### 投稿類別：

### 2023校務研究議題競賽&交流工作坊暨媒合會

篇名：

蒐集有關中央大學的新聞並加以整理分類

作者：

中央大學資工系二年級　林詮宗

中央大學資工系二年級　林楷傑

中央大學資工系二年級　侯詠倫

隊伍名稱：

中央新聞挖挖哇

議題合作行政單位：

秘書室

指導老師：

張嘉惠老師

**壹、前言**

**一，摘要**

在這篇小論文，我們首先提出了研究的動機，指出在信息時代的今天，新聞報導的即時性和多樣性日益成為社會不可或缺的一部分，同時也面臨著龐大的資訊量、信息散亂和假新聞的散播等挑戰。為了因應這些挑戰，我們提出了以中央大學新聞蒐集的自動化為焦點的研究，旨在探討其在節省人力與時間、以及處理假新聞方面的潛在優勢。

在前言部分，我們清晰地分析自動化新聞蒐集的優勢，包括提高效率、節省人力成本以及對抗假新聞的能力。隨後，研究方法一節詳細描述了兩種主要的研究方法，分別是課堂學習和網路查詢。通過參與相關課程、與教授對話、文獻回顧和案例分析，隊伍獲得了關於中央大學新聞蒐集的深入了解。

接著，正文部分介紹了研究的執行方式，包括瀏覽器初始化和網頁搜尋、新聞標題和內容定位、Excel資料彙整、啟動爬蟲和資料儲存等步驟。透過Python程式碼和Selenium技術，我們成功實現了對多個新聞網站的中央大學相關資訊的自動化蒐集，並以結構化的方式存儲在Excel檔案中。

在困難與挑戰一節，我們分享了在執行過程中遇到的一些問題，並提供了解決方案，展現了解決問題的能力。最後，簡介部分提到了計畫成果和程式碼。

最後在結論與展望的部分，我們提出了未來的展望和改進方向，包括界面優化、加入更多新聞網站、與中央大學網站結合以及技術優化等建議，希望以後能有多元的發展與實踐。

**二、研究動機**

隨著信息時代的快速發展，新聞報導的即時性和多樣性日益成為我們社會中不可或缺的一部分。然而，傳統的人工新聞蒐集方式面臨著諸多挑戰，包括龐大的資訊量、信息散亂以及假新聞的散播。因此，本研究將聚焦於中央大學新聞蒐集的自動化，以探討其在節省人力與時間、以及處理假新聞方面的潛在優勢。

首先，自動化新聞蒐集的引入被視為提高效率的一大利器。傳統人工搜尋常常耗費大量時間和人力，而自動化系統能夠在瞬息之間分析龐大的新聞資訊，不僅提高了蒐集速度，同時也有效減少了蒐集過程中的人力成本。這種高效率的蒐集方式有助於確保新聞報導的即時性，讓使用者能夠更及時地獲取到最新的資訊。

其次，自動化蒐集系統的應用可以極大地改進對抗假新聞的能力。隨著社交媒體等平台的崛起，假新聞的散播速度變得極為迅猛，這不僅影響了公眾的信息判斷，也對社會穩定產生了潛在威脅。自動化系統能夠快速辨識新聞中的可疑特徵，實現對假新聞的早期發現和處理。透過機器學習技術，系統能夠不斷學習和優化，提升辨識的準確性，進而加強整體新聞資訊的可信度。

因此，本研究旨在深入探討中央大學新聞蒐集自動化的可行性及其對人力與時間的節省效益，同時評估其在處理假新聞方面的應用潛力。透過本研究，我們期望能為新聞業務提供更有效的工具，同時促進對抗假新聞的技術發展，以確保學生與教師職員們能在本校新聞網上，獲取到更為可靠的新聞資訊。

**三、研究方法**

本研究主要採用兩種主要方法進行，即課堂學習和網路查詢，以深入了解中央大學新聞蒐集的自動化應用，並探討其對節省人力與時間和處理假新聞的實際效果。

參與相關課程： 透過參與與新聞、自動化技術相關的課程，深入學習自動化 蒐集技術的基本原理和應用實例。

與教授對話： 與相關領域的教授進行深入對話，瞭解自動化技術在新聞領域的最新發展，並取得對於中央大學新聞自動化的見解和建議。

文獻回顧： 進行文獻回顧，探討自動化新聞蒐集的先前研究成果，特別是在節省人力與時間和處理假新聞方面的應用。搜尋相關的學術論文、期刊文章和業界報告，以建立研究的理論基礎。

案例分析： 查詢中央大學相關的新聞蒐集案例，了解目前是否已有自動化技術應用在該校新聞蒐集中，並蒐集案例資料作為研究的實際參考。

**貳、正文**

**一、執行方式**

瀏覽器初始化和網頁搜尋：

在本次研究中，我們使用了Python程式碼，利用Selenium的動態爬蟲技術以及利用Chrome WebDriver初始化google瀏覽器，搜尋至中央大學的新聞網頁，為確保網頁完全載入，我們使用了time.sleep(5)的函數，使程式能在該頁面等待5秒。

新聞標題和內容定位：

利用Selenium的定位方法，我們找到了新聞頁面中的標題和內容元素。這可以使用CSS選擇器或其他Selenium提供的元素定位方式，具體視網頁結構而定。

Excel資料彙整：

我們使用了pandas套件的python套件，初始化一個Excel工作簿，創建一個工作表，並將爬取到的新聞標題和內容資訊寫入Excel表格中。

啟動爬蟲：

透過執行整個爬蟲程式碼，我們啟動了自動化的新聞蒐集程序，模擬瀏覽器行為。

資料儲存：

新聞標題和內容被彙整並儲存到Excel檔案中，這提供了一個結構化的方式來保存和管理收集到的資訊

**二、蒐集對象**

我們的新聞蒐集對象包括多個知名的新聞網站，以確保涵蓋多元且廣泛的新聞資訊。以下是新聞蒐集對象的詳細描述：

Google 新聞：

新聞範疇或類別： 包含全球、國內、財經、科技、娛樂等多個領域的新聞。

聯合報新聞網：

新聞範疇或類別： 提供全面的國際、政治、社會、生活等新聞報導。

中央社：

新聞範疇或類別： 提供即時的國內外新聞、財經、文教、科技等報導。

自由時報：

新聞範疇或類別： 涵蓋政治、社會、國際、娛樂、體育等多元領域的新聞。

中時新聞資料庫：

新聞範疇或類別： 包括國內外即時新聞、政治、財經、生活、文化等豐富內容。

透過選擇上述新聞網站來蒐集中央大學的新聞有諸多益處，舉凡能以更多元的視角，了解中央大學在國際舞台上的影響，再者，這些新聞網的即時性能確保獲得最新的相關資訊，總的來說，這些不同來源的新聞網站能夠提供更全面、即時、多元且可靠的中央大學新聞資訊，確保蒐集到的資訊具有多方位的視野。

**三、困難與挑戰**

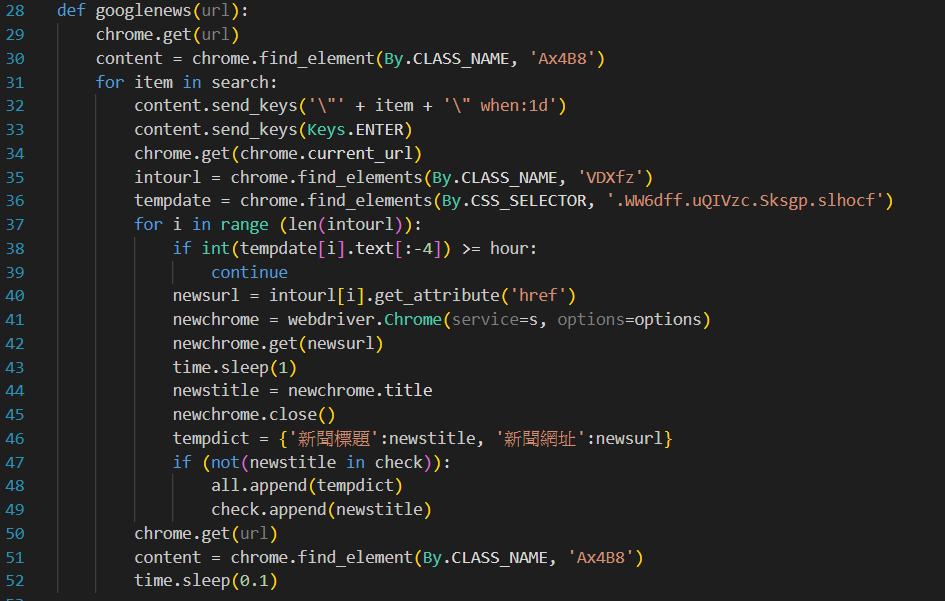
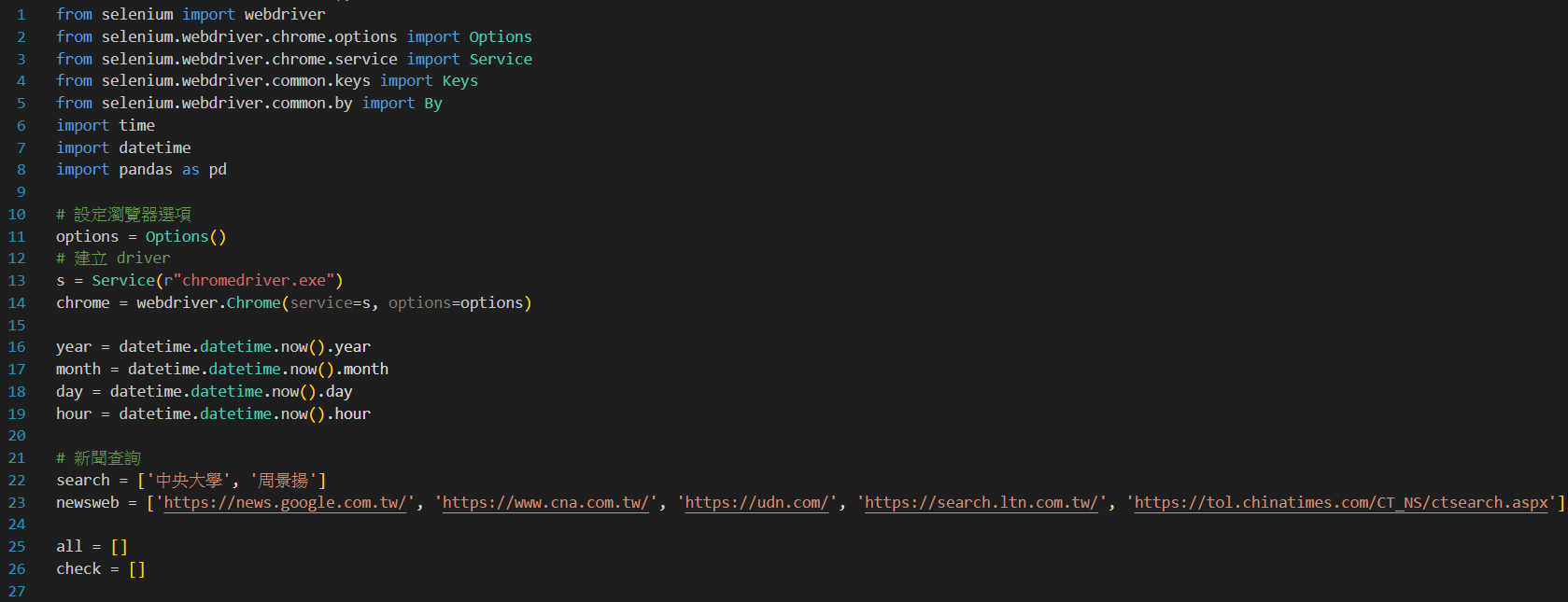
1. 擷取新聞的標題時，因為是從html裡擷取的，因此標題會有不完整的情形/開新的webdriver從裡面擷取title標籤下的文字
2. 在Google news中用的是一天內的新聞來查詢，但會有昨天的新聞出現/擷取html中的時間，並跟當天日期進行比較即可
3. 會有重複的資料出現，使得excel裡的東西太多/將當天爬到的資料進行判斷，若一樣則不存人當中
4. 新聞網站的搜尋欄是JavaScript操控的，不知如何處理/將搜尋後的url複製起來，去看哪裡不同，再去進行修改即可

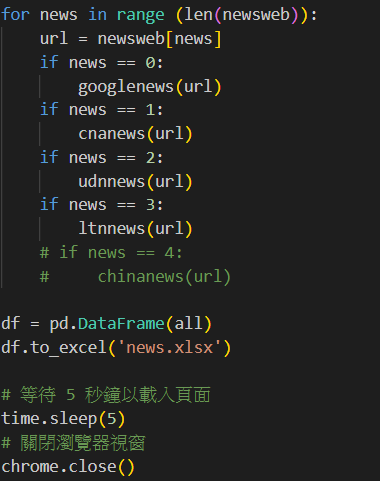
**四、計畫成果與程式碼**

excel檔：



程式碼(部分)：





**參、結論與展望**

透過蒐集多個新聞網站的中央大學相關資訊，我們成功建立了一個具有潛力的新聞資訊收集系統。這種多角度的資訊來源確保了我們能夠從不同的視角瞭解中央大學的最新動態、學術成就以及校園生活。同時，自動化的資訊蒐集方式大大提升了效率，節省了人力成本，達成合作的需求。

為了進一步優化資訊蒐集與呈現，我們提出以下展望與改進方向：

**界面優化：**

增加使用者友好的界面，提升系統的易用性，使使用者能夠更輕鬆地瀏覽和尋找所需的新聞資訊。

加入圖片、多媒體等元素，提升視覺體驗。

**增加新聞網：**

持續擴展新聞網站的涵蓋範圍，加入更多具有影響力的新聞平台，以確保更全面的新聞報導。

考慮加入特定領域的新聞網站，以滿足不同使用者對於特定主題的需求。

**與中央大學網站結合：**

整合中央大學官方網站的新聞動態，確保我們蒐集到的資訊更為全面且即時。

建立與中央大學網站的互動機制，使使用者能夠更方便地存取學校的相關資訊。

**技術優化：**

探索新的技術，提升爬蟲系統的穩定性和效率。

使用機器學習技術進行資訊分析，提高對於重要新聞的辨識和呈現。

這些建議將進一步提升我們的新聞資訊蒐集系統，使其更符合使用者需求，同時更有效地呈現中央大學的最新動態。

通過技術的不斷升級和優化，我們期待此專題之後能有進一步的發展，在校務研究議題上，有多方面的運用。