

資料生成

目標

- 測試資料只有93筆問答，若拿來訓練模型或評估模型回答成效並不太夠，所以使用Gemini來生成更多與測試資料相同格式的訓練資料集
- 資料格式(“instruction”: ,”input”: ,”output”):

```
{  
  "instruction": "AI709項目1110源來自會科須調整改抓會科數字，測試改抓會科，結果不符預期，是什麼原因？",  
  "input": "",  
  "output": "AI709報表「本月損益」之處理方式為使用當月抓取之「當年度累計」扣除上月抓取之「當年度累計」。"  
},
```

方法

- 請大型語言模型（large language model, LLM)依照提供的檔案生成出與測試案例相同格式的更多訓練資料集（training data set）。
- 希望生成的訓練資料集不要只侷限在93筆測試資料的知識，所以將所有可上雲的檔案當作參考資料提供給模型
 - 中央銀行_申報系統表單與細項清單
 - 本國銀行財報類檔案目錄暨各申報截止日
 - Mantis_AI問題分析測試案例的所有試算表

流程

- 上傳三個參考檔案
- 請Gemini根據參考檔案的資訊，依照prompt生成如測試案例格式的200筆對應問答
- 重複生成n次
- 檢查是否有重複生成的問答，將沒有重複的結果存入json檔中

上傳三個參考檔案

1. 中央銀行_申報系統表單與細項清單
2. 本國銀行財報類檔案目錄暨各申報截止日
3. Mantis_AI問題分析測試案例

重複n次

請Gemini根據參考檔案的資訊，依照prompt生成如測試案例格式的200筆對應問答

檢查是否有重複生成的問答，將沒有重複的結果存入json檔中

資料生成的提示模板與結果

```
prompt = """
根據以下資料表 (df1, df2, df3, df4)，請生成200筆類別-問題-解答對，每筆類別-問題-解答對的格式參考df1中的類別、問題和解答欄位。
每筆生成的解答不超過50字。
請確保每個問題和解答是有意義的，並能夠反映資料中的資訊。

將每一筆生成的類別-問題-解答對以JSON格式輸出，輸出格式如下：

{"類別": "類別1", "問題": "問題1", "解答": "解答1"}\n
{"類別": "類別2", "問題": "問題2", "解答": "解答2"}\n
...
{"類別": "類別200", "問題": "問題200", "解答": "解答200"}

請在生成問題和解答時參考df1, df2, df3, df4中的資料。
"""
```

方法	分次生成，扣掉重複
模型	gemini-1.5-flash
seed tasks	共三個檔案，全部試算表皆使用
prompt	每次一模一樣
總共筆數	6413
格式	指令、輸入、輸出(輸入全為空白)

- 使用 Gemini 生成的訓練資料集可能存在幻覺的問題，所以需要經由人工標註是否有幻覺，但應資料涉及跨部門，所以人工標註有難度

分析債券公開說明書與財報

分析募款目的與實際用途是否一致

- 從債券公開說明書中找出use of proceeds的段落，結合RAG的技術，並請Gemini回答公司發行這檔債券的目的

```
with open('Apple_2016_424B2.txt', 'r', encoding='utf-8') as file:
    content = file.read()
prompt = ("請根據參考資料簡短回答公司發行債券的目的\n"
          "請用繁體中文回答\n"
          "以下為參考資料:\n")
prompt += (content)
print(prompt)
```

公司發行債券的目的是為了籌措約9.952億美元的資金，用於投資符合資格的環保專案。這些專案著重於減少氣候變遷影響、使用更環保的材料以及資源保育。

- 從公司的財務報表中找出實際的資金用途
- 最後計算兩者之間的相似度，確認募款目的與實際用途是否一致