結果說明

	accuracy	onlyGPT	onlyGPTfull	onlyHKBDB	onlyHKBDBfull
last	0.9069	2.5118	4.3397	1.4309	2.3452
now	0.9305	2.4351	1.9373	0.9794	1.5063
	+0.0236	-0.0767	-2.4024	-0.4515	-0.8389

- accuracy 只算計匹配到的資料 (GPT 及 HKBDB 都有，且欄位名稱相同)
- onlyGPT 計算 GPT 多的欄位 (GPT 有, HKBDB 沒有 (既定欄位情況下)) (full 代表多的整筆資料)
- onlyHKBDB 計算 HKBDB 多的欄位 (GPT 沒有, HKBDB 沒有 (既定欄位情況下)) (full 代表多的整筆資料)
- 所有匹配到的欄位其準確率 (相似度) 為 93.05 %
- 平均每個表單 GPT 能多擷取 2.43 筆欄位
- 平均每個表單 GPT 能多擷取 1.93 筆 "事件" (也就是二維資料的筆數)
- 平均每個表單 HKBDB 多 0.97 筆欄位
- 平均每個表單 HKBDB 多 1.50 筆 "事件"

MSE & RMSE

Mean Square Error(MSE) and Root Mean Square Error(RMSE) are evaluation functions in regression problems. But there are no so-called "actual values" and "predicted values" in our case. We just only "HKBDB answers" and "GPT answers", so first we compare the similarity of the two answers, and we get a value between 0-1. We assume 1 minus this value is the difference between the actual value and the predicted value, and then MSE and RMSE can be calculated.

	MSE	RMSE
BasicInformation	0.01552	0.12459
Education	0.01363	0.11676
Work	0.01354	0.11636
Publication	0.01722	0.13123
Article	0.03966	0.19914
RelativeEvent	0.00953	0.09762

	MSE	RMSE
Honor	0.01386	0.11773
RelatedOrganizations	0.01015	0.10074
Connections	0.01508	0.12279

此次調整 (細項請參考 excel)

- 刪除 publication 中屬於"編輯者"的事件
- 刪除 HKBDB 重複描述的事件
- 工作經驗中刪除部分 (如:電影、劇本、主持)
- GPT 在 publication 加入 [翻譯簡目] & [研究資料書目]
- openai 模型有更新，故部分舊結果不好的有重跑
- HKBDB 中日期相關欄位，拿掉 "-" (1999-10-30 -> 19991030)
- HKBDB 中地點相關欄位，拿掉"__PLA*"

---

model

1. Define output template, define here what we want to capture and its description

```
example_json = {
        "NameInformation": [
            {
                "常見名稱": "常見名稱",
                "號": "像是蘇軾號 '東坡居士'",
                "共同筆名": "多個人一起共同使用的筆名，有些人會有多個筆名，但都不是
與他人共用的筆名",
                "本名": "本名",
                "筆名": "筆名1, 筆名2...",
                "原始資料": "原始資料",
                "字": "像是屈原，名平字原"
            }
        ]
    }
```

2. System message, the model will listen to system messages more, so we can give the model a role here to make it more integrated into the situation.

```
system_message = """你是一位厲害的私人偵探，接下來會給你一段文章，並根據文章內容填入
主角相關資訊(Name Information)。
    資料格式應該像是: """ + json.dumps(example_json) + """。
```

```
    請確實找到文章內容再填入，資料正確很重要，一步一步慢慢做，即使多欄位空白也沒關係"""
```

3. User message, give the data

```
user_message = """將下列文章填入格式中，並以json輸出: """ + data (文章內容)
```

4. Define rules, make the output of the model more consistent with the set conditions

```
rule = """
    請遵循以下規則:
    1.以繁體中文回答
    2."原始資料" 請放文章內的引用，要簡短，不可超過20字。
    3.我們是調查文學作家，所以其他作品不包含文章、著作、職務變更、創業、公司相關類型。
    4.其他作品如 "音樂"、"電影"、"論文"、"劇本"
    5.不一定每位作家都有其他作品，沒有就不要填
    """
```

5. Create the model

```
response = client.chat.completions.create(
        model="gpt-4o",
        temperature=default_temperature,
        response_format={"type": "json_object"},
        messages=[{
            "role": "system",
            "content": system_message
        }, {
            "role": "user",
            "content": user_message
        }, {
            "role": "user",
            "content": rule
        }])
```

- model: model name to use.
- temperature: a number between 0 and 2. The temperature is used to control the randomness of the output. When set it higher, will get more random outputs. When set it lower, towards 0, the values are more deterministic.
- response_format: let model output the json format.
- messages: the above three prompts for the model.

The use of prompt words can be summarized into the following points:

- Play a specific model or role.
- Do it slowly, step by step, etc. keywords.
- Important words are enclosed in quotation marks.
- Give examples.
- Describe the task in detail.

計算公式 (accuracy)

- 完全匹配: 一模一樣才給分
    - (台灣教育局, 台灣教育所) = 0
    - (台灣教育局, 台灣教育所) = 100
- 是否包含: 只要有一方包含在內，即得滿分，否則獲得相似分數
    - (金融學系, 金融系) = 100
- 無序匹配: 以空格為斷行, 應用在如筆名 (多個答案放在 list, 順序不重要)
    - ([查理, 姚馥蘭], [姚馥蘭, 查理]) = 100

計算方式 (onlyGPT & onlyHKBDB)

- 一維資料，一欄位算一分
- 二維資料分兩種計分
    - 同一筆事件中多填的欄位，一欄位加一分
    - 若是直接多一整筆事件，加一分，記在 full 表單

---

compare flow

- 刪除空白的表單
- 刪除無 "關鍵欄位" 的二維資料，如 "article" 表單有大量無 "文章名稱" 的資料，會將其刪除
- 判斷目前計算的表單屬於一維還二維資料
    - 一維: 從 data&ans 取得 key 對應的值,
        - 若兩種都取到，計算相似度 (accuracy)
        - data 有, ans 沒有, 計算 onlyGPT
        - data 沒有, ans 有, 計算 onlyHKBDB

    - 二維: 因二維是由多個事件構成，彼此沒有順序, 直接比對可能會比錯筆資料如
GPT	HKBDB
Data 1	Data 2
Data 2	Data 1
        - 先找到 data 的關鍵欄位 (可能像是學校名稱、公司名稱)
        - 比對 ans 所有的關鍵欄位，以找到相關事件
        - 接著比對方法如一維資料
- 若有某一 excel 少對方 sheet, 用 onlyGPT or onlyHKBDB 計算