Baseline 程式碼說明

本次比賽提供 baseline 程式碼給參賽者做參考，在初賽中我們採用 BM25 作為資訊檢索的方法，參賽者可以基於這個 baseline 程式碼進行修改，並根據自己的需求進行調整，以提高效能和準確性，或者選擇其他資訊檢索方法來參賽。

BM25 概述

BM25 是一種常用的資訊檢索演算法，基於詞頻和文件的逆向文件頻率來計算文件與查詢的相關性。

環境建立

```
pip install -r requirements.txt
```

Baseline使用方法：

以範例 150 題資料為例 (路徑請自行修正)，並可透過:

```
python bm25_retrieve.py \
    --question_path ../dataset/preliminary/questions_example.json \
    --source_path ../reference \
    --output_path ../dataset/preliminary/pred_retrieve.json
```

其中 bm25_retrieve.py為主程式，負責處理資料檢索和答案生成，其參數分別為

- question_path: 提供問題的JSON檔案路徑。
- source_path: 需要檢索的參考資料的路徑。
- output_path: 產生的預測答案將被儲存在這個路徑中。

修正方向

參賽者可以根據比賽需求，對 baseline 程式碼進行擴展或修改。以下是幾個建議的方向：

- 資料預處理: 在讀取資料之前，增加自定義的資料預處理步驟，以提高模型的輸入品質。
- 資訊檢索方式改進: 嘗試使用其他演算法來提升檢索效能。
- 多模態資料處理: 處理圖片、表格等多種資料型態，以提供更清晰的語意輸入給模型。