

Gemini Chat

幫我寫一個 python 爬蟲程式，根據以下的網址，擷取每一部電影的資訊，包含電影名稱、電影圖片 URL、評分（如有）、類型（如有），並儲存成 movie.csv

```
https://ssr1.scrape.center/page/1  
https://ssr1.scrape.center/page/2  
https://ssr1.scrape.center/page/3  
https://ssr1.scrape.center/page/4  
https://ssr1.scrape.center/page/5  
https://ssr1.scrape.center/page/6  
https://ssr1.scrape.center/page/7  
https://ssr1.scrape.center/page/8  
https://ssr1.scrape.center/page/9  
https://ssr1.scrape.center/page/10
```

◆ 這是一個完整的 Python 爬蟲程式，使用了 `requests` 抓取網頁，`BeautifulSoup` 解析 HTML，並使用 `pandas` 將資料整理儲存為 CSV 檔案。

程式碼說明

1. **自動翻頁**：程式會自動跑完 1 到 10 頁。
2. **資料擷取**：針對每一部電影抓取標題、圖片連結、評分與類型。
3. **防護機制**：加入了 `time.sleep(1)`，避免請求過快對伺服器造成負擔。
4. **檔案儲存**：最後將資料存為 `movie.csv`，並使用 `utf-8-sig` 編碼，確保用 Excel 開啟時中文不會變成亂碼。

1. 安裝必要套件

在執行程式前，請確保已安裝以下套件：

Bash

```
pip install requests beautifulsoup4 pandas
```

2. Python 爬蟲程式碼 (`scrape_movies.py`)

Python

```
import requests  
from bs4 import BeautifulSoup  
import pandas as pd  
import time  
  
def scrape_movies():  
    # 建立一個列表來儲存所有爬取到的電影資料  
    movie_data = []  
  
    # 基礎 URL 結構  
    base_url = "https://ssr1.scrape.center/page/{}"
```

```

# 迴圈遍歷第 1 頁到第 10 頁
for page in range(1, 11):
    url = base_url.format(page)
    print(f"正在爬取第 {page} 頁: {url}")

try:
    # 發送 GET 請求
    response = requests.get(url)

    # 檢查請求是否成功 (狀態碼 200)
    if response.status_code == 200:
        soup = BeautifulSoup(response.text, 'html.parser')

        # 找到該頁面上所有的電影卡片 (根據網站結構，通常在 el-card class 中)
        items = soup.find_all('div', class_='el-card')

        for item in items:
            # 1. 摘取電影名稱 (在 h2 標籤內)
            title_tag = item.find('h2')
            title = title_tag.text.strip() if title_tag else "未知名稱"

            # 2. 摘取圖片 URL (在 img 標籤的 src 屬性)
            img_tag = item.find('img', class_='cover')
            img_url = img_tag['src'] if img_tag else "無圖片"

            # 3. 摘取評分 (在 p 標籤 class='score' 內)
            score_tag = item.find('p', class_='score')
            score = score_tag.text.strip() if score_tag else "無評分"

            # 4. 摘取類型 (在 class='categories' 內的 button 標籤)
            categories_div = item.find('div', class_='categories')
            if categories_div:
                # 找到所有按鈕並取得文字，然後用逗號合併
                buttons = categories_div.find_all('button')
                categories = ", ".join([btn.text.strip() for btn in buttons])
            else:
                categories = "無分類"

            # 將這部電影的資訊加入列表
            movie_data.append({
                '電影名稱': title,
                '評分': score,
                '類型': categories,
                '圖片 URL': img_url
            })
        else:
            print(f"無法讀取第 {page} 頁，狀態碼: {response.status_code}")

except Exception as e:
    print(f"爬取第 {page} 頁時發生錯誤: {e}")

# 禮貌性延遲 1 秒，避免對伺服器造成過大負擔
time.sleep(1)

# 將資料轉換為 DataFrame
df = pd.DataFrame(movie_data)

# 儲存成 csv 檔案
# encoding='utf-8-sig' 是為了讓 Excel 開啟時中文能正常顯示
df.to_csv('movie.csv', index=False, encoding='utf-8-sig')

print("-" * 30)
print(f"爬取完成！共抓取 {len(df)} 部電影。")
print("檔案已儲存為 movie.csv")

```

```
if __name__ == "__main__":
    scrape_movies()
```

執行結果預覽

程式執行完畢後，你會在資料夾中看到一個 `movie.csv`，內容格式如下：

電影名稱	評分	類型	圖片 URL
霸王別姬 - Farewell My Concubine	9.6	劇情, 愛情	https://p0.meituan.net/...
這個殺手不太冷 - Léon	9.5	劇情, 動作, 犯罪	https://p1.meituan.net/...
...

如果有任何執行上的問題（例如網頁結構更新導致抓不到資料），請隨時告訴我！