**附件 一、資料集說明文件格式(範本)**

**生活言談資料集**

**資料集說明文件 v0.9**

**建置團隊資訊**

**子計畫名稱：\_可信賴言談感知人類語言技術\_**

**子計畫主持人：\_\_\_\_\_\_\_陳信希\_\_\_\_\_\_\_**

**說明文件版本：\_\_\_\_\_\_\_\_\_v-1\_\_\_\_\_\_\_\_\_**

**日期：\_\_\_2023-5-01\_\_\_\_\_\_**

1. **資料集基本說明**

|  |  |
| --- | --- |
| **資料集名稱**  | 生活言談資料集 |
| **資料集摘要 (含目的/用途/價值/標註重點等)** | 我們通過那些由眾人在兩個不同時間寫的關於現實生活事件的故事，並比較兩次敘述的差異。 |
| **資料集聯絡人** | 林佑恩/yelin@nlg.csie.ntu.edu.tw/0976512652陳柏君/pcchen@nlg.csie.ntu.edu.tw |
| **資料類型** | 文字 |
| **資料** | **資料集類型** | 純文字資料集，並標記事件的文字範圍及類型 |
| **儲存格式** | json |
| **來源機構** | Microsoft |
| **收集期間** | 2021.7~2021.11 |
| **案例數** | 1260 |
| **筆數/案例** | 1260 |
| **總共容量** | 24 MB |
| **適法性基礎****(本項暫不對外公開)** | □ A類(新收案)：OOO例□ B類(曾取得概括同意，已補告知程序)：OOO例□ C類(未曾取得同意，已補告知程序)：OOO例□ D類(資料已去識別，無法進行告知)：OOO例 E類(資料無個資，不需進行告知)：1260例 |
| **其它說明** |  |
| **標註資料** | **標註工具** | 自建網頁系統 |
| **標註方式** | □自動標註人工標註□自動標註+人工校正□其他：\_\_\_\_\_\_\_\_\_\_ |
| **註記方式** | □有註記：\_\_\_\_\_\_\_\_\_\_□其他：\_\_\_\_\_\_\_\_\_\_ |
| **標註儲存格式** | json |
| **標註情形** | 全部標註完成：1260例□僅標註部份項目：OOO例□僅原始資料，未加標註：OOO例 |
| **其它說明** |  |
| **與其他資料庫的串接** | □ 保留與本團隊其他資料庫串接可能性：未來可透過假id串接本團隊其他的資料庫□ 保留與國網中心資料庫串接可能性：與上傳國網中心的資料採同樣的假ID，以利後續資料更新/維護 (但僅限由本團隊主動進行更新與維護) 無串接規劃□ 其他： \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

1. **案例相關資訊欄位說明**

| **案例資訊項目** | **欄位名稱****(呈現於檔案的名稱)** | **欄位格式****(數值/字串/文字等)** | **範例/備註說明** |
| --- | --- | --- | --- |
| 個案編號 |  |  |  |
| 基本資料 | 出生日期 |  |  |  |
| 性別 |  |  |  |
| 其他 |  |  |  |
| … |  |  |  |
| 資料取得設備資訊 | 取得日期 |  |  |  |
| 資料取得設備型號 |  |  |  |
| 其他 |  |  |  |
| … |  |  |  |
| 每筆資料描述(選填) | 代碼 |  |  |  |
| 分類 |  |  |  |
| 其他 |  |  |  |
| … |  |  |  |
| 延伸資訊 | 報告 |  |  |  |
| 其他 |  |  |  |
| … |  |  |  |

1. **標註說明範例**

使用到下述檔案

|  |  |  |
| --- | --- | --- |
| **標註項目** | **資料檔** | **標註檔** |
| **生活事件** | **hippocorpusV2.csv** | **NIR.json** |
|  |  |  |

* 1. 資料與標註資料的對應關係
	2. 標註內容

標註欄位說明

|  |  |  |
| --- | --- | --- |
| **欄位名稱** | **說明** | **檔案截圖** |
| pair\_id | 文章配對 ID |  |
| story\_type | 事件來源文章之類型 |  |
| explicitness | 事件是否是顯性 |  |
| subject\_token\_ids | 主詞詞序 |  |
| predicate | 謂詞 |  |
| predicate\_token\_ids | 謂詞詞序 |  |
| object\_token\_ids | 受詞次序 |  |
| time\_token\_ids | 時間詞序 |  |
| event\_type | 事件類型 |  |
| supports | 關聯事件代號 |  |

1. **工具**

可以使用下列工具從儲存的檔案中讀取資料

| **資料類別** | **公開/免費軟體名稱/版本** | **商用軟體名稱/版本** |
| --- | --- | --- |
| 資料 | 登入微軟帳號即可免費取得 |  |
| 標註資料 | 任意文字閱讀器都可 |  |

1. **去識別化**

|  |  |
| --- | --- |
| 去識別化方式 | 遵循本專案共同制訂方式 |
| * 本資料集自行定義方式 (勾選本項請逐欄位說明各欄位之處理作法，使用工具，參考資料與連結)

說明： |