Bachelor Degree Program School of Big Data Management Soochow University
Report of practices

指導老師：蘇明祥 教授

巨資四A 09170141 蘇睿詮（組長）
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

自動化醫學詞彙檢索系統

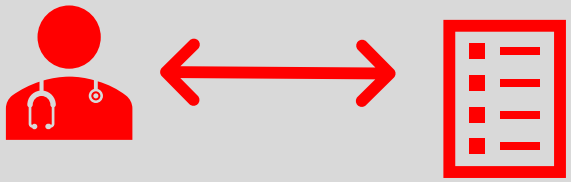
東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

資料科學系
Department of Data Science

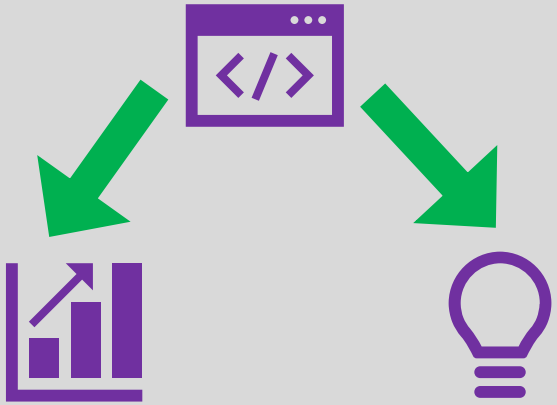
專題成果

本專題針對傳統人工標註方法缺點，提出使用機器學習和深度學習技術，對醫學文本進行醫學詞彙標註，以提高醫學詞彙辨識能力。

在醫學領域中，應用了自然語言處理技術(NLP)來建立醫療詞彙標註系統。



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，自動化的詞彙標註系統可以快速、準確地從文本提取有價值的醫學信息。



使用所獲得的模型，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。

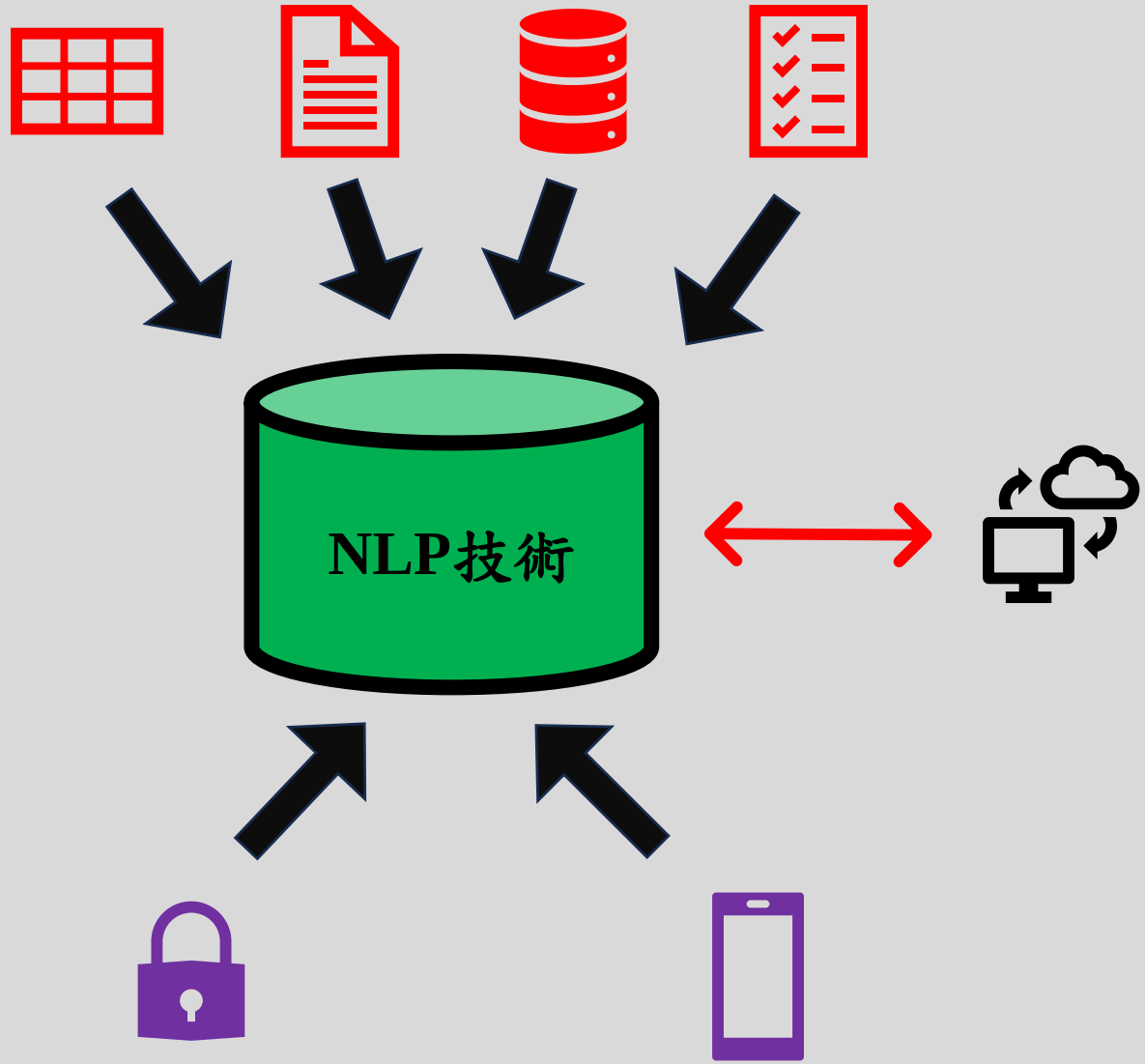
東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

Data Science
資料科學系
Department of Data Science

醫學資訊

其中文獻、臨床記錄和病歷等，蘊含大量醫學專業詞彙。因此建立一個自然語言處理（NLP）技術的醫療詞彙標註系統。應包含高效性、時間節省、準確性、正確辨識、和多種詞彙類別識別。

1. 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。
2. 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



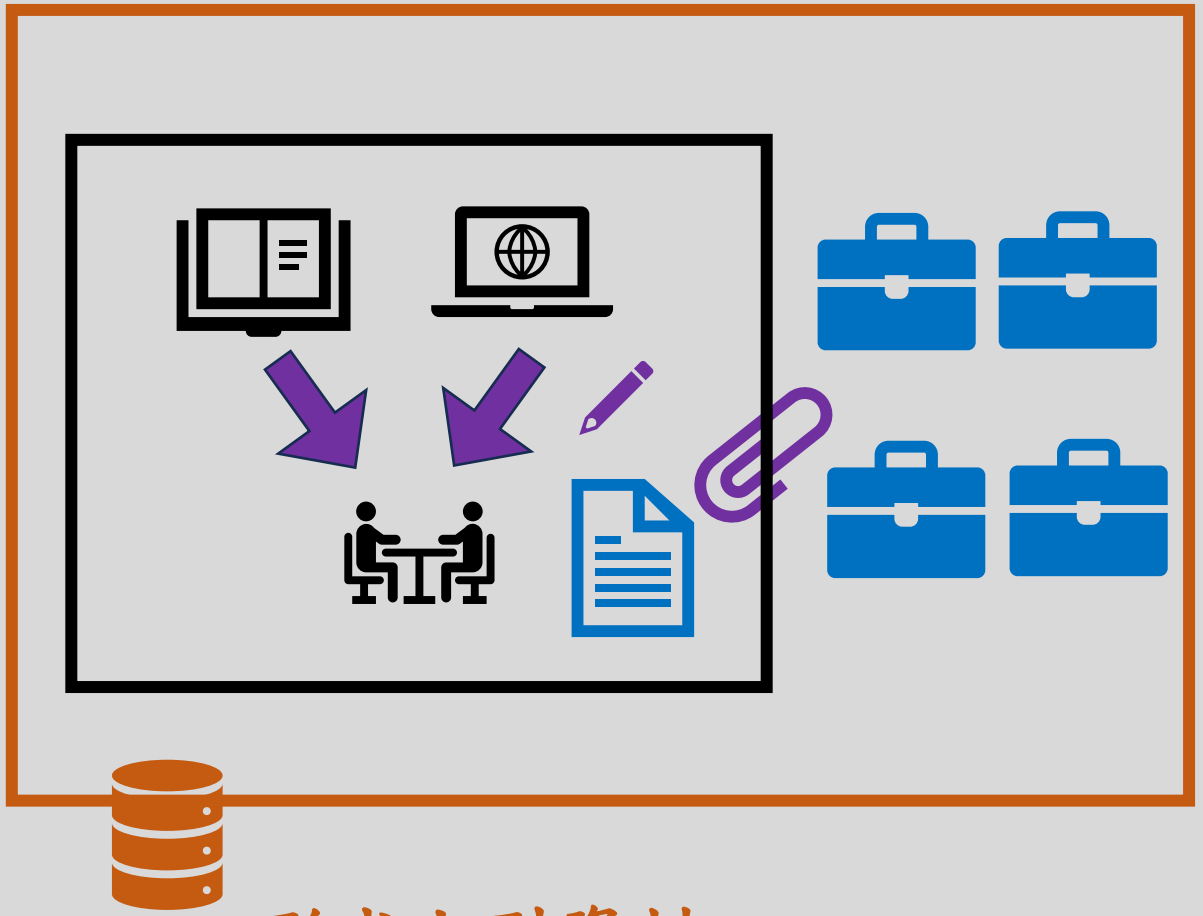
資料蒐集

1. 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
2. 共涵蓋了10種重要的實體標籤
3. 標籤: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
4. 公開的醫療資料庫，像是KingNet國家網路醫藥網站、網路上醫學言論、有機化學的專業書籍
5. 利用Python的套件(eg. BeautifulSoup、Scrapy）來開發網路爬蟲獲取相關醫療文本資料
6. 書籍:人工打至電腦為主。

人工標註的對象:專注挑選那些包含常見或者冷門相關詞彙的句子。

擴充句子:串接或者連接擴展。

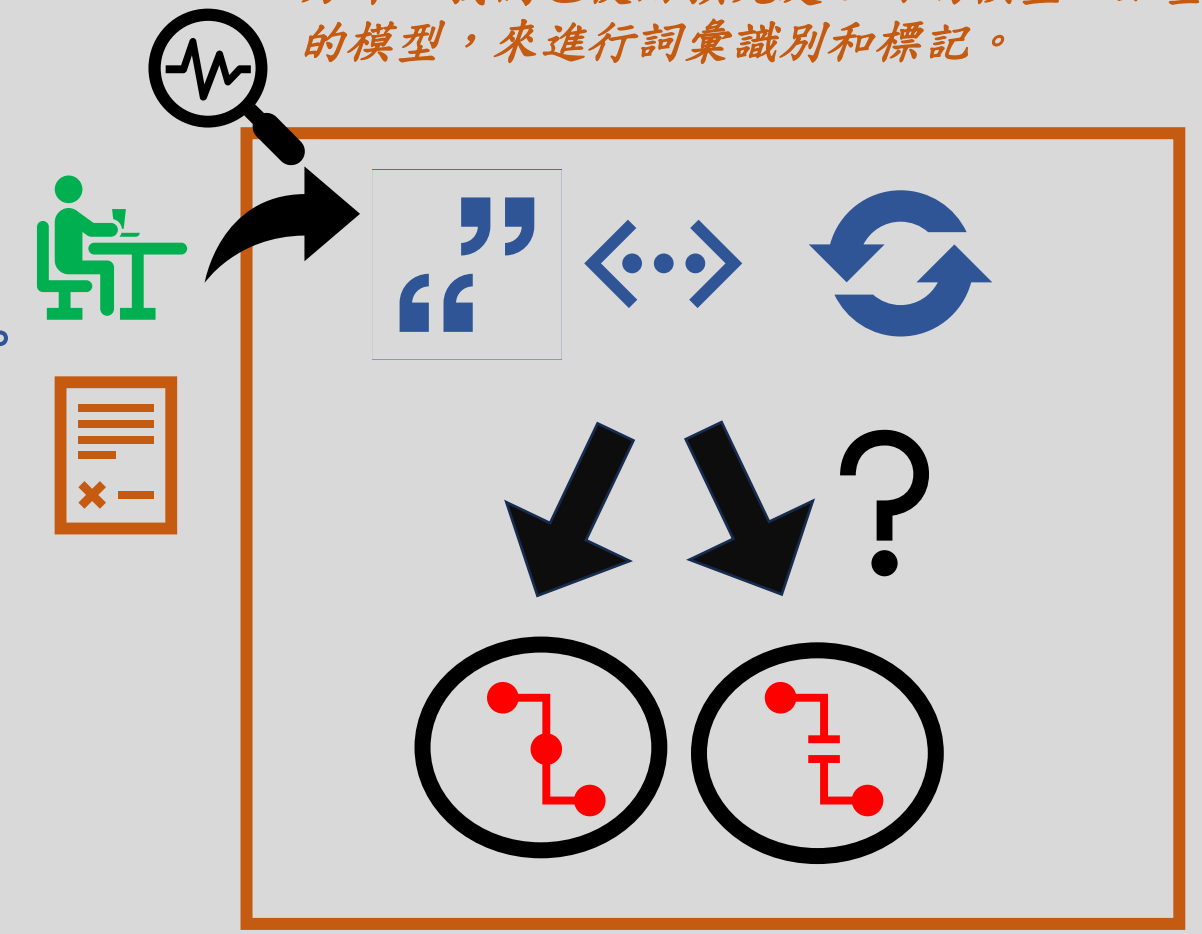
(採用了BIO標註的方法相同的10種標籤體系)



形成大型資料

資料前處理

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。

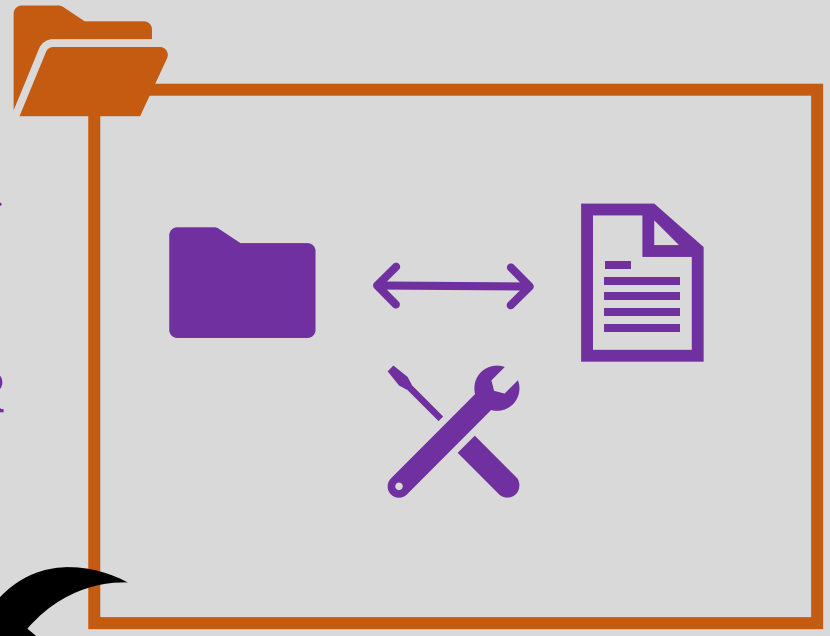


1. 清理和標準化：使用Python的函式、正則表達式，來進行資料的清理，(移除特殊字元、標點符號、多餘的空格)。同時我們對單詞的大小寫標準化、進行拼寫修正等操作。
2. 分割成句子：手動或自然語言處理工具，助於理解文本的結構、內容。(句子分割器)
3. 詞彙識別和標記：手動來標註文本中的詞彙，並使用已經標註好的資料集的格式。對資料訓練前做標記。使用標記方式是BIO (Begin, Inside, Outside) 序列標註，它有助於識別實體和詞彙的邊界。

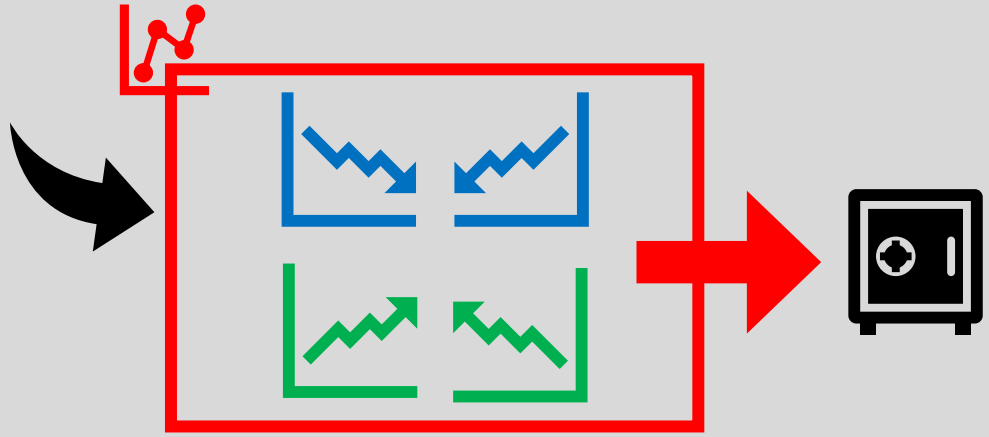
訓練集與測試集之分配

挑選模型(第一階段):使用了HealthNER資料集，以並獲得每個模型之指標，並選擇合適的模型。訓練集和測試集的比例4:1。

完成網站(第二階段):確定了最佳的模型後，將HealthNER資料集和我們自己手工標註的資料集結合在一起，全部用於模型的訓練。



獲取特徵
(將文字轉成向量處理)



獲取特徵(將文字轉成向量處理)

文本處理:建立了一個專屬的醫療字典，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。

文字轉向量:透過word2vec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。

東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

Data Science
資料科學系
Department of Data Science

訓練暨選擇最佳模型

不同模型進行訓練，選擇最適合的模型，以用於放入網站來使用

機器學習

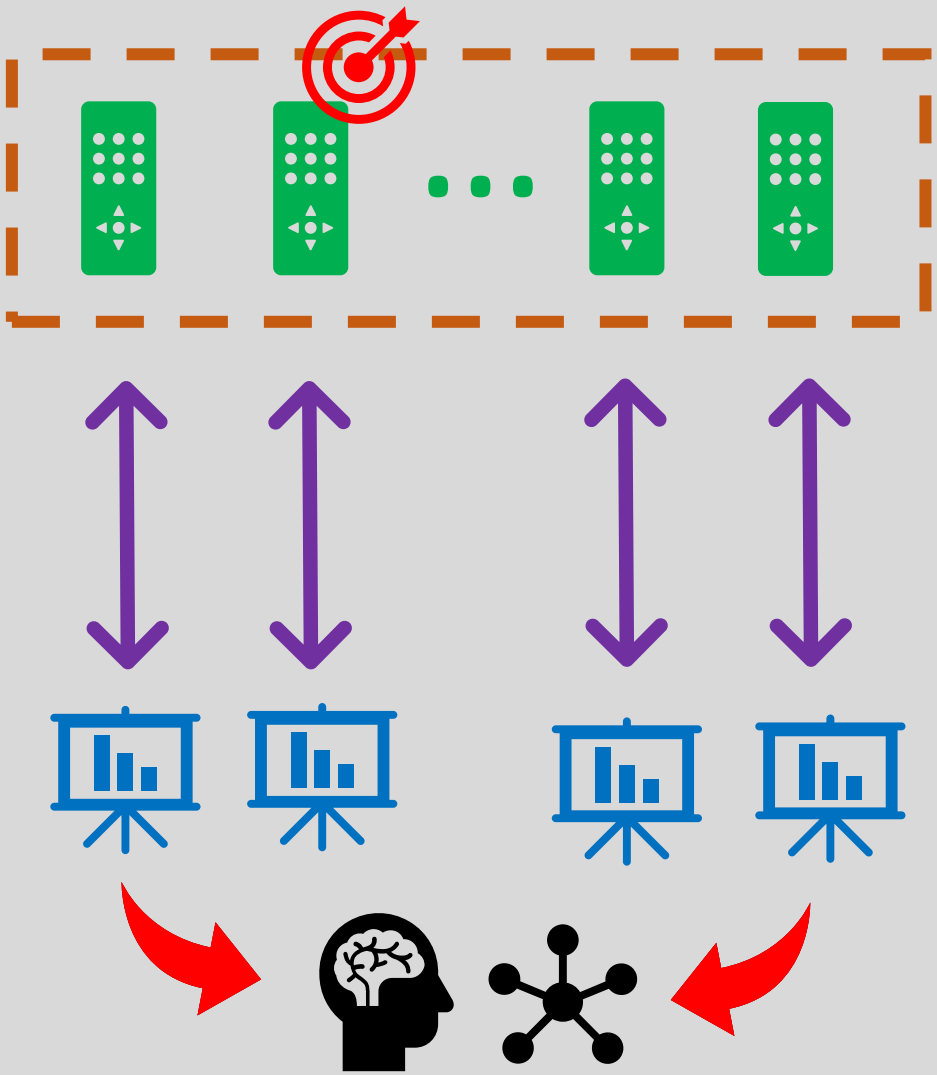
決策樹、隨機森林、支持向量機

深度學習

1. BERT
2. RoBERTa
3. BERT+LSTM+CRF
4. RoBERTa+LSTM+CRF
5. BERT+Bi-LSTM+CRF
6. RoBERTa+Bi-LSTM+CRF

評估指標: 均方誤差（MSE）、決定係數（R^2）、F1得分（FB1）和準確度（Accuracy）

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。



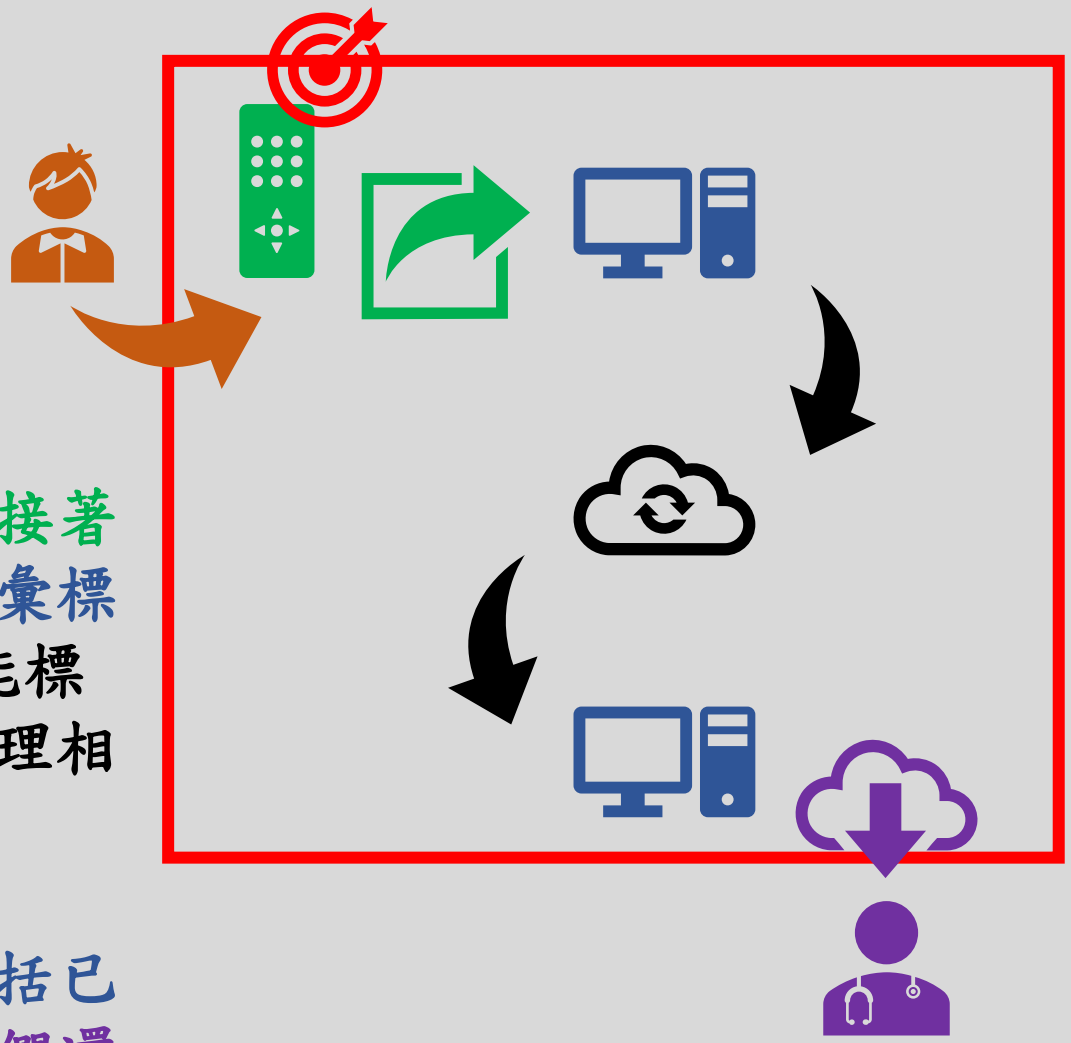
醫療詞彙標註系統網站

使用兩個架構整合開發一個網頁:

1. Django框架
2. AWS雲端

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。這個系統將再AWS雲端進行智能標註，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。

最後，處理結果會在網頁上呈現給使用者，有包括已標註的醫療詞彙，以及相關醫療詞彙的整理，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。



東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

Data Science
資料科學系
Department of Data Science

參數設定

1. 決策樹 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 Word2Vec每個單字表示為100維度的向量，前後5個單字進行訓練。分類器方面用高斯核函數，此外採用了"one-vs-rest"策略來處理多類別分類問題。
3. 隨機森林 指定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，確保每次運行時都得相同的結果。
4. BERT+MLP 句子最大長度為訓練集中的最大長度句子，epoch設定為3，學習率為2e-5，批次大小為16，權重衰減為 0.01。
5. Roberta+MLP 句子最大長度為訓練集中的最大長度句子，epoch設定為3，學習率為2e-5，批次大小為16，權重衰減為 0.01。

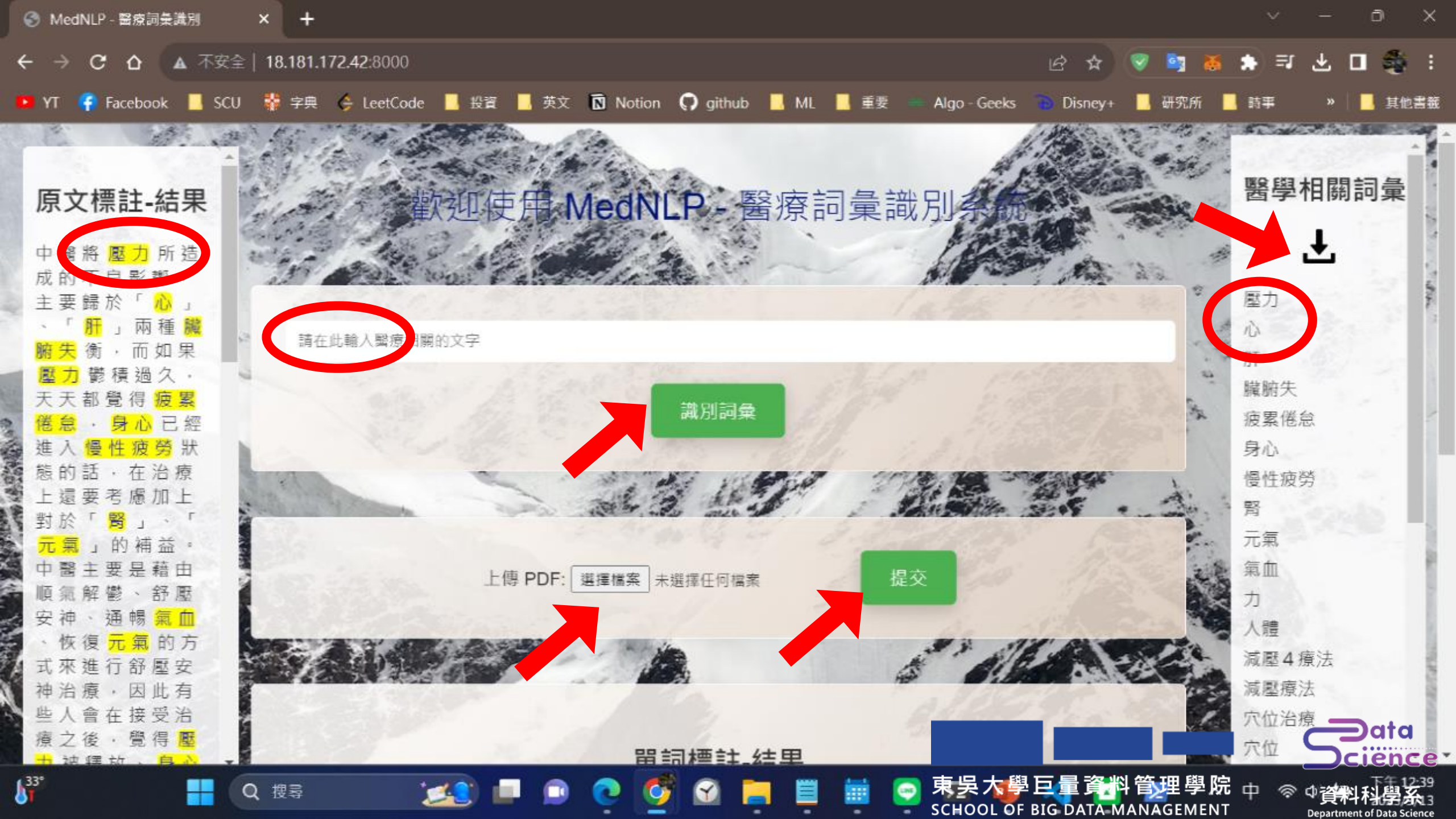
另外，在後兩種模型分別增加Bi-LSTM、CRF

6. BERT+Bi-LSTM+CRF，Roberta+Bi-LSTM+CRF

句子最大長度為150字，epoch設定為12，學習率為0.012，批次大小為16，權重衰減為1e-5，最後梯度下降為隨機梯度下降。

模型指標

	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


MedNLP - 醫療詞彙識別
18.181.172.42:8000
YT
Facebook
SCU
字典
LeetCode
投資
英文
Notion
github
ML
重要
Algo - Geeks
Disney+
研究所
時事
其他書籤
原文標註-結果
中醫將壓力所造成的
主要歸於「心」、「肝」兩種臟腑失衡，而如果壓力鬱積過久，天天都覺得疲累倦怠，身心已經進入慢性疲勞狀態的話，在治療上還要考慮加上對於「腎」、「元氣」的補益。中醫主要是藉由順氣解鬱、舒壓安神、通暢氣血、恢復元氣的方式來進行舒壓安神治療，因此有些人會在接受治療之後，覺得壓
歡迎使用 MedNLP - 醫療詞彙識別系統
請在此輸入醫療相關的文字
識別詞彙
上傳 PDF:
選擇檔案
未選擇任何檔案
提交
醫學相關詞彙
壓力
心
肝
臟腑失
疲累倦怠
身心
慢性疲勞
腎
元氣
氣血
力
人體
減壓4療法
減壓療法
穴位治療
穴位
Data Science
東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT
資料科學系
Department of Data Science

使用者操作步驟

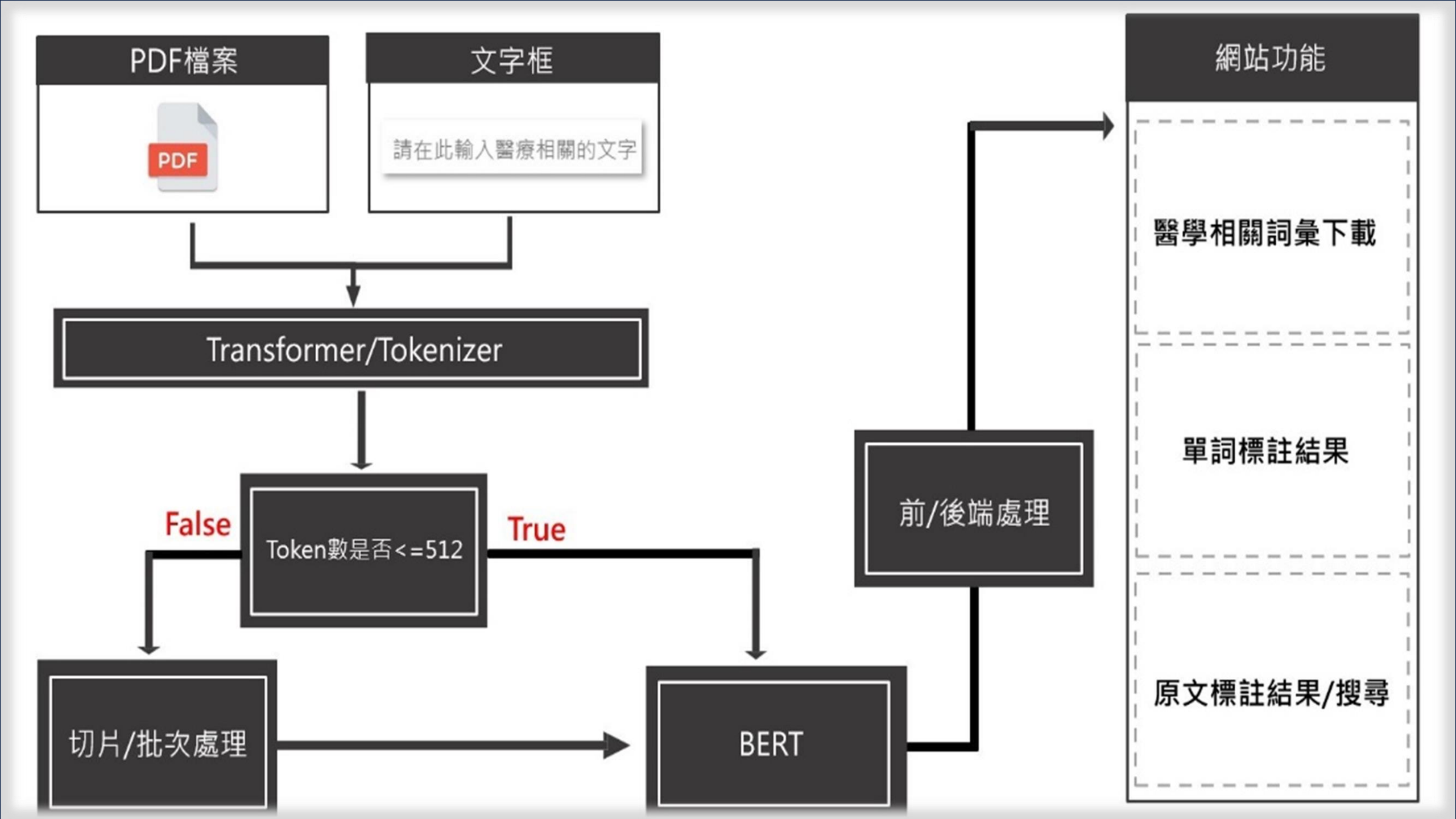
1. 使用者於文字框輸入醫學文章or上傳醫學相關文獻PDF檔
2. 點選識別詞彙或提交後，透過訓練完的AI模型，即可產出BIO標註完的結果

網站最終產出結果

1. 顯示使用者所輸入的完整全文
2. 辨識出的BIO醫療相關詞彙標註結果、黃色凸
3. 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. 列出所有文章中所出現的醫學詞彙，點選即可下載所有標記出醫學詞彙

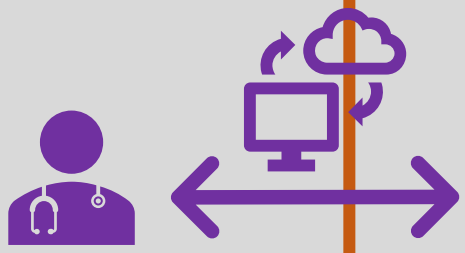
東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

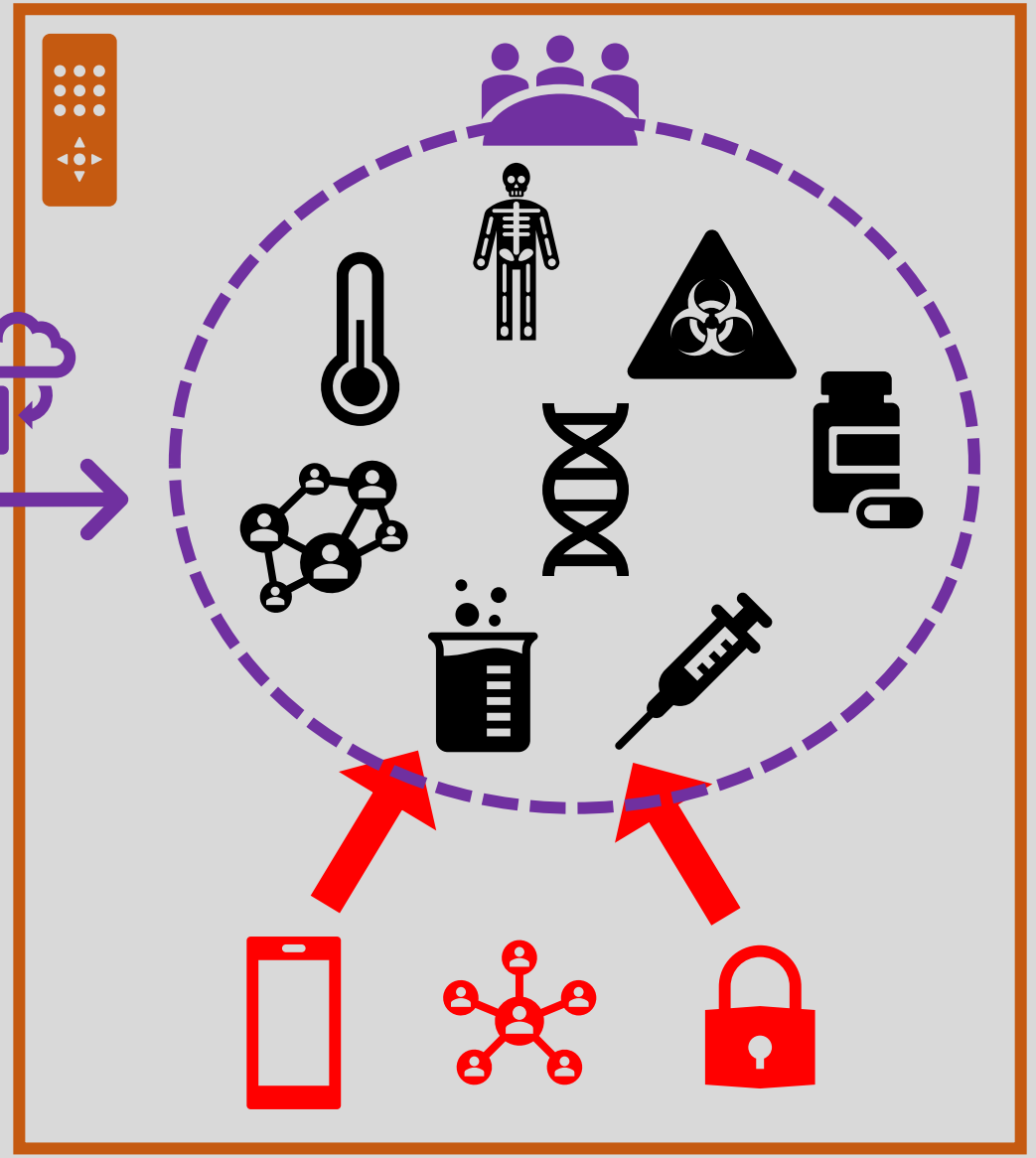
Data Science
資料科學系
Department of Data Science


PDF檔案
PDF
文字框
請在此輸入醫療相關的文字
Transformer/Tokenizer
Token數是否<=512
False
True
切片/批次處理
BERT
前/後端處理
網站功能
醫學相關詞彙下載
單詞標註結果
原文標註結果/搜尋

研究目的

在本研究中，我們強調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。利用機器學習和深度學習技術，實現快速、準確的詞彙標註，提高醫學專業人員臨床決策和研究效率。系統包括多種詞彙類別，同時注重使用者友好性和隱私安全，以廣泛應用並保護醫學資訊的機密性。





利用機器學習和深度學習技術

多種詞彙類別:

身體 症狀 儀器 檢查 化學 疾病
藥物 補充 治療 時間

東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

Data Science
資料科學系
Department of Data Science

總結

醫療詞彙標註系統網站可以用於管理和分類大量的臨床文檔(病歷、檢查報告、處方等)。

在當今信息過載的時代，這項技術的關鍵

1. 自動識別診斷
2. 治療方案
3. 患者數據

讓醫護人員容易找到所需之信息，協助進行診斷、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

東吳大學巨量資料管理學院
SCHOOL OF BIG DATA MANAGEMENT

Data Science
資料科學系
Department of Data Science