網站地圖基礎架構

壹、 主程式

甲、 基礎邏輯

程式會產生兩個 txt 檔案，分別為「links.txt」、「links_all.txt」，links.txt 存放網站地圖的網址，links_all.txt 存放基礎網址+衍生網址。將基礎網址分離是為了往後比對新增資料時可以更容易發現。

```
if __name__ == "__main__":
    # 每週開始抓前先刪除文件
    file_name = 'links.txt'
    if os.path.exists(file_name):
        os.remove(file_name)

    file_all = 'links_all.txt'
    if os.path.exists(file_all):
        os.remove(file_all)

    find_href()
    find_sidebar_links()
    database()
```

貳、 find_href()

甲、 基礎邏輯

抓取網站地圖(https://www.twse.com.tw/zh/sitemap.html)網址。

將結果寫入 links.txt

篩選條件為 非

1.TIB

2.CSR

3./downloads/

4.文件結尾

5.sitemap

若非完整網址，則將正確的前綴補上(https://www.twse.com.tw)。

參、 find_sidebar_links()

甲、 基礎邏輯

打開 links.txt，利用基礎網址進一步找出標籤<navigation>，<main>中的網址。其中會呼叫 find_main_links()，分離尋找<main>的程式碼。

肆、 find_main_links()

甲、 基礎邏輯

遞迴蒐羅所有網址，返回網址陣列到 find_sidebar_links()，再進行判斷。最後將 sitemap，<navigation>、<main>中的網址寫入 links_all.txt

伍、 Database()

甲、 基礎邏輯

讀 links_all.txt 開始抓取 map_route。

```
if count > 10 : # 每 10 筆網址隨機休息
                count =0
                random_number = random.randint(1,30)
                time.sleep(random_number)
```

每爬取 10 個網頁，隨機休息(1~30s)。

```
if  FORM_FG:
```

判斷有沒有下拉選項，是則特殊處理，否則直接爬取。

```
            if  FORM_FG:
                soup_list = date_range_select(url) # 返回各種選項產生的
soup_list
                for soup in soup_list:
                    form_fg(soup) # 取 dict
                    table_fg(soup)
                    table_attachment_fg(soup)
                    csv_and_html_button_fg(soup)
                    find_crumbs_yet = find_Crumbs(soup,url)
                    if find_crumbs_yet == Exception:
                        queue.append(url)

            else:
                form_fg(soup) # 取 dict
                table_fg(soup)
                table_attachment_fg(soup)
                csv_and_html_button_fg(soup)
                find_Crumbs(soup,url)
                if find_crumbs_yet == Exception:
                    queue.append(url)
            data_list.append(data_dict)
```

這邊生成各種資料需要的資料。

陸、 form_fg(soup)

甲、 基礎邏輯

判斷頁面有沒有表單，有返回 True，沒有返回 False，並更改字典。

柒、 table_fg(soup)

甲、 基礎邏輯

判斷頁面有沒有 table，有返回 True，沒有返回 False，並更改字典。

捌、 csv_and_html_button_fg(soup)

甲、 基礎邏輯

判斷頁面有沒有載點，有返回 True，沒有返回 False，並更改字典。

玖、 table_attachment_fg(soup)

甲、 基礎邏輯

判斷 table 有沒有載點，有返回 True，沒有返回 False，並更改字典。

壹拾、 remove_last_segment_and_append(soup)

甲、 基礎邏輯

把相對連結處理成完整連結，存入資料庫才會是可用的連結。

壹拾壹、 date_range_select(soup)

甲、 基礎邏輯

因為有 form 的頁面網址不會改變，但會在同一個頁面會產出許多不同的頁面，所以需要額外一個視窗來點擊處理。

```
        try:
            popup = driver.find_element(By.CLASS_NAME,"ok")
            popup.click()
        except:
            pass
```

近期有彈跳式視窗，若有則點即 OK，使原始頁面露出方便抓取。

```
  if number_yy == 1:
        elif number_mm == 2:
        elif number_dd == 3:
```

以 year 下拉式選單出現的次數來判斷頁面點選方式。

==1 表示需要迭代每個日期

==2 表示日期是一個區間

==3 是唯一一個特殊網址

```
            try:
                WebDriverWait(driver, 1).until(
                    EC.visibility_of_element_located((By.XPATH,'//*[@nam
e="yy"]'))
                )
              exist_yy = True
          except: exist_yy = False
```

```
        try:
            WebDriverWait(driver, 1).until(
                EC.visibility_of_element_located((By.XPATH,'//*[@name="mm"]'))
            )
            exist_mm = True
        except: exist_mm = False
        try:
            WebDriverWait(driver, 1).until(
                EC.visibility_of_element_located((By.XPATH,'//*[@name="dd"]'))
            )
            exist_dd = True
        except: exist_dd = False
```

用等待頁面是否顯示元素的方式來檢視網站是否以 CSS 方式隱藏存在 html 中的元素。由此判斷網站允許使用者使用的元素為何。