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：

```
mailto:jlwang@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=8801523
mailto:sychiang@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9200804
mailto:dshung@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9500116
mailto:dllee@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9300475
mailto:cwtsai@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9500871
mailto:sylung@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9600513
mailto:chingyi@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9500132
mailto:deer@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=1011055
mailto:cwchangg@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9400714
mailto:sclai@mail.mcu.edu.tw
```

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
```

In [4]:

```
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
```

```
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=8200662
mailto:jlwang@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=8801523
mailto:sychiang@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9200804
mailto:dshung@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9500116
mailto:dllee@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9300475
mailto:cwtsai@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9600513
mailto:chingyi@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9500132
mailto:deer@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=1011055
mailto:cwchangg@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9400714
mailto:sclai@mail.mcu.edu.tw
http://www2.mcu.edu.tw/ePortfolio/Teacher/html/ePortfolioPost_Any.asp?lang=C&tno=9600521
mailto:cykaoo@mail.mcu.edu.tw
```

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：

```
http://www.ite.mcu.edu.tw/?page_id=23
http://www.ite.mcu.edu.tw/?page_id=58
http://www.ite.mcu.edu.tw/?page_id=124
http://www.ite.mcu.edu.tw/?page_id=199
http://www.ite.mcu.edu.tw/?page_id=201
http://www.ite.mcu.edu.tw/?page_id=359
http://www.ite.mcu.edu.tw/?page_id=209
http://www.ite.mcu.edu.tw/?page_id=60
http://www.ite.mcu.edu.tw/?page_id=217
http://www.ite.mcu.edu.tw/?page_id=219
http://www.ite.mcu.edu.tw/?page_id=221
http://www.ite.mcu.edu.tw/?page_id=226
http://www.ite.mcu.edu.tw/?page_id=228
http://www.ite.mcu.edu.tw/?page_id=230
http://www.ite.mcu.edu.tw/?page_id=234
http://www.ite.mcu.edu.tw/?page_id=232
http://www.ite.mcu.edu.tw/?page_id=238
http://www.ite.mcu.edu.tw/?page_id=236
```

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)
```

In [5]:

```
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)
```

```
http://www.ite.mcu.edu.tw/
http://www.ite.mcu.edu.tw/?cat=3
http://www.ite.mcu.edu.tw/?cat=4
http://www.ite.mcu.edu.tw/?page_id=25
http://www.ite.mcu.edu.tw/?page_id=25
http://www.ite.mcu.edu.tw/?page_id=53
http://www.ite.mcu.edu.tw/?page_id=129
http://www.ite.mcu.edu.tw/?page_id=76
http://www.ite.mcu.edu.tw/?page_id=739
http://www.ite.mcu.edu.tw/?page_id=658
http://www.ite.mcu.edu.tw/?page_id=141
http://www.ite.mcu.edu.tw/?page_id=660
http://www.ite.mcu.edu.tw/?page_id=661
http://www.ite.mcu.edu.tw/?page_id=668
http://www.ite.mcu.edu.tw/?page_id=670
http://www.ite.mcu.edu.tw/?page_id=665
http://www.ite.mcu.edu.tw/?page_id=675
http://www.ite.mcu.edu.tw/?page_id=676
http://www.ite.mcu.edu.tw/?page_id=677
```

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：

```
mailto:jlwang@mail.mcu.edu.tw
mailto:sychiang@mail.mcu.edu.tw
mailto:dshung@mail.mcu.edu.tw
mailto:dllee@mail.mcu.edu.tw
mailto:cwtsai@mail.mcu.edu.tw
mailto:sylung@mail.mcu.edu.tw
mailto:chingyi@mail.mcu.edu.tw
mailto:deer@mail.mcu.edu.tw
mailto:cwchangg@mail.mcu.edu.tw
mailto:sclai@mail.mcu.edu.tw
mailto:cykaoo@mail.mcu.edu.tw
mailto:jbchen@mail.mcu.edu.tw
mailto:clchou@mail.mcu.edu.tw
mailto:wckuo@mail.mcu.edu.tw
mailto:htchen@mail.mcu.edu.tw
mailto:grace_peng@mail.mcu.edu.tw
mailto:ariel@mail.mcu.edu.tw
mailto:peiying@mail.mcu.edu.tw
mailto:yaudong@mail.mcu.edu.tw
```

```
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)
```

In [1]:

```
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)
```

```
mailto:jlwang@mail.mcu.edu.tw
mailto:sychiang@mail.mcu.edu.tw
mailto:dshung@mail.mcu.edu.tw
mailto:dllee@mail.mcu.edu.tw
mailto:cwtsai@mail.mcu.edu.tw
mailto:chingyi@mail.mcu.edu.tw
mailto:deer@mail.mcu.edu.tw
mailto:cwchangg@mail.mcu.edu.tw
mailto:sclai@mail.mcu.edu.tw
mailto:cykaoo@mail.mcu.edu.tw
mailto:jbchen@mail.mcu.edu.tw
mailto:clchou@mail.mcu.edu.tw
mailto:wckuo@mail.mcu.edu.tw
mailto:htchen@mail.mcu.edu.tw
mailto:grace_peng@mail.mcu.edu.tw
mailto:fcaibi@gmail.com
mailto:ariel@mail.mcu.edu.tw
mailto:peiying@mail.mcu.edu.tw
mailto:yaudong@mail.mcu.edu.tw
```

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：

```
mailto:jlwang@mail.mcu.edu.tw
mailto:sychiang@mail.mcu.edu.tw
mailto:dshung@mail.mcu.edu.tw
mailto:dllee@mail.mcu.edu.tw
mailto:cwtsai@mail.mcu.edu.tw
mailto:sylung@mail.mcu.edu.tw
mailto:chingyi@mail.mcu.edu.tw
mailto:deer@mail.mcu.edu.tw
mailto:cwchangg@mail.mcu.edu.tw
mailto:sclai@mail.mcu.edu.tw
mailto:cykaoo@mail.mcu.edu.tw
mailto:jbchen@mail.mcu.edu.tw
mailto:clchou@mail.mcu.edu.tw
mailto:wckuo@mail.mcu.edu.tw
mailto:htchen@mail.mcu.edu.tw
cm@mail.sju.edu.tw
mailto:grace_peng@mail.mcu.edu.tw
mailto:ariel@mail.mcu.edu.tw
mailto:peiying@mail.mcu.edu.tw
mailto:yaudong@mail.mcu.edu.tw
```

```
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"))
```

In [6]:
```
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"))
```

```
mailto:jlwang@mail.mcu.edu.tw
mailto:sychiang@mail.mcu.edu.tw
mailto:dshung@mail.mcu.edu.tw
mailto:dllee@mail.mcu.edu.tw
mailto:cwtsai@mail.mcu.edu.tw
mailto:chingyi@mail.mcu.edu.tw
mailto:deer@mail.mcu.edu.tw
mailto:cwchangg@mail.mcu.edu.tw
mailto:sclai@mail.mcu.edu.tw
mailto:cykaoo@mail.mcu.edu.tw
mailto:jbchen@mail.mcu.edu.tw
mailto:clchou@mail.mcu.edu.tw
mailto:wckuo@mail.mcu.edu.tw
mailto:htchen@mail.mcu.edu.tw
cm@mail.sju.edu.tw
mailto:grace_peng@mail.mcu.edu.tw
sammi1126s@gmail.com
mailto:fcaibi@gmail.com
amgwang@gmail.com
mailto:ariel@mail.mcu.edu.tw
mailto:peiying@mail.mcu.edu.tw
mailto:yaudong@mail.mcu.edu.tw
```

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：

```
jlwang@mail.mcu.edu.tw
sychiang@mail.mcu.edu.tw
dshung@mail.mcu.edu.tw
dllee@mail.mcu.edu.tw
cwtsai@mail.mcu.edu.tw
sylung@mail.mcu.edu.tw
chingyi@mail.mcu.edu.tw
deer@mail.mcu.edu.tw
 cwchangg@mail.mcu.edu.tw
sclai@mail.mcu.edu.tw
cykaoo@mail.mcu.edu.tw
jbchen@mail.mcu.edu.tw
clchou@mail.mcu.edu.tw
wckuo@mail.mcu.edu.tw
htchen@mail.mcu.edu.tw
cm@mail.sju.edu.tw
grace_peng@mail.mcu.edu.tw
ariel@mail.mcu.edu.tw
peiying@mail.mcu.edu.tw
yaudong@mail.mcu.edu.tw
```

```
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)
```

In [8]:

```
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)
```

```
jlwang@mail.mcu.edu.tw
sychiang@mail.mcu.edu.tw
dshung@mail.mcu.edu.tw
dllee@mail.mcu.edu.tw
cwtsai@mail.mcu.edu.tw
chingyi@mail.mcu.edu.tw
deer@mail.mcu.edu.tw
 cwchangg@mail.mcu.edu.tw
sclai@mail.mcu.edu.tw
cykaoo@mail.mcu.edu.tw
jbchen@mail.mcu.edu.tw
clchou@mail.mcu.edu.tw
wckuo@mail.mcu.edu.tw
htchen@mail.mcu.edu.tw
cm@mail.sju.edu.tw
grace_peng@mail.mcu.edu.tw
sammi1126s@gmail.com
fcaibi@gmail.com
amgwang@gmail.com
ariel@mail.mcu.edu.tw
peiying@mail.mcu.edu.tw
yaudong@mail.mcu.edu.tw
```