利用網路爬蟲
改善法務的工作效率

By Tinsley Wu

目 錄

動 機

之前在事務所擔任律師助理，
其中一項職務內容：
幫律師搜集並整理好
與案件背景相似度高的判決，
以便律師判斷案件勝訴率，
通常要花一個多小時
在網路上用關鍵字慢慢搜尋，
這是沒有效率的方法，
所以想利用網路爬蟲改善。

方 法

利用Python及Selenium
自動搜集資料，
並經由Pandas加以整理，
最後將資料存放至My SQL，
以便後續方便應用。

驗 證

Step 1

用網路爬蟲搜尋關鍵字，自動點擊換頁，並儲存原始資料。

Step 2

用Pandas過濾與清理原始資料。

Step 3

將最終資料存放至My SQL。

Github Code : https://github.com/TinsleyWu/Enhance-efficiency-for-paralegal

In [1]:
```
from selenium import webdriver
from selenium.webdriver.common.by import By
import time
import pymysql
import pandas as pd
```

In [2]:
```
driver = webdriver.Chrome()
driver.set_window_size(1300, 1080)
driver.get('https://judgment.judicial.gov.tw/FJUD/default.aspx')
driver.find_elements(By.XPATH, '//input[@id="txtKW"]')[0].send_keys('強盜殺人 強制性交')
time.sleep(2)
driver.find_elements(By.XPATH, '//input[@id="btnSimpleQry"]')[0].click()
res_df = pd.DataFrame({})
driver.switch_to.default_content() # change to main page
driver.switch_to.frame('iframe-data')

for i in range(int(driver.find_elements(By.XPATH,
                                        '//div[@id="plPager"]')[0].text.split('/ ')[1].split(' ')[0])):
    html = driver.page_source
    df = pd.read_html(html)
    df = df[0]
    list_summary = []
    for j in range(len(df['裁判字號 （內容大小）']) // 2):
        list_summary.append(df['裁判字號 （內容大小）'][2*j + 1])
        df = df.drop(2*j + 1, axis = 0)
    df['摘要'] = list_summary
    df = df.drop('序號', axis = 1)
    res_df = pd.concat([res_df, df])
    if (int(driver.find_elements(By.XPATH,
                                 '//div[@id="plPager"]')[0].text.split('/ ')[1].split(' ')[0]) -1) != i:
        driver.find_elements(By.XPATH, '//a[@id="hlNext"]')[0].click()
    time.sleep(0.2)

res_df = res_df.reset_index(drop=True)
```

In [18]:

```
# 資料庫設定
db_settings = {
    "host": "127.0.0.1",
    "port": 3306,
    "user": "root",
    "password": "",
    "db": "Crawler_Data",
    "charset": "utf8"
}
try:
    # 建立Connection物件
    conn = pymysql.connect(**db_settings)
    # 建立Cursor物件
    with conn.cursor() as cursor:
    #資料表相關操作
        # 新增資料SQL語法
        command = "INSERT INTO resumes(det_name, det_date, reason, summary)VALUES(%s, %s, %s, %s)"
        for i in range(res_df.shape[0]):
            cursor.execute(
                command, (res_df['裁判字號 （內容大小）'][i], res_df['裁判日期'][i], res_df['裁判案由'][i],
                          res_df['摘要'][i]))
        # 儲存變更
        conn.commit()
except Exception as ex:
    print(ex)
```

結 論

以上述270筆資料為例，
過去用手動搜尋關鍵字方法
要花1個多小時，
而網路爬蟲卻只需要20秒
即可搜集並整理好，
大幅提高工作效率。

0911-602-992

tinsleywu1204@gmail.com

THANKS FOR WATCHING