Final Project for the Biomedical Information Retrieval Course

資工所 P78101514 黃仁鴻

https://github.com/toonnyy8/ncku/tree/bir2021/final

系統環境

程式語言		JavaScript, Python 3
開發環境		Node.js 16.10.0、esbuild 0.13.3、ts-node 10.4.0
函式庫	前端	Vue 3.2.19
	文本處理	spacy 3.2.0
	xml 解析	xml2js 0.4.23
	深度學習	pytorch 1.10.0+cu102

資料集與前處理

本作業使用 LitCovid 上與病毒變異株相關研究論文共 5206 篇的摘要做為 CBOW 的訓練資料集，並從中抽選各 2500 篇進行展示，並進行變異株之間的相似度探討。

在訓練 CBOW 前會先透過 spacy 進行 Lemmatization、Tokenization 後轉換為小寫，並去除出現次數小於 5 的 token。詳細 CBOW 參數於表一。

Embedding dim	300
Number of context	6+6=12
Vocab size	7966
Min frequency	5

表 1 CBOW 參數

相似度加權

本次使用 TF-IDF 作為權重，配合 CBOW 訓練所得的 word embedding 計算出 document embedding 後，與兩個 Topic 計算相似度，並可藉由相似度計算兩者相關性。

相似度探討

在比對過文檔相關性，反而發現變異株之間出現時間越接近相關性越高，反而未在症狀上呈現出相似度。在實作過後發現，LitCovid 的文章中以變異株為主題的研究摘要與症狀為主題的研究摘要，兩者間內容相關性低，因此難以用症狀的相似性來進行變異株的區別

	b.1.1.529	b.1.617.2	p.1	b.1.351	b.1.1.7	original
b.1.1.529		0.57	0.53	0.5	0.48	0.3
b.1.617.2			0.72	0.65	0.52	0.32
p.1				0.59	0.71	0.34
b.1.351					0.54	0.37
b.1.1.7						0.38

表 2 相關性