**Python程式設計**

**範圍： 網路爬蟲**

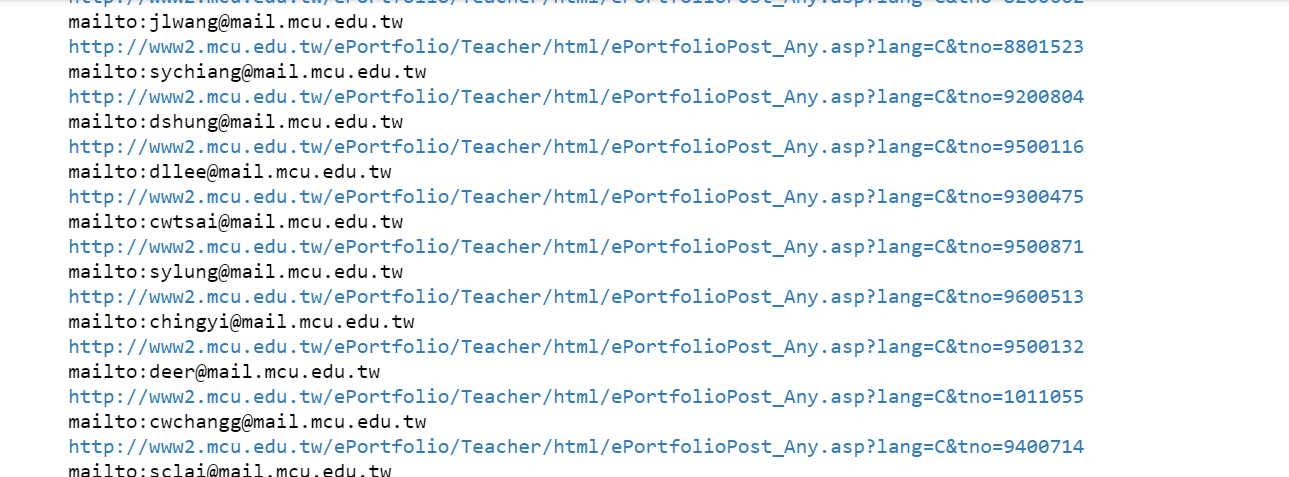
**銘傳大學電腦與通訊工程系**

|  |  |
| --- | --- |
| 班 級 | 電通四乙 |
| 姓 名 | 陳昱叡 |
| 學 號 | 04052474 |
| 成 績 | 應繳作業共 5 題，每題20分，滿分為100分  共完成 5 題，應得 100 分 |
| 授課教師 | 陳慶逸 |

※請確實填寫自己寫完成題數，並且計算得分。填寫不實者(如上傳與作業明顯無關的答案，或是計算題數有誤者)，本次作業先扣50分。

EX 1: 讀取銘傳電通系**師資介紹**網頁中所有的 <a> 超連結，並顯示所有的 href 屬性內容。

期望輸出：



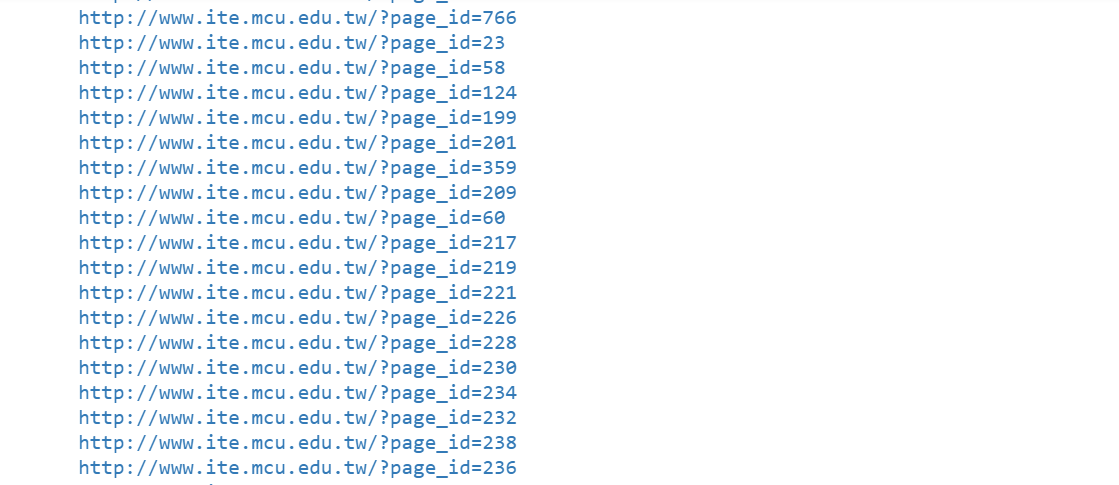
程式碼：

|  |
| --- |
| **from bs4 import BeautifulSoup**  **import requests**  **url = 'http://www.ite.mcu.edu.tw/?page\_id=76'**  **html = requests.get(url)**  **html.encoding="utf-8"**  **sp=BeautifulSoup(html.text,"html.parser")**  **links=sp.find\_all("a") # 讀取 <a>**  **for link in links:**  **print(link.get("href")) # 讀取 href 屬性內容** |



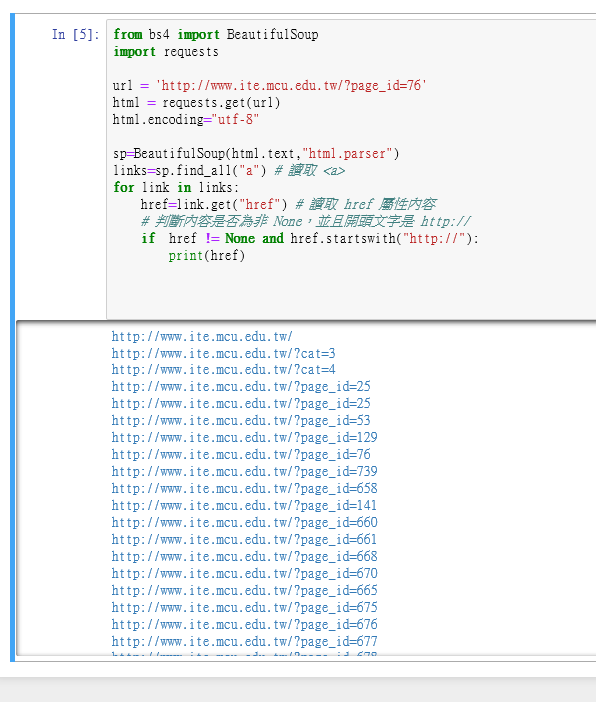
EX 2: 讀取銘傳電通系**師資介紹**網頁中所有的 <a> 超連結，並顯示所有的 href 屬性內容中以http://開頭的連結。

期望輸出：



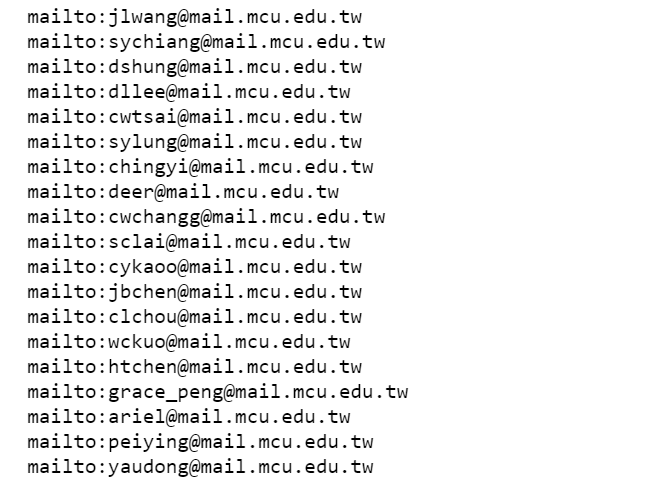
程式碼：

|  |
| --- |
| **from bs4 import BeautifulSoup**  **import requests**  **url = 'http://www.ite.mcu.edu.tw/?page\_id=76'**  **html = requests.get(url)**  **html.encoding="utf-8"**  **sp=BeautifulSoup(html.text,"html.parser")**  **links=sp.find\_all("a") # 讀取 <a>**  **for link in links:**  **href=link.get("href") # 讀取 href 屬性內容**  **# 判斷內容是否為非 None，並且開頭文字是 http://**  **if href != None and href.startswith("http://"):**  **print(href)** |

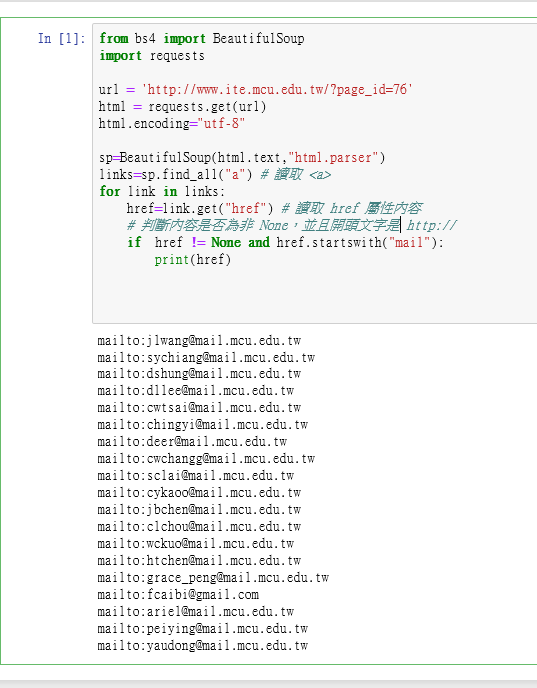


EX 3: 讀取銘傳電通系**師資介紹**網頁中所有的 <a> 超連結，並顯示所有的 href 屬性內容中所有的email。

期望輸出：

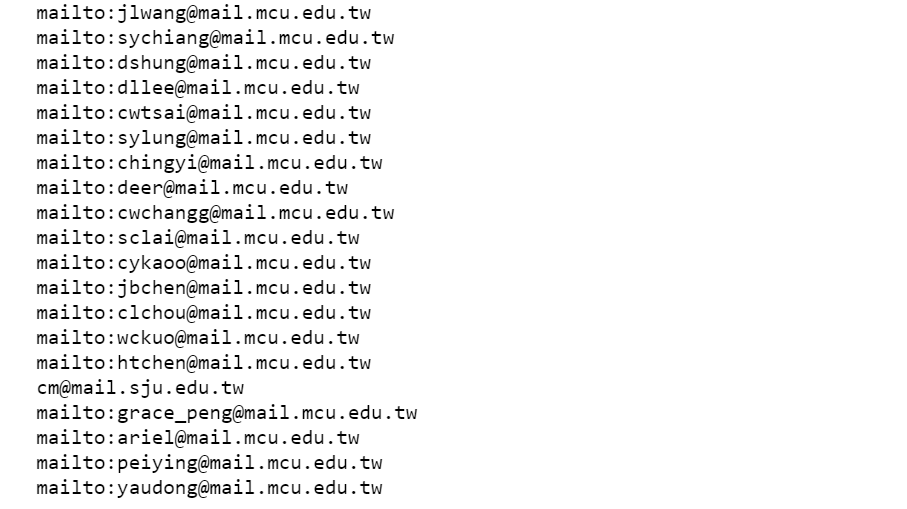


|  |
| --- |
| **from bs4 import BeautifulSoup**  **import requests**  **url = 'http://www.ite.mcu.edu.tw/?page\_id=76'**  **html = requests.get(url)**  **html.encoding="utf-8"**  **sp=BeautifulSoup(html.text,"html.parser")**  **links=sp.find\_all("a") # 讀取 <a>**  **for link in links:**  **href=link.get("href") # 讀取 href 屬性內容**  **# 判斷內容是否為非 None，並且開頭文字是 http://**  **if href != None and href.startswith("mail"):**  **print(href)** |

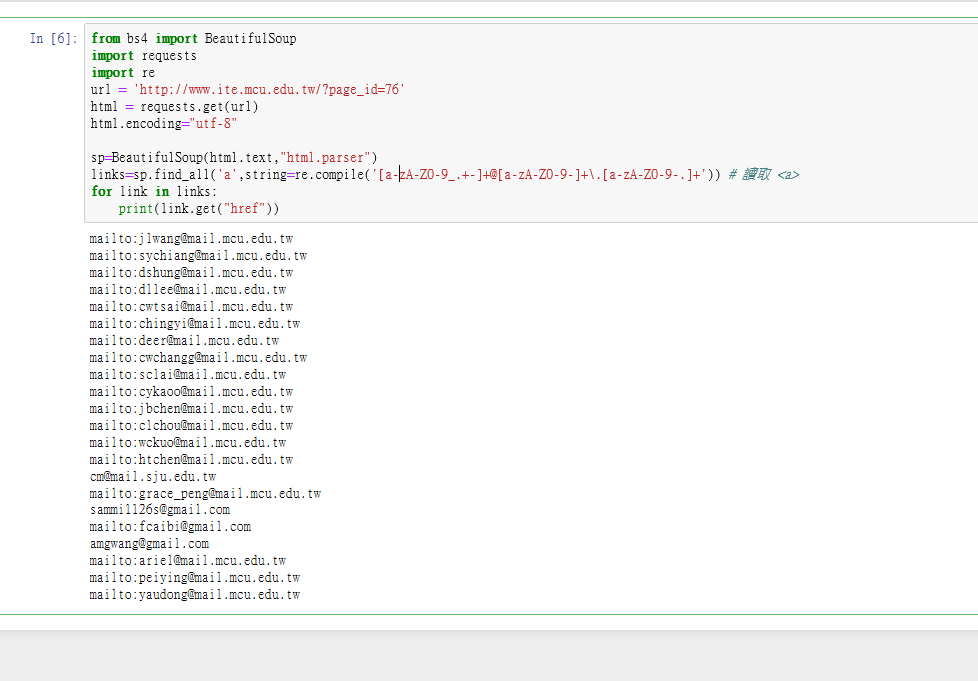


EX 4: **試以“正規表示法搭配文字內容”來取得電通系師資介紹網頁中所有的e-mail:**

期望輸出：

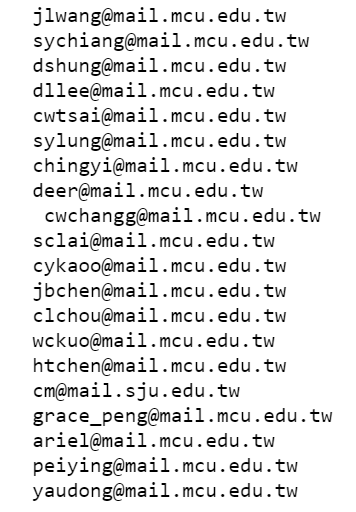


|  |
| --- |
| **from bs4 import BeautifulSoup**  **import requests**  **import re**  **url = 'http://www.ite.mcu.edu.tw/?page\_id=76'**  **html = requests.get(url)**  **html.encoding="utf-8"**  **sp=BeautifulSoup(html.text,"html.parser")**  **links=sp.find\_all('a',string=re.compile('[a-zA-Z0-9\_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+')) # 讀取 <a>**  **for link in links:**  **print(link.get("href"))** |



EX 5: **試以“正規表示法搭配文字內容”來取得電通系師資介紹網頁中所有的e-mail，並去掉“mailto”:**

期望輸出：



|  |
| --- |
| **from bs4 import BeautifulSoup**  **import requests**  **import re**  **url = 'http://www.ite.mcu.edu.tw/?page\_id=76'**  **html = requests.get(url)**  **html.encoding="utf-8"**  **sp=BeautifulSoup(html.text,"html.parser")**  **links=sp.find\_all('a',string=re.compile('[a-zA-Z0-9\_.+-]+@[a-zA-Z0-9-]+\.[a-zA-Z0-9-.]+')) # 讀取 <a>**  **for link in links:**  **print(link.string)** |

