

期末專案名稱 (建議)：AI 美食指南：Google Maps

餐廳評論速讀與摘要

41147022S 謝宇宸

一、 使用模型介紹

本專案核心採用了 Google 開發的輕量級、開源大型語言模型 (LLM) google/gemma-7b-it。Gemma 模型系列基於與其更大型的 Gemini 模型相同的研究和技術建構，旨在提供高效能且易於研究人員和開發人員使用的選擇。選擇 gemma-7b-it (70 億參數指令微調版本) 的主要原因是它在理解指令、生成文本方面具有相對優良的表現，同時透過量化技術，有機會在 Google Colab 這樣資源相對有限的環境中運行。

為了在 Google Colab 的 GPU (NVIDIA T4) 環境中高效運行 gemma-7b-it，我採用了 4-bit 量化技術。這透過 Hugging Face transformers 函式庫合 bitsandbytes 實現，具體設定是使用 BitsAndBytesConfig 來配置 load_in_4bit=True、bnb_4bit_quant_type="nf4" 以及 bnb_4bit_compute_dtype=torch.bfloat16。這種方式能夠顯著降低模型的 VRAM 佔用 (從原始 float16 的約 14GB 降至約 5-6GB)，使得在 Colab T4 GPU (約 15GB VRAM) 上進行推論成為可能，同時盡可能保留模型的效能。

在專案初期，我也曾嘗試過參數規模更小的 google/gemma-2b-it 模型進行原型開發和測試，為後續升級到 7B 模型積累了經驗。

二、 專案目的

在資訊爆炸的時代，消費者在選擇餐廳前往往會參考網路上的顧客評論。然而，熱門餐廳的評論數量可能非常龐大，逐一閱讀費時費力。本專案旨在解決這一痛點，達成以下目的：

1. 提升決策效率：開發一個工具，讓使用者能快速從大量 Google Maps 餐廳評論中獲取核心觀點和整體評價。
2. 自動化摘要生成：利用大型語言模型的自然語言處理能力，自動將多條零散的顧客評論提煉成一段簡潔、易懂的摘要。

3. 便捷的使用體驗：透過 Gradio 建構一個簡單直觀的 Web 使用者介面，方便使用者輸入餐廳名稱並獲取 AI 生成的摘要。
4. 技術實踐與學習：整合外部 API (Google Maps Platform)、探索大型語言模型的應用、掌握模型量化技術，並在資源限制下完成專案部署。

本專案的目標使用者是那些希望在短時間內了解餐廳口碑，以便做出更明智用餐選擇的普通消費者，特別是針對像台北市大安區這樣餐廳密集、資訊豐富的區域。

三、專案重點與過程

1. 系統架構與 Gradio 介面設計：本專案的核心流程是：使用者透過 Gradio 介面輸入餐廳名稱，後端 Python 程式呼叫 Google Places API 獲取餐廳資訊和評論，接著將評論文本傳給 Gemma-7B 模型進行摘要生成，最後將結果回傳並在 Gradio 介面上展示。介面設計力求簡潔，主要包含餐廳名稱輸入框、提交按鈕，以及用於展示餐廳基本資料、原始評論摘錄和 AI 評論總結的輸出區域。

2. Google Maps API 整合與評論獲取：我使用 googlemaps Python 函式庫來串接 Google Places API。首先需要設定 API 金鑰，並啟用 Places API 服務。過程中遇到的挑戰包括：

- 初期遇到 REQUEST_DENIED 錯誤，經查閱文件發現是呼叫了舊版 API 端點，需確保專案啟用了正確的 Places API 版本。
- API 對於單次請求回傳的評論數量有限（通常最多 5 則「最相關」評論），這也限制了我 AI 分析的原始資料量。

3. 大型語言模型 (Gemma-7B) 的選擇、載入與量化：選擇 gemma-7b-it 是為了在 Colab 環境下追求更好的生成效果。載入此模型需要：

- 在 Hugging Face 網站上同意 Gemma 模型的使用條款。
- 產生 Hugging Face Access Token，並在 Colab Secrets 中設定 HF_TOKEN 以供身份驗證。
- GPU 的重要性：初期曾忘記在 Colab 中啟用 GPU，導致 gemma-2b-it 運行極其緩慢（單次請求長達 5 分鐘以上）。啟用 GPU (T4) 後，2B 模型的速度得到大幅改善，也為後續挑戰 7B 模型奠定了基礎。
- 量化探索歷程：

- 嘗試無量化 gemma-7b-it (float16)：即使在 T4 GPU 上，也因 VRAM 不足 (約需 14GB+，超出 T4 上限) 而無法直接載入，或 device_map="auto" 將部分模型放到 CPU 導致極慢。
- 嘗試 8-bit 量化 gemma-7b-it：模型權重約 7-8GB，可以載入 T4 GPU。但在實際進行推論 (生成文本) 時，由於激活值 (activations) 和 KV Cache 的 VRAM 消耗，仍然遇到了 CUDA out of memory 錯誤，尤其是在處理多條評論拼接成的較長輸入時。
- 最終採用 4-bit 量化：透過 BitsAndBytesConfig 設定 load_in_4bit=True、bnb_4bit_quant_type="nf4" 等參數，成功將 gemma-7b-it 的 VRAM 佔用降低到約 5-6GB，使其能在 Colab T4 GPU 上穩定運行並進行推論。

4. 提示工程 (Prompt Engineering) 的迭代與優化：初期，即使模型成功載入，LLM 生成的摘要品質也不理想，例如只回傳無意義的符號 (-) 或重複部分輸入內容。透過在程式碼中加入詳細的 print 除錯語句，我追蹤到：

- Google Maps API 回傳的評論內容是否正確傳入。
- LLM 接收到的完整提示詞 (prompt) 結構。
- LLM 未經後處理的原始輸出 (Full decoded result before processing)。
- 經過後處理 (result[len(prompt_text):].strip()) 後的最終輸出。

針對這些問題，我對提示詞進行了多次迭代優化，主要策略包括：

- 賦予明確角色：例如「作為一個客觀的美食評論分析員...」。
- 清晰的任務指令：詳細描述期望的輸出格式、內容要點 (如包含正面與負面評價)、字數限制。
- 否定性約束：指示模型不要做什麼 (如「不要自行計算平均分數」)。
- 引導性的輸出標記：在提示詞末尾加上「專業評論總結：」來引導模型開始生成。經過優化後，模型生成摘要的相關性和品質得到了明顯改善。

5. Gradio 介面功能調整：根據專案進度和 LLM 在多任務上的實際表現，決定將功能聚焦。原先構想的「AI 風格食記片段」和「AI 提取推薦菜色」功能，由於在有限時間和資源下，提示工程的調校難度較高，且初步效果不如「評論摘要」穩定，因此在最終版本中予以移除，使專案核心目標更為明確。相應地，gradio_interface 函式和 Gradio UI 元件也進行了簡化。

四、專案最終成果

本專案成功開發出一個名為「AI 美食指南：Google Maps 餐廳評論速讀與摘要」的 Gradio Web 應用程式原型。其主要成果如下：

1. 功能實現：
 - 使用者可在介面輸入位於台北市大安區的餐廳名稱。
 - 應用程式能透過 Google Places API 自動獲取該餐廳的基本公開資訊（店名、地址、總體評分）以及最多 5 則顧客評論。
 - 利用 4-bit 量化的 google/gemma-7b-it 大型語言模型，對獲取到的評論進行分析，並生成一段約 50-80 字的「AI 評論總結」。
 - 所有資訊（餐廳資料、原始評論、AI 摘要）均清晰地展示在 Gradio 介面上。
2. 效能與環境：
 - 整個應用程式在 Google Colab GPU (T4) 環境下運行。
 - 透過模型量化和啟用 GPU，單次請求的處理時間（從提交餐廳名稱到顯示 AI 摘要）相較於初期 CPU 運行或未優化時有顯著縮短，達到了可接受的互動水平。
3. 學習與價值：
 - 本專案完整實踐了從數據獲取、AI 模型處理到使用者介面呈現的流程。
 - 深入學習了大型語言模型的載入、量化 (4-bit/8-bit)、提示工程以及在資源受限環境下的部署策略。
 - 掌握了 Gradio 建立互動式 AI 應用的方法。
 - 鍛鍊了 API 串接、錯誤處理和效能除錯的能力。
 - 最終成品可作為一個實用的工具原型，幫助使用者快速了解餐廳評價，提升生活效率。

遇到的主要困難及解決方法回顧：

- API 權限與金鑰設定：透過查閱官方文件逐步完成 Google Maps API 和 Hugging Face Token 的設定。
- 模型「Gated」存取：在 Hugging Face 網站同意模型使用條款，並使用 Token 進行身份驗證。
- Colab GPU 未啟用：初期因忘記啟用 GPU 導致運行極慢，後續修正並體會到 GPU 的重要性。

- VRAM 不足 (CUDA OOM)：在嘗試 7B 模型時，從無量化、8-bit 量化到最終選擇 4-bit 量化，不斷調試以在 Colab T4 有限的 VRAM 中成功運行模型。
- LLM 輸出品質不佳：大量投入時間進行提示工程的優化，透過明確指令、角色扮演等方式引導模型生成符合期望的內容。

截圖：

🍷 AI 美食評論家 (台北市大安區)

輸入大安區的餐廳名稱，AI 將從 Google Maps 獲取評論，並為您生成評論摘要！

請輸入餐廳名稱 (預設搜尋台北市大安區)

篩選

提交，讓 AI 分析！

餅圖蟹酥麵 本店

地址：106台灣台北市大安區師大路33巷14號 Google 總體評分：4.2 星

Google Maps 原始評論摘錄 (最多5條)

- Joanna Wang (評分: 4星): 餅圖蟹酥麵有超碎的酥末，內用座位不多，不過可以立馬品嚐現炸超酥的蟹酥麵，其他小點大部分不錯，不過胡椒鹽給得有點少，內用若覺得味道不夠，可以另外自己加胡椒鹽，外酥內嫩，覺得第一名是雞霸麵再來就是雞重豆腐，九層塔有一點點悶，內用的蔬菜是分鐘也是備油，沒有吃完的也可以自己打包！

- You (評分: 3星): 之前吃過四門店覺得他們的炸的都恰恰的，很好吃，就來總店吃看看。

食物沒什麼含油(油潤的乾淨)一樣恰恰的口感。
喜歡他們的豬血糕，炸的外酥內軟，
甜不辣就正常表現。
蟹酥麵是難得的也是正常表現。
魷魚不是魷魚麵有Q筋的那種，魷魚是整塊都是肉口感很軟，我個人沒很喜歡。

調味上稍偏鹹。

付款方式：現金、全支付、台灣pay、街口匯多支付方式。
- Max CCW (評分: 3星): 瞭解你們想透過賣胡椒鹽多掙點微薄利潤，但可不可以至少在灑上少少的胡椒鹽時可以讓的均勻一些，不然過鹹的和「原味」的蟹酥麵，花枝腳交替著吃實在不是很可口。
- CHI (評分: 5星): 支付方式多元、可內用

推薦湯圓、蟹酥麵、芋條條、玉米筍、雞重豆腐

- Arisa (評分: 5星): 咖哩飯有附湯喝飲料

雞塊很好吃 咖哩飯配角 (?)
內用環境讓人放鬆
內用環境: 2024-09-06

AI 評論總結

這家餐廳的評價良好，主要原因是其食物的美味和價格，以及內用環境的舒適。顧客的評價主要集中在食物的口感、油潤的乾淨、外酥內嫩等方面，以及內用環境的舒適。唯一負面評價是，在調味上稍偏鹹。

整體來說，這家餐廳提供優質的食物和良好的服務，價格合理，值得試驗。

透過 API 使用 使用 Gradio 建構 設定

🍷 AI 美食評論家 (台北市大安區)

輸入大安區的餐廳名稱，AI 將從 Google Maps 獲取評論，並為您生成評論摘要！

請輸入餐廳名稱 (預設搜尋台北市大安區)

增雞滷味

提交，讓 AI 分析！

增雞滷味 創始總店

地址：106台灣台北市大安區師大路43號 Google 總體評分：4.4 星

Google Maps 原始評論摘錄 (最多5條)

- Alina Lin (評分: 4星): 師大增雞滷味一直都是沒什麼味道 價格划算
味道很淡 只有很淡的滷味 雞味也淡 所以可以當成主食吃 主打的就是人流量大 常見的食材自然就新鮮了 要口味重就是要點菜時加辣了

常見學生點諸多蔬菜麵雞食當成主食吃 (現在也差不多)
以前常覺得人多擁擠可怕 衛生不會太好
但是現在師大夜市被整治過後覺得的人潮超少的 只剩下周邊學生居民為主了 再也不是觀光地區 店面也大大整治看起來乾淨明亮許多 至少眼前的乾淨是比過往好了許多

據說子都不再是塑膠盤上面套一個塑膠袋了
而是上面還有增雞滷味印刷啊
我們這些老同儕們的時代眼淚

而且以前那時候還有時令蔬菜特價呢
真的有一種時光流逝的強烈感覺啊

用餐在地下室 也有簡式馬槽
樓下還有主式散餐收盤子
還有30塊可樂罐自助暢飲
桌上竟然還有小時侯泡沫紅茶店的投幣星座運勢抽籤機

旁邊還有個小店看起來是共用地下室的 只是開的天數只有一半

AI 評論總結

這家餐廳的評價主要集中在調味和衛生方面。顧客對調味的評價主要集中在味道淡、雞味淡、新鮮度高等方面，對衛生方面則主要集中在店面乾淨、衛生設施良好等方面。

整體來說，這家餐廳的調味和衛生方面都有優點，但價格可能比較高，因此不適合那些預算比較低的顧客。

透過 API 使用 使用 Gradio 建構 設定

🍷 AI 美食評論家 (台北市大安區)

輸入大安區的餐廳名稱，AI 將從 Google Maps 擷取評論，並為您生成評論摘要！

請輸入餐廳名稱 (預設搜尋台北市大安區)

瑞安豆漿大王

提交，讓 AI 分析！

瑞安豆漿大王

地址：106台灣台北市大安區瑞安街69號 Google 總體評分：4 星

Google Maps 原始評論摘錄 (最多5條)

- Beck H. (評分: 5星): 培根蛋餅皮軟嫩，還會加上炒得香脆嫩嫩的豆芽菜，這樣的組合非常好吃、有特色，是我個人會想再來吃的餐點。看到還有其他顧客點了炒豆芽，頗讓人心動！

鹹豆漿中規中矩還算好吃；生煎包普通。

餐點種類確實，之後若有機會再來，會想嘗試其他餐點。

- Yuna Lin (評分: 4星): 湯包口味不錯，肉很豐富。

湯包不錯，口味剛好，滑嫩的豆腐口感。

蛋餅有個菜香，還不錯。

鹹豆漿豆腐很多，但喝起來都是豆腐跟菜味，有的帶鹹味，整體口感不合口味，不推。

- Wei-jia Lin (評分: 4星): 大概晚上10點到的，位子很多，但是翻桌率很快通常等10分鐘就會有位子，適合很多人一起吃完後。太吵阿該會管秩序

✔ 豆芽菜蛋餅 \$35

每次必點的品項 (要是想吃肉培根蛋餅裡也有豆芽菜) 這邊的醬油膏不會太鹹真的很讚

✔ 雞排 \$40

看到有人點餐覺得很好吃，所以也點了一份，單上面加油條真的是很好的搭配，是屬於比較清淡流軟的那可以自己加碼。

- CT T (doubleespresso) (評分: 2星): 簡單的說是半成品加熱的早餐店，寧靜點這「實」，可能只能吃蛋餅。有別於其他早餐店標榜的鹹和甜風，至少乾乾淨淨，所以多給一星！

- YUNCHING YANG (評分: 2星): 路過看評價數很多且高達四星才來試試看，沒想到在地人的口味讓我驚訝了。鍋貼很多人點，不過是冷的，表面不太像用煎的，有點像是先蒸熟再過一下油。麵皮很厚，溼濕軟軟的，真的很像吃餛飩，餛飩特別酥硬，不是酥鬆。裡面幾乎沒水分，吃起來很乾，還好薄薄的一塊，勉強給過。豆腐跟湯圓，給過，而且可能是店裡最有水準的餐點。湯包沒湯，皮上還黏著一隻蚊子。湯包吃起來感覺像回鍋蒸過，皮也是沒彈性，吃起來口感很慘。來這裡喝豆漿就好

AI 評論總結

這家餐廳提供多種餐點，包括培根蛋餅、湯包、生煎包、湯麵、蛋餅等。顧客的評價主要集中在培根蛋餅、湯包和湯包等餐點上，其中培根蛋餅和湯包的評價良好。顧客對餐點的評價主要集中在味道、價格和服務等方面，其中味道的評價良好，價格的評價良好，服務的評價良好。整體來說，這家餐廳提供優質的餐點，價格合理，服務良好，值得去嘗試。

透過 API 使用 ，使用  建構 ，設定 

Google Colab 連結：

https://colab.research.google.com/drive/13CdvIF_CQENOVxDfOC7MS5e-wo0_MPU0?usp=sharing