# Psychoinformatics - Week 5 (Exercises)

by boyonglin (r10945002@ntu.edu.tw)

## 進一步搜尋 Boy-Girl 版資訊 (8 points)

1. index.html 右上角 [< 上頁 ] 中包含了總頁數資訊，請用 LXML 抓出此經常變動的數字。(2 points)

In [ ]:
```
!pip install scrapy
!pip install bs4
```

In [ ]:
```
from bs4 import BeautifulSoup as BS
import requests
import re

def get_pages(URI):
    r = requests.get(URI)
    s = BS(r.text, 'html.parser')
    for link in s.find_all('a'):
        if link.string == '‹ 上頁':
            href = link.get('href')
            pages = int(re.search(r'\d+', href).group())
    return s, pages

s, pages = get_pages("https://www.ptt.cc/bbs/Boy-Girl")

print('Total pages in Boy-Girl:', pages + 1)
```

```
Total pages in Boy-Girl: 6191
```

2. 請用 LXML 找出距離現在時間最近的一篇[爆]文標題與 URN (有可能需要翻頁) 。(3 points)

In [ ]:
```
URI = "https://www.ptt.cc/bbs/Boy-Girl"
s, page = get_pages(URI)

target_found = False

while page > 0:
    for element in s.find_all(class_='nrec', string='爆'):
        title_element = element.find_next_sibling(class_='title')
        page_string = f'(/index{page}.html)'
        print(element.string, title_element.text.strip(), page_string)
        target_found = True

    if target_found:
        break
    else:
        page -= 1

    URI = "https://www.ptt.cc/bbs/Boy-Girl" + '/index' + str(page) + '.html'
```

```
    r = requests.get(URI)
    s = BS(r.text)
```

爆 [討論] 被說玩奈良的行程在亂排 (/index6185.html)

3. 請用 Selenium 在 index.html 往前翻三頁，每頁拍一張照片，在notebook內顯示。(3 points)

In [ ]:
```
!pip install selenium
```

In [ ]:
```
from selenium import webdriver
from selenium.webdriver.edge.service import Service
from selenium.webdriver.common.by import By

URI='https://www.ptt.cc/bbs/Boy-Girl/'
driver=webdriver.Edge(service=Service("./msedgedriver.exe"))
driver.get(URI)

for i in range(3):
    btn=driver.find_element(By.XPATH, "//*[text()='‹ 上頁']")
    btn.click()
    driver.save_screenshot(f'index-{i + 1}.png')

driver.close()

# print(driver.page_source)
```

In [ ]:
```
from IPython.display import Image, display

image1 = Image(filename="./index-1.png")
image2 = Image(filename="./index-2.png")
image3 = Image(filename="./index-3.png")

display(image1)
display(image2)
display(image3)
```


批踢踢實業坊 › 看板 Boy-Girl
聯絡資訊 關於我們
看板 精華區
最舊 ‹ 上頁 下頁 › 最新
搜尋文章…
3 (本文已被刪除) [ilv1181023]
- 10/08
1 Re: [討論] 台灣其實很需要外籍配偶吧？
hayuyang 10/08 …
3 [討論] 跟曖昧對象
peterasd 10/08 …
4 Re: [討論] 請問，看這個版對脫單有幫助嗎！
chirex 10/08 …
2 Re: [討論] 為什麼男生比較不敢對抗社會壓力？
breathair 10/08 …
32 Re: [討論] 為什麼男生比較不敢對抗社會壓力？
dreamingyou 10/08 …
3 Re: [討論] 台灣其實很需要外籍配偶吧？
dreamingyou 10/08 …
4 Re: [討論] 為什麼男生比較不敢對抗社會壓力？


批踢踢實業坊 › 看板 Boy-Girl
聯絡資訊 關於我們
看板 精華區
最舊 ‹ 上頁 下頁 › 最新
搜尋文章…
13 [討論] 朋友發生這種情況算是被PUA嗎？
Kakehiko 10/07 …
39 [討論] 台灣其實很需要外籍配偶吧？
PPAPwww 10/07 …
Re: [討論] 請問，看這個版對脫單有幫助嗎！
takizawa5566 10/07 …
7 Re: [討論] 我的條件嚴苛？還是男生太魯蛇？
darkdick 10/07 …
2 Re: [討論] 台灣其實很需要外籍配偶吧？
season2011 10/07 …
26 [分享] 整形就對了 CP值超高 爽快
Fucker5566 10/07 …
6 Re: [閒聊] 台男認識印尼女網友變成老婆的故事
b122771 10/07 …
74 [討論] 這是特例，還是常態


Please submit your Jupyter Notebook in PDF before next Friday (10/13/2023).