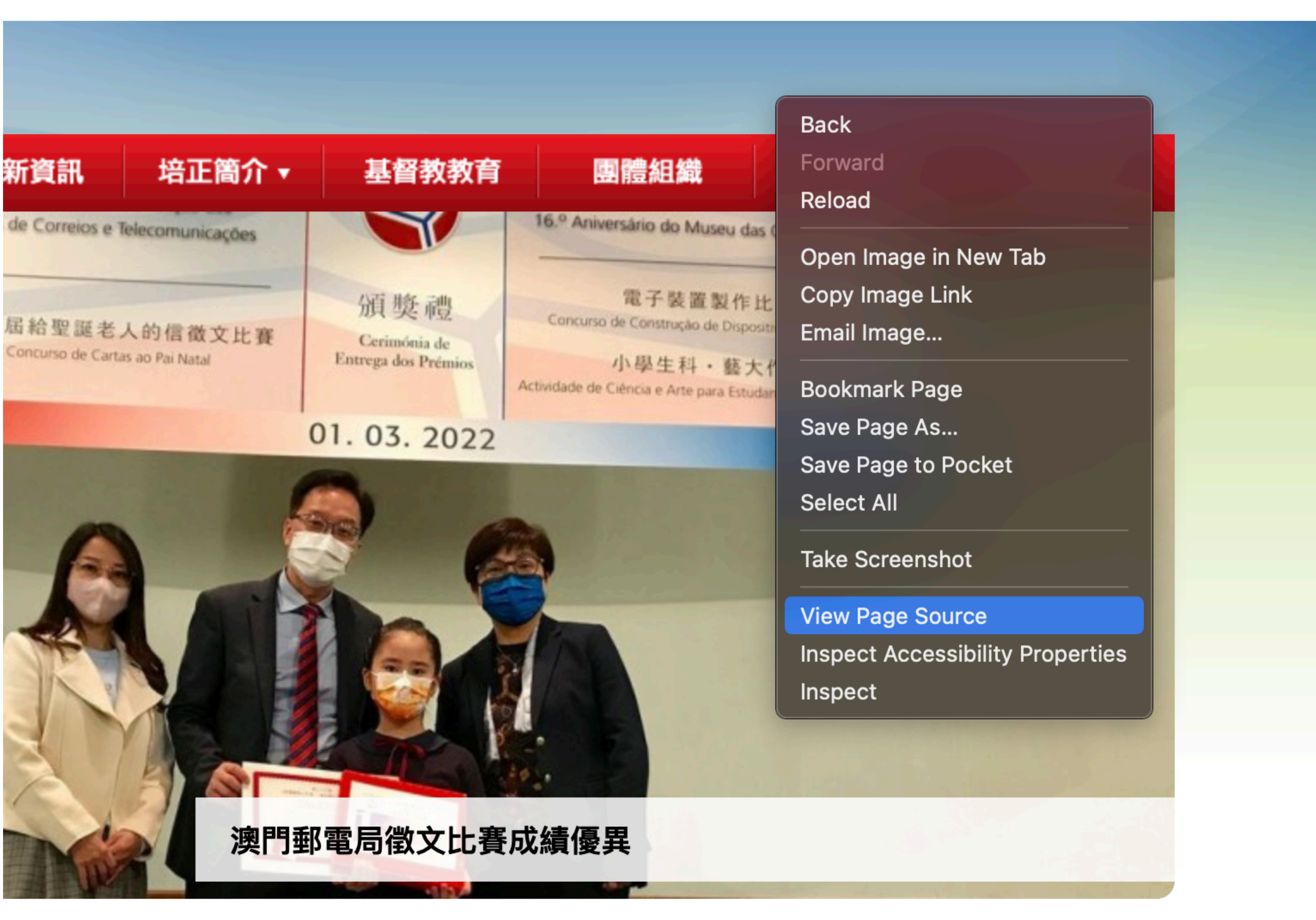
> 取得內容 ("瀏覽"網頁) ➡️ 進行處理 (解析內容)

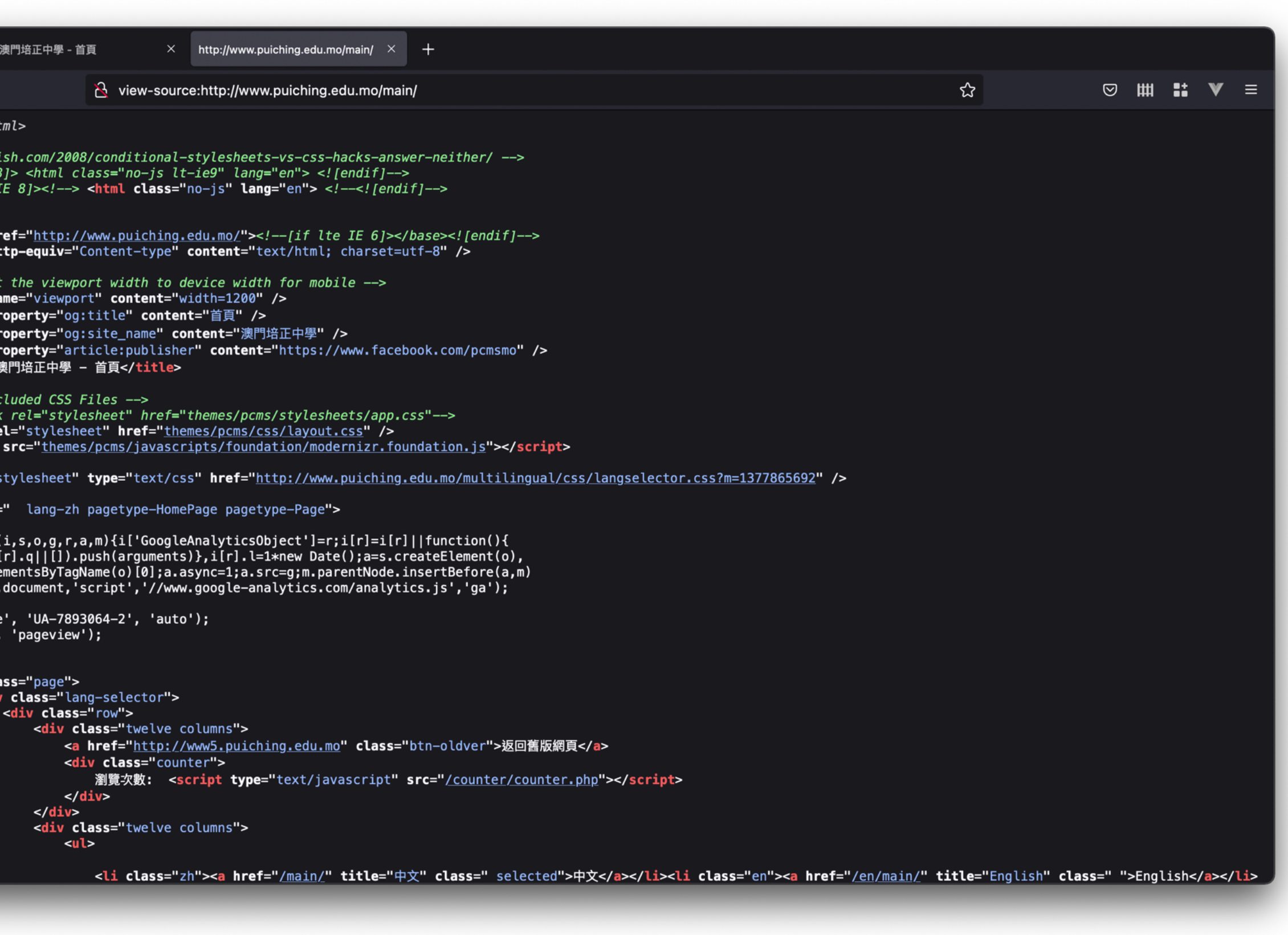
主要可使用兩種技術

1. Requests + Beautiful Soup
   ➡️ 用於靜態產生(HTML)的網頁
2. Selenium + Web Driver
   ➡️ 用於動態產生(JavaScript)的網頁

靜態產生 (Static) vs 動態產生 (Dynamic) 網頁

➡️ 演示





Requests 和 Beautiful Soup 是甚麼

1. Requests ➡️ 取得原始內容
2. Beautiful Soup ➡️ 進一步解析內容

使用 Requests

In [ ]:
```
# 安裝 Requests
```

In [ ]:
```
pip install requests
```

```
DEPRECATION: Configuring installation scheme with distutils config files is deprecated and will no longer work in the near future
. If you are using a Homebrew or Linuxbrew Python, please see discussion at https://github.com/Homebrew/homebrew-core/issues/7662
1
Requirement already satisfied: requests in /usr/local/lib/python3.9/site-packages (2.27.1)
Requirement already satisfied: charset-normalizer~=2.0.0 in /usr/local/lib/python3.9/site-packages (from requests) (2.0.12)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.9/site-packages (from requests) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /usr/local/lib/python3.9/site-packages (from requests) (1.26.9)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.9/site-packages (from requests) (3.3)
WARNING: You are using pip version 21.2.4; however, version 22.0.4 is available.
You should consider upgrading via the '/usr/local/opt/python@3.9/bin/python3.9 -m pip install --upgrade pip' command.
Note: you may need to restart the kernel to use updated packages.
```

使用 requests "瀏覽" 網頁

In [ ]:
```
import requests

requests.get('http://www.puiching.edu.mo') # 注意加 http://
## -> 返回 Response Object
```

Out[ ]:
```
<Response [200]>
```

如何得知成功?

In [ ]:
```
response = requests.get('http://www.puiching.edu.mo/news/')

if response:
    print('Success')
else:
    print('Error')
```

```
Success
```

試試瀏覽一個不存在的網頁

例如 http://www.puiching.edu.mo/news/asdfasdf

In [ ]:
```
# 使用 requests.get 瀏覽網頁，使用上面的 if 做檢查，得出 Success 還是 Error?
pass

# 試試印出 response.status_code
pass
```

網頁的狀態代碼

常見的代碼 200, 301, 302, 404, 500

200 至 <400 - 成功 ✅
400-500或以上 - 失敗 ❌

參考 https://developer.mozilla.org/en-US/docs/Web/HTTP/Status

In [ ]:
```
# Status code 為 301 的演示

response = requests.get('http://google.com')
print(response.history) # response 經過的歷程

# http://www.google.com ➡️ http://www.google.com/intl/zh-TW/ ➡️ http://www.google.com/?hl=zh-TW

print(response.status_code)
print(response.url)

if response:
    print('Success')
else:
    print('Error')
```

```
[<Response [301]>, <Response [302]>, <Response [302]>]
200
http://www.google.com/?hl=zh-TW
Success
```

GET, POST?

GET - 不傳送資料或資料不會改變狀態時使用 (例如搜尋keyword)
POST - 傳送比較大量資料及會(在伺服器)記錄資料時會使用

取得網頁內容

使用 response.text

In [ ]:
```
response = requests.get('http://www.puiching.edu.mo/news/')
response.text
```

Out[ ]:
```
'<!DOCTYPE html>\n\n<!-- paulirish.com/2008/conditional-stylesheets-vs-css-hacks-answer-neither/ -->\n<!--[if IE 8]> <html class=
"no-js lt-ie9" lang="en"> <![endif]-->\n<!--[if gt IE 8]><!--> <html class="no-js" lang="en"> <!--<![endif]-->\n\n<head>\n    <ba
se href="http://www.puiching.edu.mo/"><!--[if lte IE 6]></base><![endif]-->\n    <meta http-equiv="Content-type" content="text/ht
ml; charset=utf-8" />\n\n    <!-- Set the viewport width to device width for mobile -->\n    <meta name="viewport" content="width
=1200" />\n    <meta property="og:title" content="最新消息" />\n    <meta property="og:site_name" content="澳門培正中學" />\n    <me
ta property="article:publisher" content="https://www.facebook.com/pcmsmo" />\n    <title>澳門培正中學 - 最新消息</title>\n\n    <!--
Included CSS Files -->\n    <!--link rel="stylesheet" href="themes/pcms/stylesheets/app.css"-->\n    <link rel="stylesheet" href=
"themes/pcms/css/layout.css" />\n    <script src="themes/pcms/javascripts/foundation/modernizr.foundation.js"></script>\n    \n<l
ink rel="stylesheet" type="text/css" href="http://www.puiching.edu.mo/multilingual/css/langselector.css?m=1377865692" />\n</head>
\n<body class="  lang-zh pagetype-NewsHolder pagetype-InfoHolder pagetype-Page">\n<script>\n  (function(i,s,o,g,r,a,m){i[\'Google
AnalyticsObject\']=r;i[r]=i[r]||function(){\n  (i[r].q=i[r].q||[]).push(arguments)},i[r].l=1*new Date();a=s.createElement(o),\n
m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)\n  })(window,document,\'script\',\'//www.google-a
nalytics.com/analytics.js\',\'ga\');\n\n  ga(\'create\', \'UA-7893064-2\', \'auto\');\n  ga(\'send\', \'pageview\');\n\n</script>
\n    <div class="page">\n        <div class="lang-selector">\n            <div class="row">\n                <div class="twelve
columns">\n                    <a href="http://www5.puiching.edu.mo" class="btn-oldver">返回舊版網頁</a>\n                    <div
class="counter">\n                        瀏覽次數:  <script type="text/javascript" src="/counter/counter.php"></script>\n
</div>\n                </div>\n                <div class="twelve columns">\n                    <ul>\n\n
<li class="zh"><a href="/news/" title="中文" class=" selected">中文</a></li><li class="en"><a href="/en/news/" title="English" cla
ss=" ">English</a></li>\n\n                    </ul>\n                </div>\n            </div>\n        </div>\n        <div cl
ass="page-wrapper  ">\n            <div class="page-top-banner"></div>\n            <div id="page-nav" class="page-nav">\n
<div class="logo">\n                    <a href="main/"><img src="themes/pcms/images/common/pcms-logo.png" /></a>\n
</div>\n                <ul class="fallback">\n                    <li class="nav-home"><a href="main/">首頁</a></li>\n
<li class="nav-news active"><a href="news/">最新資訊</a></li>\n                    \n                    <li class="nav-intro">\n
<a href="intro/" onclick="return false;" class="nolink">培正簡介</a>\n                        <ul class="sub-nav">\n
<li><a href="intro/emblem/">校徽及校歌</a></li>\n                            <li><a href="intro/school-intro/">學校簡介</a></li>\n\n
\n                            <li><a href="intro/school-report/">校務簡報</a></li>\n                            <li><a href="intro
/calendar/">校曆資訊</a></li>\n                            <li><a href="intro/school-map/">校園地圖</a></li>\n
</ul>\n                    </li>\n                    <li class="nav-christianed">\n                        <a href="christian-ed
ucation/" target="_blank">基督教教育</a>\n                    </li>\n                    <li class="nav-organizations"><a href="org
anizations/">團體組織</a></li>\n                    \n                    \n                    <li class="nav-links"><a href="lin
ks/">網址精選</a></li>\n                    <li class="nav-contact"><a href="contact-us/">聯絡我們</a></li>\n                </ul>\n
</div>\n            \n            <div class="page-contents extra-margin">\n                <div class="row">\n
<div class="five columns">\n                        \n                        <div class="sideblock" id="block-livecamera">\n
<h3>校園即時影像</h3>\n                            <div class="block-bottom">\n                                <div class="block-repeat
">\n                                    <div class="block-content">\n                                        <a href="http://www.puiching
.edu.mo/webcam/index.html" class="btn-livecamera">校園即時影像</a>\n                                    </div>\n
</div>\n                            </div>\n                        </div>\n                        <div class="sideblock" id="bl
ock-eservices">\n                            <h3>網上服務</h3>\n                            <div class="block-bottom">\n
<div class="block-repeat">\n                                    <div class="block-content">\n                                        \n
\n                                        <a href="http://admission.puiching.edu.mo" class="btn-sidenav"><div><span>入學及收費資訊</spa
n></div></a>\n                                        \n                                        <a href="https://eclass.puiching.edu.mo/" cla
ss="btn-sidenav"><div><span>eClass</span></div></a>\n                                        \n                                        <a hre
f="http://mo.evi.com.hk/" class="btn-sidenav"><div><span>EVI</span></div></a>\n                                        \n
<a href="http://students.puiching.edu.mo/library/" class="btn-sidenav"><div><span>圖書館</span></div></a>\n
\n                                        <a href="http://home.puiching.edu.mo/~cc" class="btn-sidenav"><div><span>中華文化館</span></d
iv></a>\n                                        \n                                    </div>\n                                </div>\n
</div>\n                        </div>\n                        <div class="sideblock" id="bl
ock-webmail">\n                            <h3>培正電郵系統</h3>\n                            <div class="block-bottom">\n
<div class="block-repeat">\n                                    <form method="post" action="http://webmail.puiching.edu.mo/cgi-bin/op
enwebmail/openwebmail.pl" enctype="multipart/form-data" name="login" target="_blank">\n                                        <div class
="row">\n                                            <div class="seven columns">\n                                                <label class="i
nline">帳號:</label>\n                                            </div>\n                                            <div class="seventeen column
s">\n                                                <input type="text" name="loginname">\n                                            </div>\n
</div>\n                                        <div class="row">\n                                            <div class="seven columns">\n
<label class="inline">密碼:</label>\n                                            </div>\n                                            <div class="s
eventeen columns">\n                                                <input type="password" name="password">\n
</div>\n                                        </div>\n                                        <input type="hidden" name="logindomain" value="we
bmail.puiching.edu.mo">\n                                        <div class="row">\n                                            <div class="twent
yfour columns actions">\n                                                <input type="submit" name="loginbutton" value="登入" class="btn-l
ogin">\n                                            </div>\n                                        </div>\n                                    </for
m>\n                                </div>\n                            </div>\n                        </div>\n                    <
/div>\n                    <div class="fourteen columns">\n                        <div class="centerblock">\n
<div class="block-bottom">\n                                <div class="block-repeat">\n                                    <div class="b
lock-tabs">\n<div class="block-tabs-left">\n    <dl class="block-tabs-bar five-up">\n        <dd class="active"><a href="news/">
全部</a></dd>\n        <dd class=""><a href="/news/category/5">學校動態</a></dd>\n        <dd class=""><a href=
"/news/category/2">學生獎項</a></dd>\n        <dd class=""><a href="/news/category/4">其他資訊</a></dd>\n        \n
<dd class=""><a href="/news/category/3">媒體報導</a></dd>\n        \n    </dl>\n</div>\n</div>\n\n\n<div class="news-item-containe
r">\n    <div class="news-item">\n        <h4 class="news-title"><a href="/news/view/3327">學界乒乓球單打賽獲五冠</a><span class="dat
e">2022-03-20</span></h4>\n        <div class="news-content">\n            \n            \n            <div class="image-containe
r"><img src="/assets/NewsPageImages/_resampled/contentfit-3327-20220316091131-22031605.jpg" alt="3327 20220316091131 22031605" />
</div>\n            \n            <p>第四十五屆學界乒乓球單打賽於3月15日晚上進行了男女子A、B組的決賽和名次爭奪階段，我校同學取得一冠一亞三季的優異
成績。</p>\n<p>男子A組<br/>亞軍：麥天恩，季軍：黃奕瑾，殿軍：張浚軒，第五名：李德彥</p>\n<p>男子B組<br/>季軍：楊文政，第六名：霍展聰</p>\n<p>女子B
組<br/>冠軍：李溈瑋，季軍：陳思澄</p>\n<p>而20日則進行了男、女子C、D、E組的決賽和名次爭奪階段，我校同學取得四冠六亞四季二殿的優異成績。獲獎學生如下：</
p>\n<p>男子C組<br/>亞軍：陳軒賢；殿軍：張兆鋒</p>\n<div class="text_exposed_show">\n<p>女子C組<br/>冠軍：李文彥；亞軍：張祉萱；季軍：譚文菁</p
>\n<p>男子D組<br/>亞軍：關曜星；季軍：林子揚</p>\n<p>女子D組<br/>冠軍：劉梓瑩；亞軍：高珮嘉；季軍：梁凱晴；殿軍：蔡家欣</p>\n<p>男子E組<br/>冠軍：
溫庭昊；亞軍：麥子俊；第六名：梁卓諾</p>\n<p>女子E組<br/>冠軍：羅昫沂；亞軍：張禎熹；季軍：梁樂昕</p>\n</div>\n            </div>\n        <div cla
ss="actions">\n            <a href="/news/view/3327" class="btn-detail">詳細</a>\n        </div>\n    </div>\n</div>\n\n<div class
="news-item-container">\n    <div class="news-item">\n        <h4 class="news-title"><a href="/news/view/3334">全澳中學生衛星概念徵集
比賽共兩隊獲獎</a><span class="date">2022-03-20</span></h4>\n        <div class="news-content">\n            \n            \n
<div class="image-container"><img src="/assets/NewsPageImages/_resampled/contentfit-3334-20220321095939-22032107.jpg" alt="3334 2
0220321095939 22032107" /></div>\n            \n            <p>本校學生參加澳門科技大學月球與行星科學國家重點實驗室主辦的 “澳門星，科學夢：假
如你有一顆衛星，你要做什麼？” 全澳中學生 “澳門科普一號” 衛星概念徵集比賽，經過初賽選拔，進入複賽，最後獲獎名單如下：</p>\n<p>三等獎：高二信 袁中原<br/>
題目：物理的秘密水知道<br/>指導老師：梁沛津</p>\n<p>優異獎：高一信 朱尉妍、梁正朗；初三愛 梁博晞<br/>題目：抗癌新方，獲優異獎<br/>指導老師：陳健鴻</p
>\n        </div>\n        <div class="actions">\n            <a href="/news/view/3334" class="btn-detail">詳細</a>\n        </div
>\n    </div>\n</div>\n\n<div class="news-item-container">\n    <div class="news-item">\n        <h4 class="news-title"><a href="
/news/view/3335">澳大附中同學來訪中華文化館</a><span class="date">2022-03-20</span></h4>\n        <div class="news-content">\n
\n            \n            <div class="image-container"><img src="/assets/NewsPageImages/_resampled/contentfit-3335-202203211002
27-22032108.jpg" alt="3335 20220321100227 22032108" /></div>\n            \n            <p>周六上午，澳大附中歷史小組高中同學二十餘人，在
黎舒琪老師的帶領下，來訪本校中華文化館，我校張柏恩老師分享了一場獨具匠心的「china 瓷」講演，簡介瓷器的歷史和獨特魅力，感受中華之美。高二善馮卓琳，高二愛梁
愫、郭曉嵐、張博旼、麥詠渝，初三望周子瑩幾位同學協助導賞。</p>\n        </div>\n        <div class="actions">\n            <a href="/news/v
iew/3335" class="btn-detail">詳細</a>\n        </div>\n    </div>\n</div>\n\n<div class="news-item-container">\n    <div class="ne
ws-item">\n        <h4 class="news-title"><a href="/news/view/3336">鄺景運校友被按立成為牧師</a><span class="date">2022-03-20</span><
/h4>\n        <div class="news-content">\n            \n            \n            <div class="image-container"><img src="/assets/
NewsPageImages/_resampled/contentfit-3336-20220321101059-22032111.jpg" alt="3336 20220321101059 22032111" /></div>\n            \
n            <div class="kvgmc6g5 cxmmr5t8 oygrvhab hcukyx3x c1et5uql ii04i59q">\n<div dir="auto">本校1994年級學社校友鄺景運，在培正學
習期間相信耶穌，在昨天被浸信宣道會按立成為牧師，擔任頌愛堂主任牧職。鄺牧師在述志中表示：願忠於傳講聖經真理，放眼世界，廣傳福音。高錦輝校長和歐陽效鴻牧師應
邀出席了按牧典禮，送上祝賀。</div>\n<div dir="auto"><br/>另外，曾在本校服務多年的梅保羅牧師，於1987年建立浸信宣道會頌愛堂，今天為該堂紀念成立35周年
的大日子，眾信徒在培正中學禮堂參與感恩崇拜，向神獻上衷心感謝和頌讚，陳敬濂副校長和歐陽效鴻牧師也出席了感恩崇拜，一起向神獻上感謝和讚美！</div>\n</div>\n
</div>\n        <div class="actions">\n            <a href="/news/view/3336" class="btn-detail">詳細</a>\n        </div>\n    </di
v>\n</div>\n\n<div class="news-item-container">\n    <div class="news-item last">\n        <h4 class="news-title"><a href="/news/
view/3331">2021年度培正中學購置資訊及實驗室設備招標開標結果</a><span class="date">2022-03-19</span></h4>\n        <div class="news-conte
nt">\n            \n            \n            \n        </div>\n        <div class="actions">\n            <a href="/news/view/33
31" class="btn-detail">詳細</a>\n        </div>\n    </div>\n</div>\n\n\n\n<!-- Pagination -->\n\n<div class="pagination">\n\n\n\n
\n    <span class="pagination-current">1</span> \n  \n\n  \n    <a href="news/?start=5" title="Go to page 2" class="pagination-li
nk">2</a> \n  \n\n  \n    <a href="news/?start=10" title="Go to page 3" class="pagination-link">3</a> \n  \n\n  \n    <a href="ne
ws/?start=15" title="Go to page 4" class="pagination-link">4</a> \n  \n\n  \n    <a href="news/?start=20" title="Go to page 5" cl
ass="pagination-link">5</a> \n  \n\n<span class="more">...</span>\n\n    <a href="news/?start=5" class="pagination-next">下一頁 &g
t;</a>\n\n</div>\n\n<!-- Pagination end -->\n                                </div>\n                            </div>\n
</div>\n                    </div>\n                    <div class="five columns">\n                        <div class="sideblock
" id="block-typhoon">\n                            <h3>颱風暴雨</h3>\n                            <div class="block-bottom">\n
<div class="block-repeat">\n                                    <div class="block-content">\n                                        <a h
ref="http://www.dsej.gov.mo/~webdsej/typhoon/cinter_typhoon.php?aa=1220" target="_blank"><img src="themes/pcms/images/common/dsej
_typhoon.gif" alt="教育局停課消息" /></a>\n                                        <a href="http://portal.dsej.gov.mo/webdsejspace/inte
rnet/Inter_main_page.jsp?id=21332" target="_blank"><img src="themes/pcms/images/common/dsej_typhoon2.gif" alt="颱風暴雨" /></a>\n
</div>\n                                </div>\n                            </div>\n                        </div>\n
<div class="sideblock" id="block-movieclips">\n                            <h3>影片集</h3>\n                            <div class
="block-bottom">\n                                <div class="block-repeat">\n                                    <div class="block-conte
nt">\n                                        \n                                        <div class="movie-container">\n
<iframe width="168" height="126" src="//www.youtube-nocookie.com/embed/vTVy85Y_-F0?rel=0&showinfo=0&controls=0" frameborder="0" a
llowfullscreen></iframe>\n                                            <div class="caption"><a href="https://youtu.be/vTVy85Y_-F0" tar
get="_blank">共建群體免疫屏障</a></div>\n                                        </div>\n                                        \n
<div>\n                                            <a href="http://www.youtube.com/user/MacauPuiChing" class="btn-youtubechannel" tar
get="_blank">關注本校的 YouTube</a>\n                                        </div>\n                                    </div>\n
</div>\n                            </div>\n                        </div>\n                        <div class="sideblock" id="bl
ock-socialnetwork">\n                            <h3>社交網站</h3>\n                            <div class="block-bottom">\n
<div class="block-repeat">\n                                    <div class="block-content">\n                                        <a href="htt
p://www.facebook.com/pcmsmo" class="btn-facebook" target="_blank">關注本校的 Facebook</a>\n                                        <a href=
"http://www.instagram.com/pcmsmo" class="btn-instagram" target="_blank">關注本校的 Instagram</a>\n                                        <
a href="/rss/" class="btn-rss">特別資訊RSS</a>\n                                    </div>\n                                </div>\n
</div>\n                        </div>\n                    </div>\n                </div>\n            </div>\n        </div>\n
<div class="page-footer  ">\n            <div class="row">\n\n                <div class="five columns mobile-button">\n
\n                </div>\n                <div class="six columns push-four footer-content end">\n                    <img src="t
hemes/pcms/images/common/footer-logo.png" />\n                    <p class="sub-one">\n                        澳門培正中學<br />\n
PUI CHING MIDDLE SCHOOL MACAU\n                    </p>\n                    <p class="sub-two">\n                        澳門高士
德馬路七號<br />\n                        Avenida de Horta e Costa, No.7, Macau<br />\n                        TEL: 28529333 &
nbsp; FAX: 28339056\n                    </p>\n                </div>\n            </div>\n        </div>\n    </div>\n\n
<!-- Included JS Files (Uncompressed) -->\n    <script src="themes/pcms/javascripts/foundation/jquery.js"></script>\n\n    <scrip
t src="themes/pcms/javascripts/foundation/foundation.min.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquer
y.event.swipe.js"></script>\n    <script src="themes/pcms/javascripts/jquery.bxslider/jquery.bxslider.min.js"></script>\n\n    <!
--<script src="themes/pcms/javascripts/foundation/jquery.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquer
y.event.swipe.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.magellan.js"></script>\n    <s
cript src="themes/pcms/javascripts/foundation/jquery.foundation.tooltips.js"></script>\n    <script src="themes/pcms/javascripts/
foundation/jquery.foundation.joyride.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.alerts.
js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.navigation.js"></script>\n    <script src="t
hemes/pcms/javascripts/foundation/jquery.foundation.topbar.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jqu
ery.event.move.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.accordion.js"></script>\n
<script src="themes/pcms/javascripts/foundation/jquery.foundation.forms.js"></script>\n    <script src="themes/pcms/javascripts/f
oundation/jquery.foundation.buttons.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.cookie.js"></script
>\n    <script src="themes/pcms/javascripts/foundation/jquery.placeholder.js"></script>\n    <script src="themes/pcms/javascripts
/foundation/jquery.foundation.reveal.js"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.tabs.js
"></script>\n    <script src="themes/pcms/javascripts/foundation/jquery.foundation.clearing.js"></script>\n    <script src="theme
s/pcms/javascripts/foundation/jquery.foundation.mediaQueryToggle.js"></script>\n    <script src="themes/pcms/javascripts/foundati
on/jquery.foundation.orbit.js"></script>\n    <script src="themes/pcms/javascripts/jquery.bxslider/jquery.bxslider.min.js"></scri
pt>-->\n\n    <!-- Application Javascript, safe to override -->\n    <script src="themes/pcms/javascripts/foundation/app.js"></sc
ript>\n</body>\n</html>\n'
```

如何取得(解析)所需內容?

取得消息標題，因為內容不確定，需要使用 regex.findall 取得內容

HTML Tag 的特性

```
<tag></tag>
<div></div>
<a href="XXX">YYY</a>
<a href="XXX" attr="ZZZ">YYY</a>
<img src="XXX" />
```

編寫 Regex 配對

可以使用 https://regex101.com/ 做實驗

In [ ]:
```
import re

response = requests.get('http://www.puiching.edu.mo/news/')
content = response.text

matches = re.findall('<a href="([^"]*)">([^>]*)</a>', content)
```

收窄範圍

- 使用 re.findall 找出所有 XXX
- 使用 for 檢查每個 match, 使用 startswith 檢查內容是否為 /news 開頭

In [ ]:
```
# re.findall
matches = re.findall('<a href="([^"]*)">([^>]*)</a>', content)

for each_match in matches:
    pass # your code here
```

結論

- 使用 requests.get 取得 Response
- 使用 .status_code 可以檢查狀態
- 使用 .text 可以返回的文字內容
- 使用 regex 略有不足，而且較難掌握。可考慮另找方法對資料做處理，例如 Beautiful Soup。