

# 比賽資料說明

本次比賽使用網路上供社會大眾查閱的公開資料，如下說明

| 資料類別 ( <code>category</code> ) | 簡介                | 參考文件 ( <code>reference</code> ) 檔案類型 |
|--------------------------------|-------------------|--------------------------------------|
| faq                            | 玉山銀行官方網站上的常見問題    | .json                                |
| insurance                      | 玉山銀行代銷的保險產品之保單條款  | .pdf                                 |
| finance                        | 公開資訊觀測站上的上市公司財務報告 | .pdf                                 |

下圖表示了初賽的資料夾結構:

```
1 | ├── dataset
2 | |   ├── preliminary
3 | | |   ├── questions_example.json
4 | | |   └── ground_truths_example.json
5 | └── reference
6 |     ├── faq
7 |     |   └── pid_map_content.json
8 |     ├── insurance
9 |     |   ├── 1.pdf
10 |    |   ├── 2.pdf
11 |    |   └── ...
12 |     └── finance
13 |         ├── 0.pdf
14 |         ├── 1.pdf
15 |         └── ...
```

- `questions_example.json` 為範例題目，共有 150 題，供參賽者練習使用。
- `ground_truths_example.json`: 範例題目的答案，供參賽者練習使用。
- 資料夾 `reference` 存放各類型資料的參考文件

## 資料格式-初賽

### 1. 主辦方發佈的題目格式 ( `questions_example.json` )

`questions_example.json` 的格式與正式比賽題目 `questions.json` 相同

```

1  {
2      "questions": [
3          {
4              "qid": 1,
5              "source": [442, 115, 440, 196, 431, 392, 14, 51],
6              "query": "匯款銀行及中間行所收取之相關費用由誰負擔?",
7              "category": "insurance"
8          },
9          // 後面題目省略...
10     ]
11 }

```

題目包在 `questions` 串列裡，每個題目皆有：

| 欄位       | 型態              | 說明  |
|----------|-----------------|---|
| qid      | integer         | 題號  |
| query    | string          | 問題  |
| source   | list of integer | 能夠回答問題的可能選項，數字的意義為文件編號 ( <code>pid</code> )，可在資料夾 <code>reference</code> 中找到對應的檔案或內容，例如：<br>- 當 <code>category</code> 為 <code>insurance</code> 或 <code>finance</code> ：pid 13 對應到文件庫中 <code>13.pdf</code><br>- 當 <code>category</code> 為 <code>faq</code> ：pid 13 對應到 <code>pid_map_content.json</code> 中，鍵為 13 的內容 |
| category | string          | 資料類型， <code>reference</code> 裡有對應的資料夾存放該類型的文件   |

## 2. 參賽者提交的答案格式 ( 請命名為 `pred_retrieve.json` )

```

1  {
2      "answers": [
3          {
4              "qid": 1,
5              "retrieve": 926
6          },
7          // 後面題目省略...
8      ]
9  }

```

參賽者提交的預測結果放在 `answers` 串列裡，每個題目皆有：

| 欄位       | 型態      | 說明  |
|----------|---------|---|
| qid      | integer | 題號  |
| retrieve | integer | 文件編號 ( <code>pid</code> )，您認為問題的答案能在此文件編號的內容中找到 |