餐廳評論文字探勘專案報告

一、專案背景與目的

Google Maps 的餐廳評論往往數量龐大，使用者只能先看到整體星等，若要理解細節（如用餐環境、服務態度、熱門菜色等）就必須閱讀大量文字。另一方面，星等可能因促銷或灌水而失真，因此需要以「文字內容」為主體進行分析。本專案透過文字探勘流程，協助萃取餐廳的優缺點，並提供業者改善方向與消費者參考資訊。

本專案的目標如下：

1. 建立完整的中文評論前處理與分析流程。
2. 建立情緒分類模型（正面/負面）。
3. 萃取評論主題與關鍵詞，呈現優缺點差異。
4. 支援「單一餐廳」查詢，輸出前三好與前三不滿。

二、資料來源與欄位

資料來源為 Google Maps 餐廳評論，主要欄位包含：

- `place_name`：餐廳名稱
- `rating`：星等
- `review_text`：原始評論文字
- `sentiment`：原始情緒欄位（可對照使用）
- `clean_text`：原始清理後文字（資料中已存在）

本專案以 `review_text` 為核心進行清理與斷詞，再依星等自動標註情緒。

三、整體流程與程式邏輯

程式主檔為 `text_mining_project.py`，主要流程如下：

1. **文字清理（clean_text）**
    - 移除網址與特殊符號
    - 保留中文、英文與數字
    - 合併重複字元、統一空白
2. **中文斷詞（jieba）**
    - 使用 jieba 進行分詞
    - 搭配停用詞表 `stopwords_zh.txt` 移除無意義詞彙
    - 產生 `processed_text`，以空白分隔詞彙
3. **情緒標註（label_sentiment）**
    - `rating > 4` 標記為 **positive**
    - `rating <= 3` 標記為 **negative**
    - 不使用 neutral
4. **情緒分類模型**

    - 特徵：TF-IDF
    - 模型：Naive Bayes 或 Logistic Regression
    - 切分訓練/測試集（8:2）
    - 輸出分類報告與混淆矩陣

5. **主題萃取（Topic Modeling）**
    - 方法：LDA
    - 產出每個主題的高權重詞彙
    - 用以解釋「評論在討論什麼」

6. **關鍵詞分析**
    - 針對正面與負面評論分別計算 TF-IDF
    - 產出正/負面高權重關鍵詞

7. **餐廳分店輸出**
    - 依 `place_name` 分組
    - 各餐廳輸出主題、正/負關鍵詞、摘要資料
    - 支援輸出「指定店家前三好/壞」

四、使用的技術與工具

- **Python**：資料處理與模型建置
- **pandas / numpy**：資料整理與統計
- **jieba**：中文斷詞
- **scikit-learn**：
    - TF-IDF 特徵萃取
    - Naive Bayes / Logistic Regression
    - LDA 主題模型
- **JSON/CSV**：輸出中介資料與報告結果

五、主要輸出成果

程式完成後會輸出以下內容（位於 `outputs/`）：

- `preprocessed_reviews.csv`：清理後的完整評論
- `sentiment_metrics.json`：模型評估指標
- `sentiment_confusion_matrix.csv`：混淆矩陣
- `sentiment_predictions.csv`：全量預測結果
- `topic_terms.csv`：主題詞彙
- `positive_top_keywords.csv` / `negative_top_keywords.csv`：正負關鍵詞
- `top_word_frequency.csv`：整體高頻詞彙
- `per_place/`：每家餐廳的細節資料夾與摘要檔

此外可透過以下方式查詢單一店家前三好與前三不好：

```
python3 text_mining_project.py --input final_reviews_for_analysis.csv --
output outputs --place-name "店名"
```

六、結果解讀範例

以單一餐廳為例，可得到：

- **前三好關鍵詞**：代表消費者最常稱讚的面向
- **前三負關鍵詞**：代表主要抱怨點或服務缺失
- **主題詞**：揭示該店最常被討論的主題（餐點、價格、服務、環境）

這些結果可做為：

- 業者改善優先順序
- 消費者快速判斷餐廳特性
- 選址與行銷策略的參考

七、限制與未來改進

1. **情緒標註依賴星等**：若星等不一致或偏差，模型會被影響。
2. **主題解讀需人工命名**：LDA 產出的主題詞仍需人工詮釋。
3. **停用詞需客製化**：不同類型餐廳需要不同停用詞與專有詞表。
4. **可加入地區或餐廳類型分析**：若資料含地區、類型欄位，可進一步分群比較。

八、結論

本專案建立了可重複執行的餐廳評論文字探勘流程，涵蓋文字清理、情緒分類、主題萃取與關鍵詞分析。使用者可以從大量評論中快速得出店家優缺點，並且支援「單一餐廳」查詢，讓分析結果更具實用價值。此流程可延伸至其他類型評論或社群文本分析。