Project 3 for the Biomedical Information Retrieval Course

資工所 P78101514 黃仁鴻

https://github.com/toonnyy8/ncku/tree/bir2021/hw3

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	NLTK 3.5、torchtext 0.11.0、spacy 2.3.5
	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 PubMed 上的10000篇躁鬱症相關研究論文的摘要做為資料集，在進行 word2vec 的訓練前，會先將資料全數轉為小寫後通過 spacy tokenizer 與 NLTK 的詞性分析及 WordNetLemmatizer 執行前處理。

CBOW 參數

Hidden layer dim	256
Embedding dim	128
Number of Context	4+4=8
Vocab size	5436
Optimizer	Adamax, lr=0.01
Epoch、batch size	100、256

相似度探討

常常會連接使用的 bipolar disorder，利用 cosine similarity 計算得出兩者之間的相似度並不高，而 disorder 與 illness、schizophrenia 及 disease 則有著極高的相關性。另外像是 hamd、ymrs、madrs 等等評估量表之間與 adhd（注意力不足過動症）、mdd（重性憂鬱疾患）這些心理疾病也具有高度相關性。由此可以發現，藉由 CBOW 的方式得到的 word embedding 確實能在一定程度上表現文字的意義。