真相只有一個: 事實文字檢索與查核競賽

評分標準

1. 提交檔案請使用.jsonl 檔，內容格式需符合比賽格式規定。上傳檔案內容請使用 UTF-8(無 BOM 檔首) 編碼，並使用 Unix 系統換行字符。請勿使用其他 Non-Printable Characters。**並注意提交答案內之"文章名稱"，其文字應與主辦單位資料庫內使用之文字一致，以避免評分失誤的可能。**
2. Leaderboard 系統會對每次的提交結果進行評測，以最高分那一次呈現於 Leaderboard。若出現參賽隊伍同分情形，以上傳繳交時間判斷排名順序。競賽期間參賽隊伍會得到 Public Leaderboard 評測的分數做為參考，Private Leaderboard 排名與分數則於競賽結束後公布，並以 Private Leaderboard 之結果為最終排名參考依據。
3. 競賽測試集 Private Dataset 將於 5/29 (一) 上午 11:00 開放下載，同時可開始上傳答案。**請注意！5/29(一)上午 00:00 至 10:59:59 之間，不提供答案上傳功能，5/29(一)上午 11:00 重新開放。**
4. 5/29 (一) 上午 11:00 至 6/2 (五) 下午 14:00 之間，可上傳 Private Dataset 預測結果，每日上傳次數上限為 3 次，逾時則不予評分。此 5 日期間，每日上傳次數以檔案計算，若 Public Dataset 與 Private Dataset 預測結果合併於同一份檔案提交，則僅計算為 1 次提交。若 Public Dataset 與 Private Dataset 預測結果各自單獨一份檔案提交，則將計算為 2 次提交。

下載格式

提供參賽者下載之檔案共分三大類：1. 訓練資料集，包含「中文維基百科資料」與「公開訓練資料集」；2. Public 測試資料集；3. Private 測試資料集。

說明如下：

1. 中文維基百科資料 (封存的版本為 Dec 2022 dump)

- 格式為 .jsonl，已被切分成數個檔案，每個檔案有 50,000 筆處理過的條目。
- 每一行 (row) 代表一個條目，每個條目都可以對應到維基百科的文章。
- “id”代表文章名稱，“text”代表處理後的文章，“lines”代表維基的原始資料。

2. 公開訓練資料集：下載的檔案為 public_train.jsonl，資料的內容包含

- 每一行 (row) 代表一個陳述句及其正確答案。
- 每一行 (row) 的資訊包含 “id”:樣本代號、“label”:陳述句的驗證類別、“claim”:陳述句文字、"evidence"：證據組。
- "evidence" 的格式為 [<annotation_id>, <evidence_id>, 文章名稱, 第幾句]，是由多層的 lists 所構成，代表一筆陳述句可能有多組證據，且每一組證據可能包含多個證據句 (可於資料集中最內層的 list 觀察到)。

  若一組證據包含多個證據句，則該組的每個證據句都會有相同的<evidence_id>。每一組證據句在資料的意義上表示可以單獨支持或反對陳述句，若一組證據有兩個以上的句子，即代表該兩句證據句需要合併才能夠支持或反對該筆陳述句。證據組中的每一個證據句都帶有相同的<annotation_id>，且若一筆陳述句的"label"是 "NOT ENOUGH INFO" 則<evidence_id>為 None。

支持或反對的範例

```
{
        “id”: 123,
        “label”: "SUPPORTS",
        “claim”: “樂山大佛修建了 90 年。”,
        "evidence": [
                [
                   [<annotation_id>, <evidence_id>, "樂山大佛", 0],
                      [<annotation_id>, <evidence_id>, "樂山大佛", 1]
                ],
                [
                      [<annotation_id>, <evidence_id>, "佛教", 0]
                ],
    ]
}
```

沒有足夠資訊的範例

```
{
        “id”: 2366,
        “label”: "NOT ENOUGH INFO",
        “claim”: “國立成功大學有 12 個健身房。”,
        "evidence": [
                [
                      [<annotation_id>, None, None, None]
                ],
    ]
}
```

3. Public 測試資料集 & Private 測試資料集：

- 格式為 .jsonl，內含多個陳述句。
- 每一行 (row) 代表一個陳述句，每個陳述句的內容包含兩個欄位："id" 與 "claim"。其中 "id" 是該筆陳述句的代號，"claim" 則是陳述句內容。
- 提交之結果檔案格式請參考下述「上傳格式」說明。

上傳格式

提交檔案請使用.jsonl 檔，檔案中的每一行 (row) 代表一個陳述句的樣本，且每一行必須包含 "id"、 "predicted_label" 以及 "predicted_evidence"，如以下範例所示：

反對且證據只有一句
{"id": 22334, "predicted_label": "SUPPORTS", "predicted_evidence": [["樂山大佛", 3]]}
支持且證據不只一句
{"id": 78123, "predicted_label": "REFUTES", "predicted_evidence": [["尼克·貝爾格", 3], ["伊拉克", 0], ["伊拉克", 1]]}
沒有足夠資訊
{"id": 12345, "predicted_label": "NOT ENOUGH INFO", “predicted_evidence”: None}

其中：

- id 代表樣本 (陳述句) 代號。每一個 id 只能上傳一個預測答案(row)，若一個 id 重複上傳超過一個預測答案(row)，則不予計分。
- predicted_label 代表該樣本的 3 種預測類別，需與訓練資料集內所提供之類別相同，包含："**SUPPORTS**"、 "**REFUTES**"、以及"**NOT ENOUGH INFO**"，大寫或小寫不影響評分結果，但需注意拼字錯誤。
- predicted_evidence 是預測該樣本類別的證據句，格式為：[“文章名稱”, 第幾個句子]。其中，文章名稱請務必與中文維基百科資料內所提供的資訊一致，並請注意括號的使用，以避免造成評分上的錯誤。第幾個句子，請參考「中文維基百科資料」內之“lines” 行號。
- 須注意，上傳答案時 predicted_evidence 證據句的上限為 5 句，答案不符合規定則不予評分，請參賽者自行將最有可能的證據句進行排序。

評分方式

參賽者需要建立系統去預測陳述句 (claim) 的類別，其中類別的部分包含「支持」(SUPPORTS)、「反對」(REFUTES)、「沒有足夠資訊」(NOT ENOUGH INFO)。

我們將以 Strict Accuracy 作為評估標準，公式如下：

$$\textit{Strict Acc} = \frac{(\text{正確預測「支持」的樣本數}) + (\text{正確預測「反對」的樣本數}) + (\text{正確預測「沒有足夠資訊」的樣本數})}{\text{總樣本數}}$$

其中 (正確預測「支持」的樣本數) 與 (正確預測「反對」的樣本數) 需提出與答案 (Ground-truths) 至少一組相符的證據句才會列入計算。換言之，我們考慮的是嚴格的準確率。

例如，當正確答案為：

```
{
    “id”: 00000,
    “claim”: “尼克·貝爾格在 2004 年前往中東尋找生意商機。”,
    “label”: “SUPPORTS”,
    “evidence”: [
        [["尼克·貝爾格", 0]],  // 第一組證據，有一個句子
        [[“尼克·貝爾格”, 3], [“伊拉克”, 0]],  // 第二組證據，有兩個句子
    ]
}
```

上傳預測答案可能分為幾種情況，各自會得到不同的評分結果，如下表：

預測範例	Case1	Case2	Case3	Case4
predicted_label	“SUPPORTS”	“SUPPORTS”	“SUPPORTS”	“REFUTES”
predicted_evidence	[["尼克·貝爾格", 0], ["尼克·貝爾格", 3], ["伊拉克", 0]]	[["伊朗高原", 7] , ["伊拉克", 0], ["尼克·貝爾格", 3]]	[["尼克·貝爾格", 3]]	[["尼克·貝爾格", 0]]
結果	正確，列入計算	正確，列入計算	錯誤，不列入計算	錯誤，不列入計算

預測範例	Case1	Case2	Case3	Case4
原因	預測類別正確，且證據句完全正確。	預測類別正確，且回答出至少一組正確的證據。	雖然預測類別正確，但沒有完全答對任何一組證據。	雖然回答出一組正確的證據，但類別預測錯誤。

由於此範例的“label”為“SUPPORTS”，因此 Case1 和 Case2 會被列入正確預測「支持」的樣本數，並且 Case3 跟 Case4 不會被列入正確預測「支持」的樣本數。

請注意！對於每筆陳述句，我們僅採納 5 個證據句進行計算，若參賽者上傳的檔案中於陳述句的“**predicted_evidence**”提供了 6 個(含)以上的證據句，將不予以評分、也不扣除當日的評分次數。因此參賽者需要自行將最有可能的證據句進行排序，確保答案符合格式要求以利後續評分。

如果是類別為「沒有足夠資訊」的樣本，則不需要考慮證據句的正確性，只需要預測的類別 (predicted_label) 正確即可。