**附件 一、資料集說明文件格式(範本)**

**「健康言談資料集」(資料集名稱)**

**資料集說明文件 v0.1**

**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：\_\_\_\_\_\_\_\_\_\_v0.1\_\_\_\_\_\_\_\_\_\_\_\_\_**

**日期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_111/09/05\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **資料集基本說明**

|  |  |  |
| --- | --- | --- |
| **資料集名稱** | | 健康言談資料集 |
| **資料集摘要 (含目的/用途/價值/標註重點等)** | | 目的：提供健康諮詢言談文字生成之學術研究  用途：訓練語言模型  價值：為目前學術領域中相對稀少的繁體中文資料集  標註重點：無標註 |
| **資料集聯絡人** | | (陳柏君/pcchen@nlg.csie.ntu.edu.tw) |
| **資料類型** | | 選擇(結構化數據，語音，文字，影像，其他)  文字 |
| **資料** | **資料集類型** | 文字 |
| **儲存格式** | .csv |
| **來源機構** | (衛生福利部 台灣e院) |
| **收集期間** | 2022/12 |
| **案例數** | 86,399 |
| **筆數/案例** | (若每個案例收錄多筆資料請註明) |
| **總共容量** | 壓縮後（.zip）126MB |
| **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  R E類(資料無個資，不需進行告知)：86,399例 |
| **其它說明** | 無 |
| **標註資料** | **標註工具** | 無 |
| **標註方式** | □自動標註  □人工標註  □自動標註+人工校正  R其他：\_\_\_\_無標註\_\_\_\_\_\_ |
| **註記方式** | □有註記：\_\_\_\_\_\_\_\_\_\_  R其他：\_\_\_\_\_無\_\_\_\_\_ |
| **標註儲存格式** | 無 |
| **標註情形** | □全部標註完成：OOO例  □僅標註部份項目：OOO例  R僅原始資料，未加標註：86399例 |
| **其它說明** | 無 |
| **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)  R 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 | |  |  | 無此項資料 |
| 基本資料 | 出生日期 |  |  | 無此項資料 |
| 性別 |  |  | 無此項資料 |
| 提問 | source | 字串 |  |
| 回覆 | target | 字串 |  |
| 原始資料網址 | URL | 字串 |  |
| 科別 | DPT | 字串 |  |
| 資料取得設備資訊 | 取得日期 | 2022/12月 |  |  |
| 資料取得設備型號 |  |  | 無此項資料 |
| 其他 |  |  |  |
| … |  |  |  |
| 每筆資料描述(選填) | 代碼 |  |  |  |
| 分類 |  |  |  |
| 其他 |  |  |  |
| … |  |  |  |
| 延伸資訊 | 報告 |  |  | 無 |
| 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** |
| 無標註 | 無標註 | 無標註 |
|  |  |  |

* 1. 資料與標註資料的對應關係

無標註

* 1. 標註內容

標註欄位說明

|  |  |  |
| --- | --- | --- |
| **欄位名稱** | **說明** | **檔案截圖** |
| 無標註 | 無標註 | 無標註 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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 | Python | Excel |
| 標註資料 | 無標註 | 無標註 |

1. **去識別化**

|  |  |
| --- | --- |
| 去識別化方式 | R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： |