

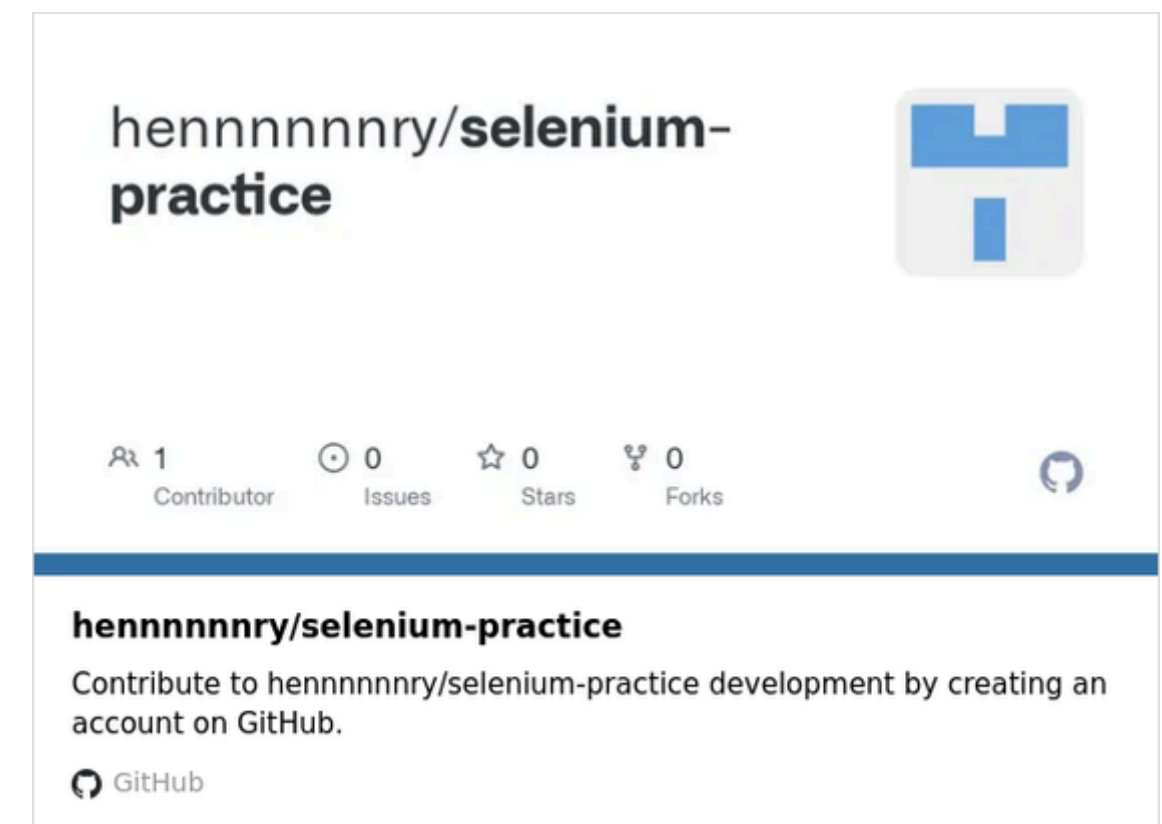
# Selenium 與資料庫

---

簡報者：陳弘蒼

# 目錄

- 構想概述.....3
- 步驟概述.....4-8



GitHub詳細程式碼

# 構想 概述

- 主要功能

更快速的瀏覽新聞是我原先想解決的痛點，主要步驟為先透過selenium爬取新聞網站的文章後，將文章匯入資料庫中，再串接openai對文章內容做摘要，節省閱讀時間。

但在串接openai的過程遇到了一些問題，API成功接上了，而因版本的更新有許多地方需要調適，在國外的論壇上有人遇到同樣問題並建議可以降版本使用。

因此在解決問題前，此處僅先呈現selenium和資料庫的使用。



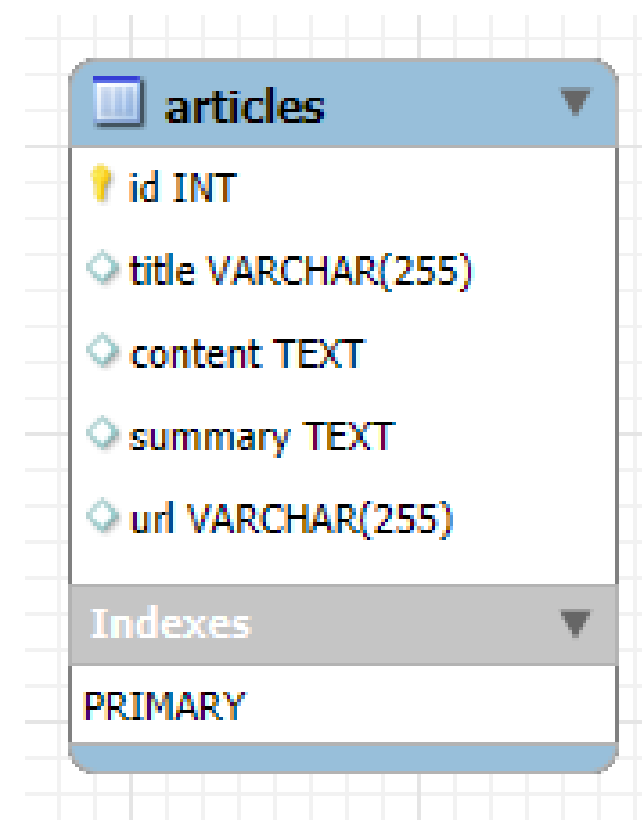
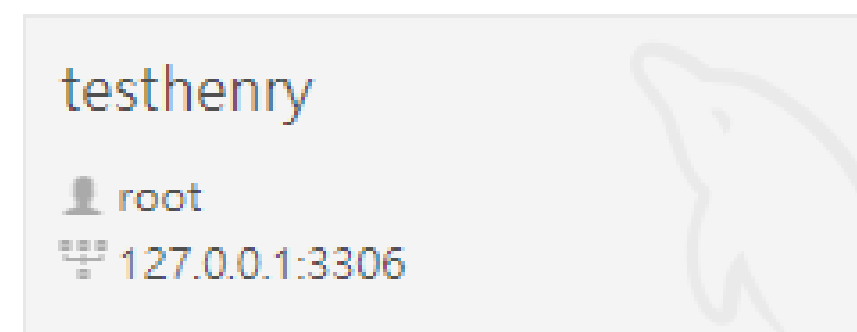
# 步驟概述

- MySQL連接設置

```
# MySQL 資料庫連接設置
db_connection = mysql.connector.connect(
    host="127.0.0.1",
    port="3306",
    user="root",
    password="",
    database="news_db"
)

cursor = db_connection.cursor()

# 建立 articles 資料表（如果尚未存在）
cursor.execute("""
CREATE TABLE IF NOT EXISTS articles (
    id INT AUTO_INCREMENT PRIMARY KEY,
    title VARCHAR(255),
    content TEXT,
    url VARCHAR(255)
)
""")
```



先與資料庫做連接，這裡使用workbench做資料庫的編輯，方便查詢、更新等操作。

# 步驟概述

- 資料蒐集

```
service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service)
```

設置chromedriver

```
driver.get('https://technews.tw/')
time.sleep(3)
```

開啟有興趣的網站

```
▼ <div class="img">
  ▼ <a href="https://finance.technews.tw/2024/09/16/samsung-begins-active-corporate-restructuring/" onclick="tnClickEvent('Album_08_A1', {'post_id':'1281765'})"> event
    Samsung 三星
  </a>
▼ <li class="spotlist">
  <a href="https://technews.tw/2024/09/17/tsmc-2nm-chips-for-iphone-17/">
    iPhone 17 Pro 系列將採 2 奈米製程晶片，台積電加快試產但有挑戰</a>
  </li>
▼ <h1 class="entry-title">
  <a href="https://technews.tw/2024/09/17/huawei-trifold-phone-replace-the-screen/" title="華為三摺機小心別摔螢幕！維修更換費用可買一台 iPhone 16 Pro" rel="bookmark">
    華為三摺機小心別摔螢幕！維修更換費用可買一台 iPhone 16 Pro</a>
  </h1>
```

使用開發者工具找尋需要的資訊

```
articles = driver.find_elements(
    By.XPATH,
    '//*[@class="img"]/a | //*[@class="spotlist"]/a | //*[@class="entry-title"]/a'
)
```

將所需的標題和連結一次爬取

- 套件使用

**selenium**：匯入webdriver自動操作及網頁爬蟲使用。

**mysql.connector**：將獲取的資料匯入資料庫。

**logging**：記錄日誌訊息方便偵錯。

# 步驟概述

- 儲存文章標題和連結

```
seen_hrefs = set()
article_data = []

for article in articles:
    href = article.get_attribute('href')
    if href not in seen_hrefs:
        seen_hrefs.add(href)
        article_data.append((article.text, href))
```

防止存取重複連結

- 參數調適

```
for article_title, article_link in article_data:
    driver.get(article_link)
    time.sleep(3)

    try:
        content = WebDriverWait(driver, 10).until(
            EC.presence_of_element_located((By.XPATH, '//*[@class="indent"]'))
        )
        print(f"Title: {article_title}")
        print(f"Content: {content.text}")
```

進去存取的連結抓取文章內容

# 步驟 概述

- 儲存文章到資料庫

```
# 清除上一次查詢的結果
cursor.reset()
# 檢查資料庫中是否已經存在具有相同標題的文章
cursor.execute("SELECT id FROM articles WHERE title = %s", (article_title,))
result = cursor.fetchone()

if result is None:
    # 將文章標題、內容和 URL 存入資料庫
    cursor.execute("""
        INSERT INTO articles (title, content, url)
        VALUES (%s, %s, %s)
    """, (article_title, content.text, article_link))
    db_connection.commit()
else:
    print(f"Article with title '{article_title}' already exists in the database.")
```







若不存在此標題，存入資料庫

# 步驟概述

## ● 確認資料庫

1 • `SELECT * FROM articles;`

2

Result Grid					
Filter Rows: <input type="text"/>					
Edit:   					
Export/Import:  					
Wrap Cell Content: 					
	id	title	content	summary	url
▶	1	一手好牌變爛牌！陸行之：Wolf speed 公司...	日前，碳化矽晶圓供應商 Wolf speed 於美股...	NULL	https://finance.technews.tw/2024/08/23/wolfs...
	2	靠科技翻身，東歐小國亞美尼亞經濟崛起	世界上有一個被遺忘的國家，正在利用科技...	NULL	https://technews.tw/2024/08/23/it-developme...
	3	三星：2027 年 2 奈米導入晶背供電，晶片...	三星電子 (Samsung Electronics Co.) 晶圓代...	NULL	https://technews.tw/2024/08/23/samsung-elec...
	4	傳三星今年投片首批 HBM4 設備，2025 年提...	三星今年稍晚時推出首款 HBM4 記憶體元件...	NULL	https://technews.tw/2024/08/23/samsung-hbm...
	5	預防美國再緊縮出口管制，中國前七個月半...	彭博社引述中國海關總署的數據表示，中國...	NULL	https://technews.tw/2024/08/23/chinas-semico...
	6	世上最快電子顯微鏡首亮相，能拍攝 1 秒繞...	一群物理學家開發出世上拍攝速度最快的電...	NULL	https://technews.tw/2024/08/22/microscope-el...
	7	中美晶旗下台特化第三季底掛牌，上半年 E...	矽晶圓大廠中美晶旗下小金雞台特化舉行上...	NULL	https://finance.technews.tw/2024/08/20/global...
	8	鈺祥台南柳營廠開幕，為全球首座先進製程...	半導體先進製程供應鏈廠商鈺祥今日台南柳...	NULL	https://technews.tw/2024/08/20/yesiang-new-...
	9	英特爾延長處理器保固要擴展到主機板，夥...	英特爾針對先前第 13、14 代 Core-i 桌上型...	NULL	https://technews.tw/2024/08/23/intel-extends-...
	10	德媒質疑高額補貼歐積電，蕭茲：德國需要...	台積電德勒斯登晶圓廠動土典禮獲高度關注...	NULL	https://technews.tw/2024/08/23/olaf-scholz-es...
	11	輝達財報在即，Hopper 供給吃緊狀況恐是變數	輝達 (Nvidia Corp.) 即將於 8 月 28 日公布...	NULL	https://finance.technews.tw/2024/08/23/as-nv...
	12	需求復甦慢！DRAM 價格漲勢暫歇，今後看...	PC、智慧手機需求復甦緩慢，DRAM 價格漲...	NULL	https://technews.tw/2024/08/23/dram-price-ris...
	13	印度月船三號探測器新數據佐證，月球曾被...	去年 8 月，印度月船三號探測器成功於月球...	NULL	https://technews.tw/2024/08/23/pragyan-luna...
	14	倒楣恆星定期供饋物質，科學家成功計算黑...	一對恆星在路過黑洞後，其中一顆被完全吞...	NULL	https://technews.tw/2024/08/21/at2018fyk-st...
	15	火星隕石雨！研究：幾乎每天都有籃球大小...	少了大氣層保護的星球會發生什麼事？就像...	NULL	https://technews.tw/2024/08/20/mars-space-r...

資料成功以標題、文章內容、URL 形式儲存進資料庫



**THE END**