東吳大學巨量資料管理學院學士專題成果報告書

Bachelor Degree Program

School of Big Data Management

Soochow University

Report of practices

自動化醫學詞彙檢索系統

巨資四 A 09170141 蘇睿詮

巨資四 A 09170171 張升祥

巨資四 B 09170281 高麒祥

巨資四 B 09170282 蘇子恩

巨資四 B 09170284 林羿帆

指導老師：蘇明祥 教授

摘要

本專題成果，主要是在廣泛的醫學領域中，應用了自然語言處理技術(NLP)來建立醫療詞彙標註系統。因為在傳統的醫學文本人工處理時，需要耗費大量時間且容易出錯，然而自動化的詞彙標註系統則可快速且準確地從文本中提取有價值的醫學信息。因此本專題主要針對傳統人工標註方法的缺點，提出利用機器學習和深度學習技術，對醫學文本進行醫學專有詞彙標註以提高醫學詞彙辨識能力，最後使用這些學習所獲得的模型，建立一個醫療詞彙標註系統網站供使用者使用。使用者輸入醫療相關句子或文本，本系統會自動返回使用者輸入內容之標註結果。整體而言，此系統不但能夠幫助醫療專業人士取得所需之醫療關鍵詞，不但能節省人工處理時間並增加工作效率，且對於未來面臨著龐大的醫學資訊，更能夠準確和有效率的處理。

目錄

第一章、 緒論

在廣泛醫學領域中存在著大量的醫學文獻、臨床記錄和病歷等豐富的醫學資訊。然而，這些資訊包含了大量的醫療專業名詞，如疾病[1]、症狀[2]、藥物[3]、手術程序[4]等。因此對於從事醫療專業人員而言，快速且準確地從這些文本中提取有價值的醫學信息對於臨床決策[5]、研究和醫療相關應用至關重要。目前這些大量醫學文本仍使用人工處理來分析與提取有用之醫學信息，但這個處理過程可能非常耗時且容易出現遺漏或錯誤的情況。因此為解決此情況產生的問題，本專題擬使用自然語言處理（natural language processing, NLP）技術來建立醫療詞彙標註系統[6]，解決傳統處理方式的缺點。其中，本專題擬利用機器學習及深度學習技術，解決模型的有效性和醫學中出現文字之複雜性，可快速有效地解決標註所花的時間，更可提升醫學專有名詞標註的正確性。

本專題所提之方法，不僅可快速對醫學文本進行詞彙之標註，更能提高對醫學詞彙的正確辨識，從而迅速對醫學信息的詳細檢索，去除傳統人工標註上所花的時間和其中可能因人為因素產生的嚴重錯誤。相比於傳統的醫學專業詞彙識別，大多需要依賴人工進行，而且耗時且容易出現錯誤。因此利用 NLP 技術建立醫療詞彙標註系統，可以快速對醫學文本進行標註以及減少人員處理所需的時間。總而言之，醫療詞彙標註系統可以幫助該領域專業人士快速從眾多的醫學文獻、臨床記錄和病歷中檢索相關的醫學信息，節省大量時間並提高工作效率。

本醫療詞彙標註系統，當使用者輸入醫療相關句子或文本時，會經過模型之演算法處理，產生對應之醫療詞彙標註結果。使用者可使用醫療詞彙標註類別應用於的醫療檢索或其他之醫學應用，並且在醫學信息處理中能夠發揮著關鍵作用。其中之醫療詞彙標註類別如下:身體（構成人或動物的整個物理結構）、症狀（由特定疾病引起的任何不適感或身體或精神變化）、儀器 (全名,INST) 、檢查(檢查並發現可能的疾病的行為)、化學（通常在人體中發現的任何基本化學元素）、疾病（由感染或健康狀況不佳而非意外事故引起的人或動物疾病）、藥物（用作藥物的任何天然或人工製造的化學品）、補充(添加到其他東西中以改善人體健康)、治療（一種用於治療疾病的行為方法）和時間（時間為定義下的代表性事件）。身體詞彙類別能夠幫助識別人或動物的生理結構，有助於準確描述病情或解釋病因。症狀類別對於追蹤特定疾病引起的不適感或身體變化，有助於早期診斷和治療。儀器和檢查類別則有助於確定醫療檢查以及診斷的關鍵細節。化學和藥物類別用於識別治療和藥物相關信息，能夠幫助選擇合適的治療方案。疾病類別使用於追蹤和分析疾病的流行病學信息。補充類別有助於識別營養補充劑，有提高健康管理效果。治療類別則用於醫療方案和手術的描述，能夠支持臨床決策。最後的時間類別，對於跟蹤事件和治療進展有著極大幫助。這些類別的應用，能夠使

得有這方面需求的醫學專業人員，更有效地管理和應用醫學信息，促進疾病診斷和治療的準確性和效率。

未來的醫學領域將會面臨著龐大的醫學資訊，包括文獻、臨床記錄和病歷等，並且其中蘊含大量醫學專業詞彙。這些專業詞彙之信息對於臨床決策、研究和醫療應用至關重要，因此建立一個自然語言處理（NLP）技術支持的醫療詞彙標註系統，應包含高效性、時間節省、準確性和正確辨識、和多種詞彙類別的識別。而且，對於未來可能需要面臨的兩點:使用者友好性、隱私安全，尤為重要。使用者友好性將確保醫療專業人員和研究人員能夠輕鬆使用系統、介面直觀、快速上手。系統嚴格的隱私安全，能夠確保醫療資訊的機密性得到妥善保護。總而言之，這個醫療詞彙標註系統將成為醫學領域的關鍵工具，有望提高效率、確保準確性，並支持多種醫學詞彙的識別，從而更好地管理和利用豐富的醫學信息資源。

第二章、 研究方法與步驟

圖 2-1 研究方法之主要步驟

第一節、 資料蒐集

在資料的蒐集上，我們透過兩種不同且有計劃的方式，總共取得了龐大的資料集。這些資料集對於我們的研究過程至關重要，能夠提供深入洞察醫療領域的實體識別以及相關關鍵詞的資訊。

首先，我們著重介紹第一種資料收集方式。這種方式是使用了 Chinese Healthcare Named Entity Recognition Dataset，簡稱 HealthNER dataset，該資料集由 NCUEE NLP 研究室的專業人員所收集與標記[7]。這個資料庫中的資料是經過嚴謹的整理和標註，採用了 BIO 標註的方法，共涵蓋了 10 種重要的實體標籤，

分別為 Body、Symptom、Instrument、Examination、Chemical、Disease、Drug、Supplement、Treatment 和 Time。這些標籤反映了醫療領域中各種不同的實體，從身體部位到藥物、治療方式以及時間等等，提供了全面的資訊。

接下來，我們描述第二種資料蒐集方式。在這個方式中，我們使用了多個來源，有包括公開的醫療資料庫，像是 KingNet 國家網路醫藥網站以及網路上發表的各種醫學言論，或者有機化學的專業書籍。其中，我們利用 Python 的套件(eg. BeautifulSoup、Scrapy）來開發網路爬蟲，並且透過分析網頁結構和使用相應的爬蟲技巧，從網頁上來獲取這些醫療相關的文本資料，來提取我們所需的資料，另一方面，對於書籍處理方式則是以人工打至電腦為主。接著，我們從中專注於挑選那些包含常見或者冷門相關詞彙的句子，並將這些句子將成為我們後續人工標註的對象。為了擴充這些句子的數量以滿足我們的研究需求，我們可能會對它們進行串接或者連接擴展。最後，我們協調了團隊成員，讓他們負責不同部分的標註工作。這個過程中，我們依然採用了 BIO 標註的方法，保持了與第一種方式相同的 10 種標籤體系。

這兩種不同的資料收集方式為我們的研究提供了多樣性和全面性。通過這些方式，我們能夠充分利用現有資源，以及精心整理的資料集，從而更好地理解和分析醫療領域的實體識別和相關詞彙。其中，圖 2-1 展現研究方法之主要步驟。

第二節、 資料前處理

接在資料預處理的過程中，我們執行了一系列步驟，來確保文本數據的品質和可用性：

1.清理和標準化：我們利用 Python 的字串之操作函式與正則表達式，來進行資料的清理，其中包含移除特殊字元、標點符號，以及多餘的空格。同時，我們進行單詞的大小寫標準化，並進行拼寫修正等操作，以確保文本的一致性和準確性。

2.分割成句子：我們通過手動或利用自然語言處理工具，例如句子分割器，將文本分割成句子。這有助於更好地理解文本的結構和內容，使後續處理更容易進行[8]。

3.詞彙識別和標記：我們使用人工方式，即手動來標註文本中的詞彙，和使用已經標註好的資料集來對資料訓練前做標記。另外，我們也使用預先處理好的模型，如基於字典的模型，來進行詞彙識別和標記。通常使用的標記方式之一是 BIO（Begin, Inside, Outside）序列標註，它有助於識別實體和詞彙的邊界。

綜觀來說，這些資料預處理步驟是確保文本資料適合進一步分析和應用的關鍵過程。通過清理、結構化和標記，我們能夠建立高品質的資料集，以供機器學習和深度學習模型進行訓練和應用。這對於自然語言處理任務和醫療文本分析等領域具有重要意義。

第三節、 特徵提取並訓練模型

第一項、 訓練集與測試集之分配

我們將資料分為兩個階段使用，分別為挑選模型和完成網站。首先，在挑選模型階段，我們使用了 HealthNER 資料集，這是我們需要的資料，以獲得指標並選擇合適的模型，以供將來放入網站使用。訓練集和測試集的比例大約為 4:1。

至於第二部分，當我們挑選並確定了最佳的模型後，我們需要訓練一個更加完善的模型供使用者在登入網站時使用。因此，我們計劃將 HealthNER 資料集和我們自己手工標註的資料集結合在一起，不再區分訓練集和測試集，而是將它們全部用於模型的訓練。這些改變將有助於我們獲得更好的模型，以滿足不同階段的需求。

第二項、 獲取特徵(將文字轉成向量處理)

我們進行文本處理的步驟如下：首先，我們建立了一個專屬的醫療字典，解決詞性和句法架構，其中包含豐富的醫學術語和專業名詞，以確保文本的精確性和專業性。接著，我們將文字轉換為向量，這是透過 word2vec [9]等詞向量技術實現的，將單詞轉換成數值形式，以便計算機處理，其中包含上下文特徵的處理。最後，我們使用詞向量方法來提取文本的特徵，這有助於我們分析和理解文本內容，並應用在不同的醫療領域。

第三項、 不同模型之訓練暨選擇最佳模型

我們在不同模型之間進行訓練，並選擇最適合的模型，以用於放入網站來使用。首先，在機器學習方面，我們將使用決策樹[10]、隨機森林[11]和支持向量機[12]進行訓練，這些是傳統的機器學習模型。另一方面，在深度學習領域，我們將嘗試多種架構，包括 BERT [13]、Roberta [14]、BERT+LSTM+CRF [15]、Roberta+LSTM+CRF、BERT+Bi-LSTM+CRF [16]和 Roberta+Bi-LSTM+CRF 等模型，以便找到最適合我們特定任務的深度學習模型，其中有些使用比傳統 RNN [17]優秀的 Bi-LSTM [18]循環神經網路提升訓練結果。

為了評估這些模型的性能，我們將使用多個指標，其中含有:均方誤差（MSE）、決定係數（R^2）、F1 得分（FB1）和準確度（Accuracy）。這些指標將幫助我們瞭解並比較每模型之間的表現，並以此選擇最佳的模型，方便我們決定最終的模型需求。總結而言，我們通過訓練和評估多種機器學習和深度學習模型，以找到最佳模型，並使用不同指標來評估其性能，以確保我們之後的應用能夠達到最佳的效果。

第四節、 建立醫療詞彙標註系統之網站

最後，我們將模型建立成一個系統[19]，用於處理醫療資料，提供準確的詞彙標註服務。我們使用了 Django 框架和 AWS 雲端整合來開發一個強大的網頁應用，並提供使用者使用。在這個系統內，使用者透過網頁介面輸入需要標註的醫療資料，接著這些資料將被傳送到後端，然後再傳輸到醫療詞彙標註系統進行處理。這個系統將進行智能標註，並識別並標記出文本中的醫療詞彙，同時整理相關的醫療詞彙，提供全面的信息。

接著，處理結果會在網頁上呈現給使用者，這結果不但包括已標註的醫療詞彙，以及相關醫療詞彙的整理，而且我們還將提供下載相關醫療詞彙整理的提取按鈕，以方便使用者來選擇是否下載來做為進一步的研究和應用。

這個醫療詞彙標註系統將成為醫學領域的寶貴工具，其中有助於醫療專業人員和研究人員更有效地處理和分析醫療文本資料，並且因為它供高度自動化的解決方案，節省時間並提高準確性，因此將對醫學研究和臨床實踐產生積極影響。

第三章、 結果

第一節、 標註成果

第一項、 各個模型結果

首先，決策樹是一種樹狀結構，可用於分類任務，透過在每個節點進行分割，將資料劃分為子集直至達到終止條件。支持向量機則是用於二元分類的演算法，目的是找到一個超平面，能夠將不同類別的樣本分分開。而隨機森林則是結合多個決策樹進行分類，透過隨機抽樣和建立決策樹，最後綜合它們的預測結果來做出最終決策。最後我們討論機器學習中的各個模型的參數設定。

1. 決策樹方面，我們創建了一個隨機森林分類器，其中設置 50 棵樹，並且每棵樹的最大深度限制為 5。
2. 支持向量機（SVM）方面，我們使用了 Word2Vec 並將每個單字表示為 100 維度的向量，並考慮了上下文，使用了前後 5 個單字進行訓練。分類器方面，我們採用了高斯核函數（rbf），通常用於處理非線性分類問題。此外，我們採用了"one-vs-rest"策略來處理多類別分類問題。
3. 隨機森林分類器的設置中，我們指定了 1000 棵樹，每棵決策樹的最大深度為 40，並設定了隨機種子為 42，以確保每次運行模型時都獲得相同的結果。

接著 BERT，即 Transformer 的 Encoder。是谷歌以無監督方式利用大量無標記文本的方法訓練而成在深度學習領域方面的各個模型的參數設定，

Roberta 則是 BERT 針對識別任務上改進得模型，因此在 BERT 以及 Roberta 分別使用模型為(ckiplab/bert-base-chinese, 768-hidden)以及哈工大的(hfl/chinese-roberta-wwm-ext, 768-hidden)模型。在前兩種實驗上，我們在模型上增加多層感知機(MLP) 進行訓練。

1. BERT+MLP: 設定句子最大長度為訓練集中的最大長度句子進行實驗，在訓練方面，epoch 設定為 3，學習率設置為 2e-5，批次大小為 16，權重衰減設定為 0.01。
2. Roberta+MLP: 設定句子最大長度為訓練集中的最大長度句子進行實驗，在訓練方面，epoch 設定為 3，學習率設置為 2e-5，批次大小為 16，權重衰減設定為 0.01。

最後，我們在後兩種模型分別增加 Bi-LSTM 以及 CRF，其中 LSTM（長短期記憶）是一種循環神經網絡，可以有效地捕捉長距離相依性，並且不容易受到梯度消失或梯度爆炸的問題影響。其中包含記憶單元（cell）和三個門（input gate、output gate、forget gate）。而 Bi-LSTM 使用兩個獨立的 LSTM 結構，一個從前向序列開始，另一個從後向序列開始。這兩個方向的 LSTM 都有自己的隱藏狀態，並且能夠捕捉前向和後向信息。而 CRF
條件隨機場，是一種概率圖模型，用於序列標記和結構預測問題。它可以建模序列中不同標記之間的相依性，並且可以考慮全局序列上下文，而不僅僅是局部信息。

1. BERT+Bi-LSTM+CRF: 設定句子最大長度為 150 字，在訓練方面，epoch 設定為 12，學習率設置為 0.012，批次大小為 16，權重衰減設定為 1e-5，最後梯度下降為隨機梯度下降。
2. Roberta+Bi-LSTM+CRF: 設定句子最大長度為 150 字，在訓練方面，epoch 設定為 12，學習率設置為 0.012，批次大小為 16，權重衰減設定為 1e-5，最後梯度下降為隨機梯度下降。

第二項、 分析並選擇最佳模型

Models	指標	
	MSE	R^2
Random Forest	4.676	**-0.03**
Decision Tree	**3.758**	-2.952
	F1	Accuracy
SVM	78.759	51.165
Bert+MLP	**79.49**	92.05
Roberta+MLP	77.46	**93.08**
Bert-BiLSTM-CRF	75.9	86.1
Roberta-BiLSTM-CRF	77.57	92.76
Bert-LSTM-CRF	75.58	89.81

Roberta-LSTM-CRF	74.83	89.91

表 3-1-2 模型結果

首先,指標以 MSE 以及 R^2 作為評估的隨機森林以及決策樹進行評估,隨機森林的 MSE 為 4.676,決策樹的 MSE 為 3.758。在這兩種模型中,低的 MSE 表示模型的迴歸表現較好,而決策樹的 MSE 較低,可能在迴歸任務上表現上較佳。在 R^2 方面,通常情況下,R^2 的取值範圍在 0 到 1 之間,數值越接近 1 表示模型的擬合效果越好。然而,隨機森林的 R^2 為-0.03,決策樹的 R^2 為-2.952。這兩個值都非常接近 0,表示模型對資料的擬合效果不佳。因此,可以得出結論,隨機森林和決策樹模型不適合用於解決此問題。接下來,關於深度學習模型,評估指標主要使用 F1 分數和準確度。 F1 分數綜合考慮了模型的精確性召回率,而準確率只關注正確分類的樣本數。Bert+MLP 和 Roberta+MLP 模型在 F1 分數和準確率方面表現良好,顯示它們在分類任務上的表現較好。而在在 F1 分數和準確率來看,Bert+MLP 模型的 F1 較高,而 Roberta+MLP 的準確率較高,而在我們所需要的結果來說,我們需要的是 F1 分數的精確性召回率,因此我們選擇 Bert+MLP 作為我們的主要模型進行訓練。表 3-1-2 展現了模型結果。

第二節、 網站呈現

第一項、 技術面背後呈現方式

1. 從使用者的 input data,透過 Transformer Tokenizer 進行分詞並轉成 Tensor 格式
2. 設定函數檢查 input data 是否超過 BERT 所限制的 512 Token:
   - 2.1. 若超過,對於輸入的文本進行切割,依照 Tensor 中所含的固定鍵值(“input_ids”、 ”attention_mask”、 ”token_type_ids”)三者進行拆分。因三個鍵值的資料是一一對照,利用串列切片方式批次選取相同數量並丟入 pre-trained 完成的 BERT 進行自動化標註。
   - 2.2. 若 input data 無超過 BERT Token limit,則回傳原來的 inputs,不需進行批次處理。
3. 利用批次方式:禁用梯度下降,而是讓模型進行批次推理方式針對數據進行標註,最後得到 logits,即是模型對每個標註(類別)的原始預測分數。
4. 使用 argmax 來選擇每個位置的預測類別:對於每個 token 或位置,在模型輸出的所有可能類別的 logits 中選擇最大的那個。
5. 得到最終預測:每個位置的預測類別將組成一個預測序列,即可得到最終標記結果。
6. 利用最終標記結果,實現”單詞標註-結果”、“原文標註-結果”、” 醫學相關詞彙檢視並下載”。
7. 圖 3-2-1 顯示了使用者在使用網站的時候,會呈現出來的畫面。

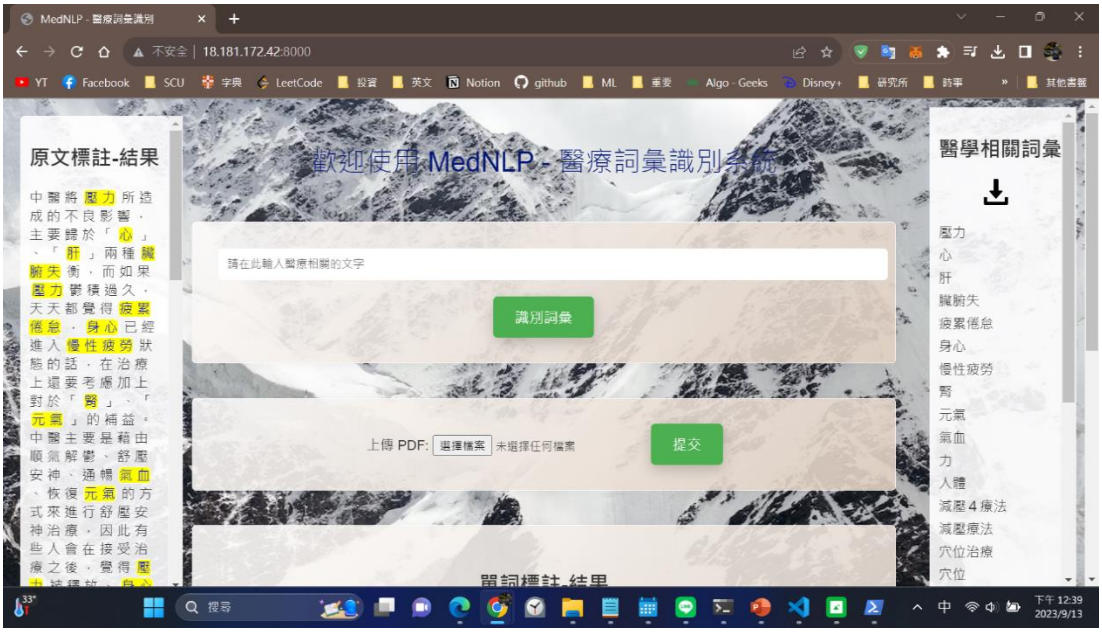


圖 3-2-1. 網站呈現圖

第二項、 使用步驟

1. 使用者於文字框輸入醫學文章或透過上傳醫學相關文獻 PDF 檔
2. 點選識別詞彙或提交後，透過訓練完的 AI 模型，即可產出 BIO 標註完的結果

第三項、 網站最終產出結果

1. **於網站下方提供(1).單詞標註-結果:**
   可顯示出各單詞對應的 BIO 詞性，一共有 10 種 Entity Type，分別為 Body(人體)、Symptom(症狀)、Instrument(醫療器材)、Examination(檢驗)、Chemical(化學物質)、Disease(疾病)、Drug(藥品)、Supplement(營養品)、Treatment(治療)、Time(時間)。圖 3-2-2 表現出其標註的結果，會出現在字的右下角。
2. **並將名詞片語(Noun Phrase, NP)，透過 BIO 的三個標記為：**（1）B-NP：名詞片語的開頭 （2）I-NP：名詞片語的中間 （3）O：不是名詞片語，而標記結果為'O'我們選擇不顯示出來

圖 3-2-2. 標註結果

3. **於網站左側提供(2).原文標註-結果:**
   2.1 顯示使用者所輸入的完整全文，並透過 BIO 將其 AI 模型所辨識出的 BIO 醫療相關詞彙標註結果用黃色凸顯出來，並透過標記位置的連續性可抓出完整多個詞彙的專業醫學名詞。

2.2. 點選標記出的黃色底醫學詞彙，即可將該詞彙連結至 google 搜尋引擎，直接查找相關醫學資訊。

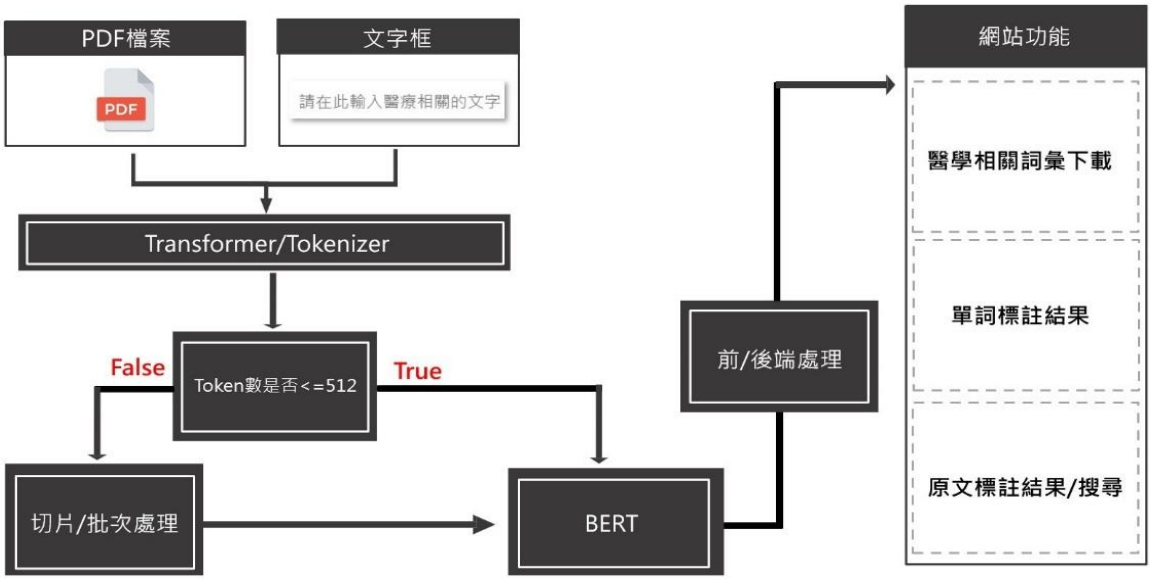
4. 於網站右側提供(3).醫學相關詞彙:

3.1. 列出所有文章中所出現的醫學詞彙

3.2. 點選下載圖標，即可提供一鍵式下載所有標記出醫學詞彙

5. 圖 3-2-3.表示出完整的流程圖:

圖 3-2-3. 流程圖



第四章、 結論

第一節、 結論

在本研究中，我們深入探討了在廣泛醫學領域中使用自然語言處理技術建立醫療詞彙標註系統的重要性和應用價值。醫學領域充斥著大量的文獻和臨床數據，其中包含豐富的醫學專業詞彙，這些詞彙對於醫學專業人員的臨床決策和研究至關重要。然而傳統的人工處理方法耗時且容易出錯，因此我們提出了建立醫療詞彙標註系統的解決方案。

此系統利用機器學習和深度學習技術，能夠快速、準確地標註醫學專業詞彙，從而節省時間並提高正確性。我們設計了多種詞彙類別，包括身體、症狀、儀器、檢查、化學、疾病、藥物、補充、治療和時間，以滿足不同醫學信息處理需求。另外，我們強調了使用者友好性和隱私安全的重要性，確保系統能夠被廣泛應用並保護醫學資訊的機密性。

其中，這項研究的終結將是一個具有巨大潛力的工具，它有望加速醫學信息

的處理和應用，並為臨床決策、研究和醫療領域的發展貢獻一份力量。未來，我們期待這個系統能夠不斷優化和擴展，以滿足不斷增長的醫學信息需求，並為改善健康保健提供更多支援。

最後，這個研究讓我們深刻體會到科學和技術的潛力，以及如何應用它們來解決現實世界中的問題。我們的工作也是對醫學領域的一種貢獻，希望它能夠改善醫療服務，提高醫學研究的效率，並最終造福人類的健康。

第二節、 未來的應用

醫療詞彙標註系統網站具有廣泛的應用前景，且在當今信息過載的時代，這項技術可以用於管理和分類大量的臨床文檔，包括病歷、檢查報告和處方等。這項技術的一項關鍵應用是自動識別診斷、治療方案和患者數據，這使得醫護人員更容易找到所需的信息，協助醫生進行診斷、制定治療計劃，並預測疾病風險。這不僅節省了時間，還提高了工作效率，有助於提供更好的醫療服務。並且同時，在研究領域，醫療詞彙標註系統的應用也是不可忽視的。它可以自動標註和分類資料中的關鍵詞，這有助於支持關聯性分析和元分析，從而推動醫學知識的不斷進步。

當然，醫療詞彙標註系統網站還有更多令人期待的未來發展。其中一個主要的發展方向是智能問答系統。例如，像最近非常熱門的 ChatGPT 這樣的系統，能夠回答使用者的疑問，並提供個性化的建議。當使用者提出與醫療相關的問題時，這些系統可以從豐富的資料庫中即時提供答案。此外，它們還可以根據使用者提供的病歷和症狀信息生成處方建議，或協助醫生更迅速地制定治療計劃。

另一個值得關注的未來趨勢是視覺化數據分析的融合。這一趨勢將有助於提高數據的可解釋性和可視化，使用戶能夠更深入地理解文本數據中的命名實體。通過將醫療詞彙標註系統識別的命名實體與圖表、圖形或儀表板中的數據相結合，用戶可以更輕鬆地分析實體之間的關係和趨勢，進而做出更明智的決策。考慮到語言多樣性，多語言支持也將成為這一趨勢的關鍵因素，以確保在全球范圍內廣泛應用這項技術。這將有助於跨越語言和文化障礙，實現更廣泛的醫療信息共享和合作。

總而言之，醫療詞彙標註系統網站在當今和未來都有著巨大的應用潛力，將不斷推動醫療領域的進步和改善醫療服務的質量。它們將成為醫療領域中不可或缺的工具，為醫護人員和研究人員提供寶貴的支持和資源。

參考文獻

[1] Hou, J. K., Imler, T. D., & Imperiale, T. F. (2014). Current and future applications of natural language processing in the field of digestive diseases. Clinical

Gastroenterology and Hepatology, 12(8), 1257-1261.
[2] Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. Journal of the American Medical Informatics Association, 26(4), 364-379.
[3] Öztürk, H., Özgür, A., Schwaller, P., Laino, T., & Ozkirimli, E. (2020). Exploring chemical space using natural language processing methodologies for drug discovery. Drug Discovery Today, *25*(4), 689-705.
[4] Mellia, J. A., Basta, M. N., Toyoda, Y., Othman, S., Elfanagely, O., Morris, M. P., ... & Fischer, J. P. (2021). Natural language processing in surgery: a systematic review and meta-analysis. Annals of Surgery, *273*(5), 900-908.
[5] Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. Journal of biomedical informatics, *42*(5), 760-772.
[6] Iroju, O. G., & Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. International Journal of Information Technology and Computer Science, *8*, 44-50.
[7] Lee, L. H., & Lu, Y. (2021). Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. IEEE Journal of Biomedical and Health Informatics, 25(7), 2801-2810.
[8] Kang, A., Ren, L., Hua, C., Song, H., Dong, M., Fang, Z., & Zhu, M. (2021). Environmental management strategy in response to COVID-19 in China: Based on text mining of government open information. Science of the Total Environment, 769, 145158.
[9] Chen, Q., & Sokolova, M. (2021). Specialists, scientists, and sentiments: Word2Vec and Doc2Vec in analysis of scientific and medical texts. SN Computer Science, *2*, 1-11.
[10] Xue, D., Frisch, A., & He, D. (2019). Differential diagnosis of heart disease in emergency departments using decision tree and medical knowledge. In Heterogeneous Data Management*,* Polystores*,* and Analytics for Healthcare*:* VLDB *2019* Workshops*,* Poly and DMAH*,* Los Angeles*,* CA*,* USA*,* August *30, 2019,* Revised Selected Papers *5* (pp. 225-236). Springer International Publishing.
[11] Amato, F., Coppolino, L., Cozzolino, G., Mazzeo, G., Moscato, F., & Nardone, R. (2021). Enhancing random forest classification with NLP in DAMEH: A system for DAta Management in eHealth Domain. Neurocomputing, *444*, 79-91.
[12] Faris, H., Habib, M., Faris, M., Alomari, M., & Alomari, A. (2020). Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines. Journal of biomedical informatics, *109*, 103525.

[13] Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D., & Jawahar, C. V. (2021, April). Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021* IEEE *18*th International Symposium on Biomedical Imaging *(*ISBI*)* (pp. 1033-1036). IEEE.
[14] Yang, X., Bian, J., Hogan, W. R., & Wu, Y. (2020). Clinical concept extraction using transformers. Journal of the American Medical Informatics Association, *27*(12), 1935-1942.
[15] Dong, X., Chowdhury, S., Qian, L., Li, X., Guan, Y., Yang, J., & Yu, Q. (2019). Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN. PloS one, *14*(5), e0216046.
[16] Qin, Y., & Zeng, Y. (2018). Research of clinical named entity recognition based on Bi-LSTM-CRF. Journal of Shanghai Jiaotong University *(*Science*)*, *23*, 392-397.
[17] Zhou, X., Li, Y., & Liang, W. (2020). CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE/ACM Transactions on Computational Biology and Bioinformatics, *18*(3), 912-921.
[18] Bohnet, B., McDonald, R., Simoes, G., Andor, D., Pitler, E., & Maynez, J. (2018). Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. arXiv preprint arXiv*:1805.08237*.
[19] Bennetot, A., Donadello, I., Qadi, A. E., Dragoni, M., Frossard, T., Wagner, B., ... & Díaz-Rodríguez, N. (2021). A practical tutorial on explainable ai techniques. arXiv preprint arXiv*:2111.14260*.

組員工作表

蘇睿詮: 20% (組長)
張升祥: 20%
高麒祥: 20%
蘇子恩: 20%
林羿帆: 20%