

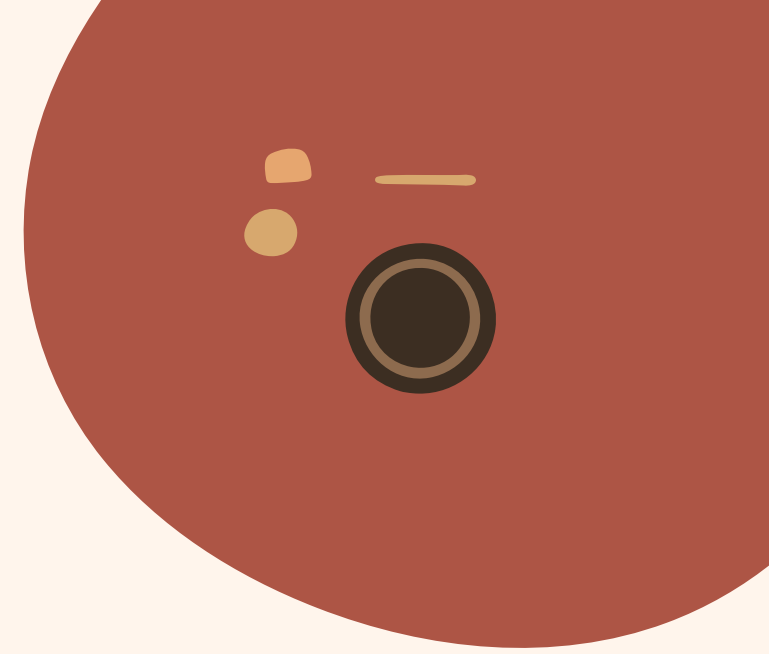


Web Crawler

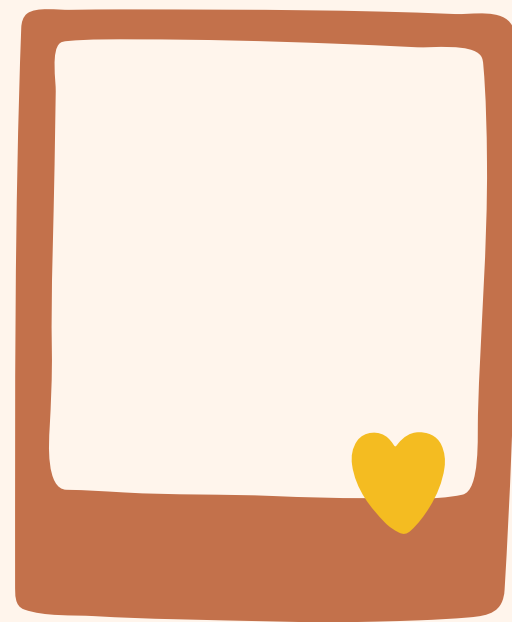


START

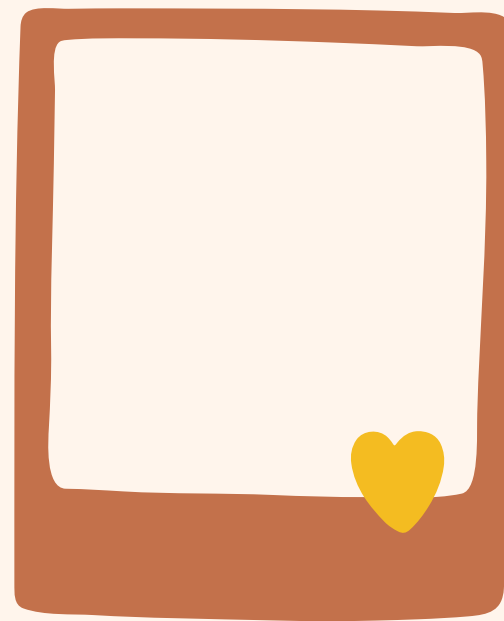




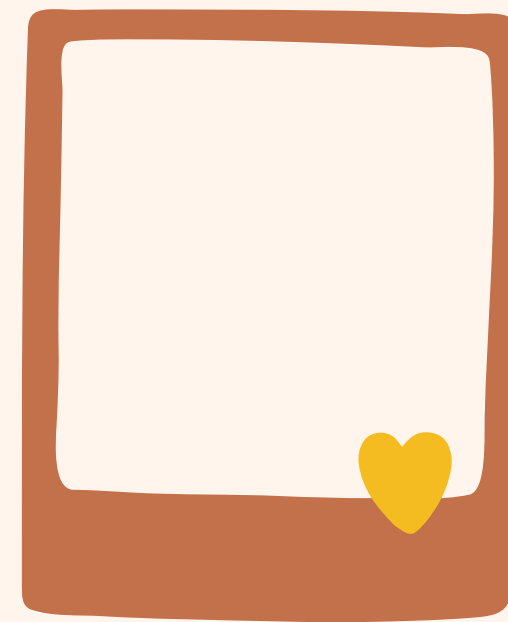
Package



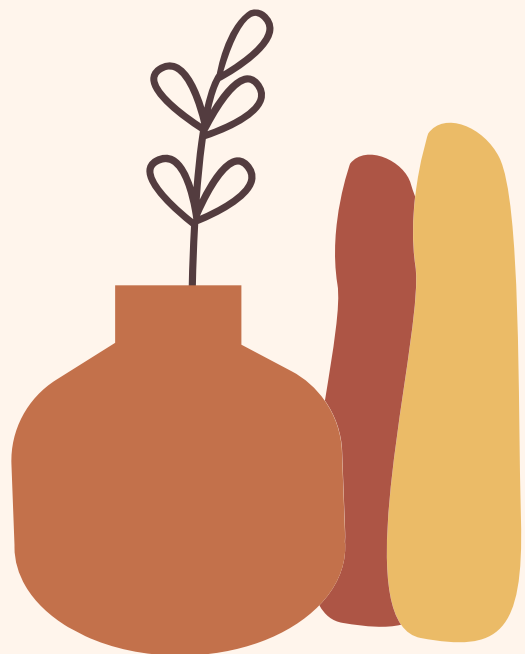
requests

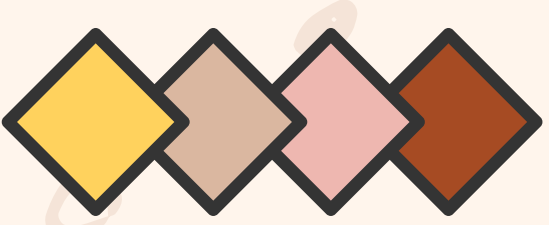


selenium



json



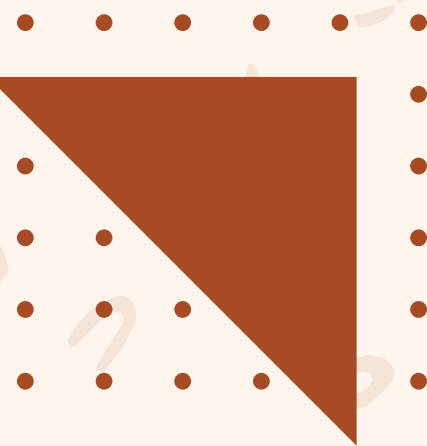


Steps

Topic 1: 爬取相關資訊

Topic 2: 載入圖片

Topic 3: 建立對應檔案的 json






Topic 1

資料篩選

位置 **按鄉鎮**  按捷運  按學校  按商圈 

 清空條件

 保存搜尋

不限 ☒ 中山區 ☒ 大安區 ☐ 信義區 ☐ 士林區 ☐ 內湖區 ☒ 中正區 ☒ 萬華區 ☐ 松山區 ☒ 大同區 ☐ 文山區 ☐ 北投區 ☐ 南港區

類型 **不限** 整層住家 獨立套房 分租套房 雅房 車位 其他

租金 **不限** ☐ 5000元以下 ☐ 5000-10000元 ☐ 10000-20000元 ☐ 20000-30000元 ☐ 30000-40000元 ☐ 40000元以上 - 元

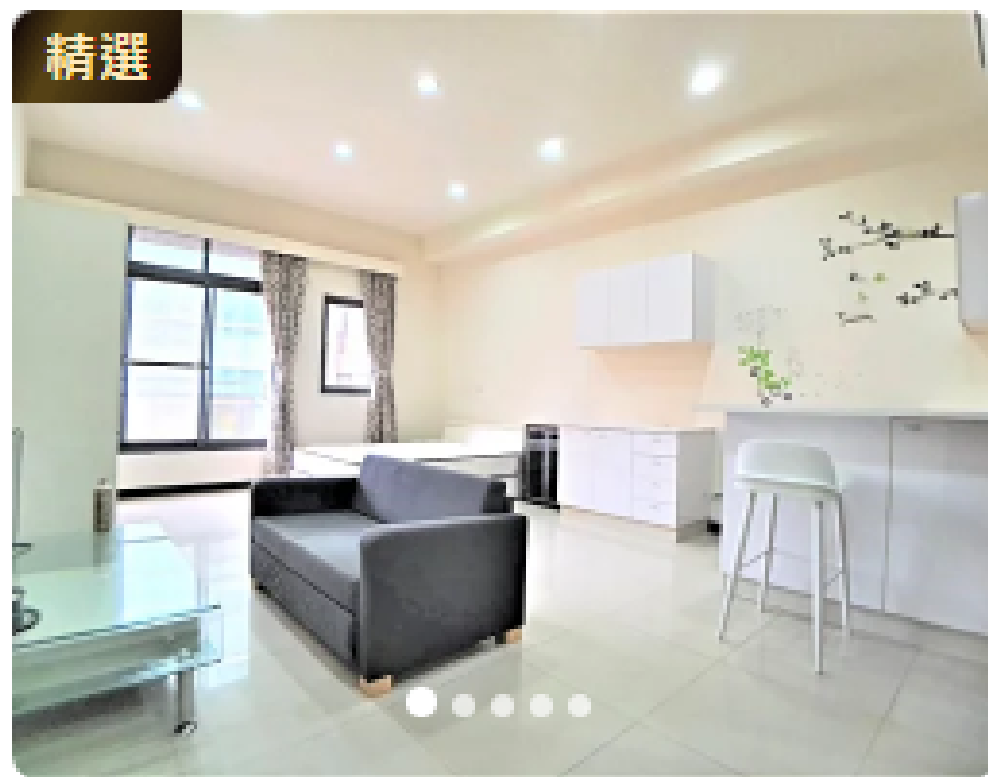
格局 **不限** ☐ 1房 ☐ 2房 ☐ 3房 ☐ 4房以上

特色 **不限** ☐ 新上架 ☐ 近捷運 ☐ 可養寵物 ☐ 可開伙 ☐ 有車位 ☐ 有電梯 ☐ 有陽台 ☐ 可短期租賃

展開選項 (型態/坪數/樓層/設備/須知) 

資料訊息

精選



✨中山國小站 ✨含租補稅新大樓 ✨挑高 ✨陽台

優選好屋

近捷運

新上架

拎包入住

近商圈

有電梯

可開伙


🏠 整層住家 | 開放式格局 | 17.5坪 | 5F/13F

📍 近林森北路中山北路三段中山北路二段雙城街 中山區-農安街

📏 距民權西路站 346公尺

25,999元/月

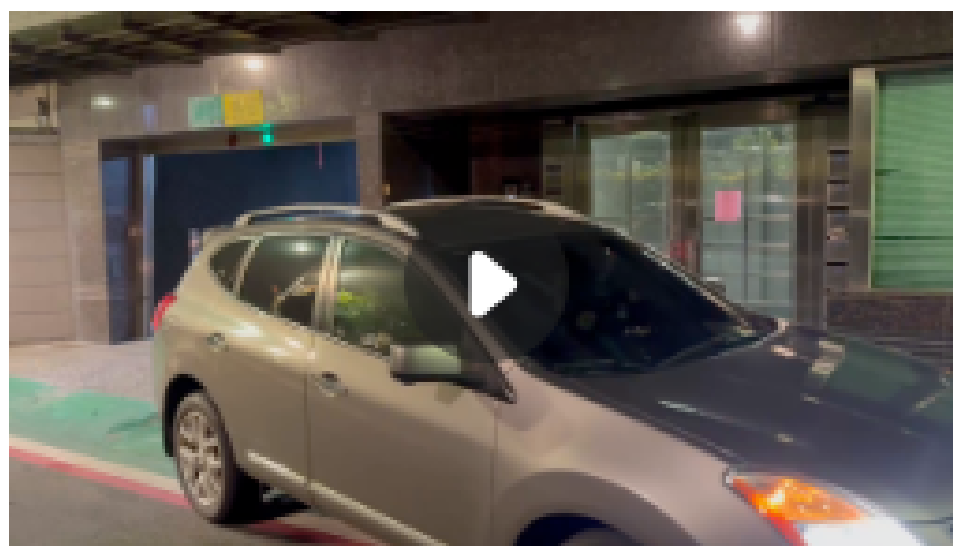
仲介 加賴專員服務 | 5天前更新 | 昨日217人瀏覽

The background features a light cream color with several organic, flowing shapes in shades of pink, peach, and brown. In the top-left corner, there is a cluster of small brown dots of varying sizes. The text "Show json file" is centered in a bold, black, sans-serif font.

Show json file

資料丟棄

置頂



車位出租重慶南路汀州路紹安街可停休旅車



屋主直租

影片房屋

近捷運



車位 | 機械式 | 5坪



頂真社區 中正區-汀州路二段12巷8號



距古亭站 622公尺

屋主 鄭小姐 | 2小時內更新 | 昨日7人瀏覽

4,000元/月

Excel fails

標題中有很多符號沒有對應的
utf-8 編碼, 存入 excel 只會變
成亂碼

Data duplicated

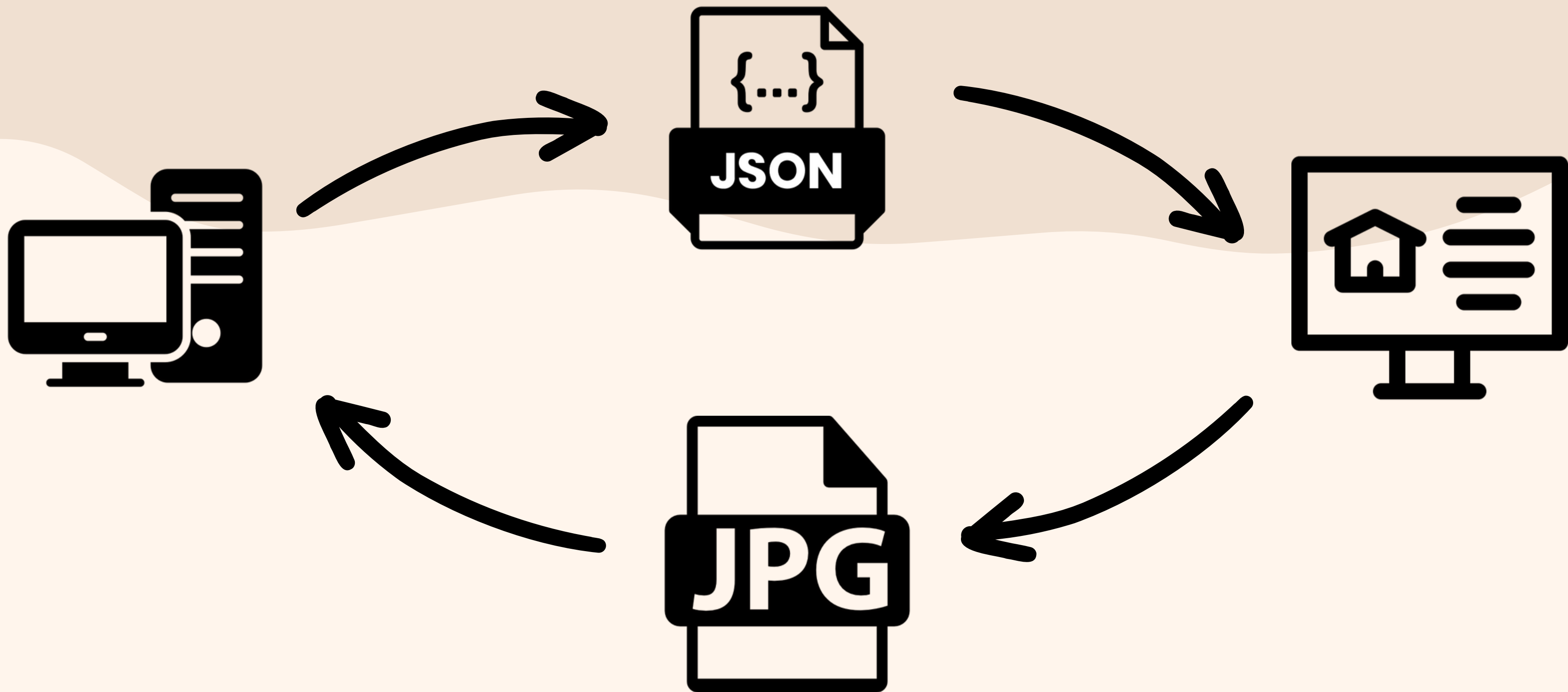
資料在爬取的過程中，網頁可能會重複呈現相同資料在不同頁面，所以才需要加入“hid”用於辨識不同資料

(這樣資料非常少，但仍然需要排除)

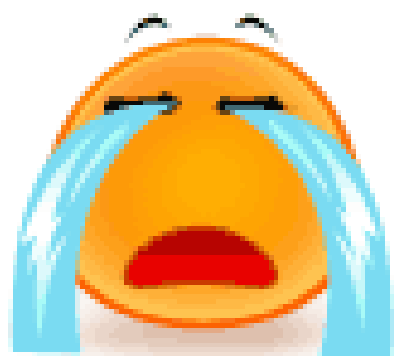


Topic 2

擷取圖片



資料丟棄



很抱歉，您查詢的物件不存在，可能已關閉或者被刪除。

您可以

1、檢查剛才的輸入

2、去[首頁](#)、[租屋](#)、[中古屋](#)、[新房屋](#)、[頂讓](#)、[會員中心](#)、[幫助中心](#)

資料儲存

1.

依照 'hid' 對檔案進行命名, 儲存對應資料的照片(檔案裡面只有照片)

2.

圖片爬取完成建立一個 json file 來儲存每筆資料的資訊及其圖片對應的檔案

The background features a light cream color with several organic, flowing shapes in muted red, peach, and brown tones. In the top-left corner, there is a cluster of small brown dots of varying sizes. The text "Show how to work" is centered in a bold, black, sans-serif font.

Show how to work

(無法一次爬取的原因)

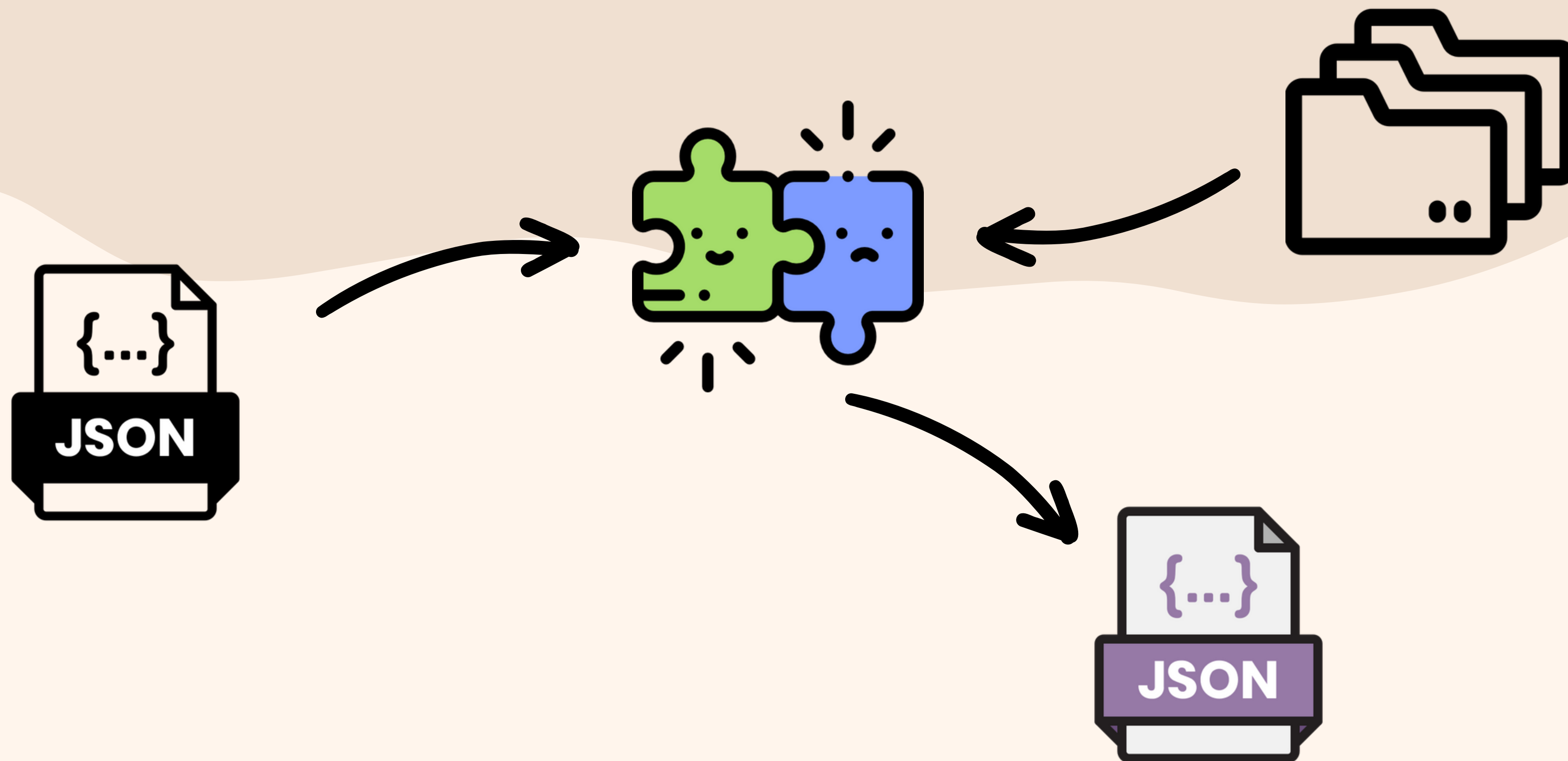
- 請求過快
- 滑動過快
- 網路卡頓
- 網頁卡住
- Exception 處理
- 網頁模板不固定

允許分段爬取圖片，不必一次完成，但後續需要再建立 json file 連結資料與各個獨立檔案的關係



Topic 3

資料比對



THANK YOU
SO MUCH!

