Project 4 for the Biomedical Information Retrieval Course

資工所 P78101514 黃仁鴻

https://github.com/toonnyy8/ncku/tree/bir2021/hw4

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 PubMed 上的躁鬱症與新冠肺炎相關研究論文各 1000 篇的摘要做為 CBOW 的訓練資料集，並從中抽選各 100 篇進行展示。

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 3 的 token。詳細 CBOW 參數於表一。

Hidden layer dim	256
Embedding dim	64
Number of context	4+4=8
Vocab size	4848
Min frequency	3

表 1 CBOW 參數

相似度加權

本次使用了基於文檔與基於句子之 TF-IDF、BM25+ 共四種方法得出的分數作為權重，配合 CBOW 訓練所得的 word embedding 計算出 sentence embedding 後，才與 query 計算相似度後排序。

BM25+

$$\ln\left(\frac{N-\left|\{j:t_i\in d_j\}\right|+0.5}{\left|\{j:t_i\in d_j\}\right|+0.5}+1\right)\cdot\frac{tf_{i,j}\cdot(k_1+1)}{tf_{i,j}+k_1\cdot\left(1-b+b\cdot\frac{\left|D_j\right|}{avgdl}\right)}$$

N 為文檔數、$\left|D_j\right|$ 是第 j 個文件的 token 數、avgdl 表示平均文件長度，$k_1=2$、$b=0.75$。

相似度探討

一般來說，BM25+的相關性得分應該會比 TF-IDF 更好，但可能是因為搭配了 word embedding 以及文檔長度較短，兩者之間並沒有明顯的優劣。另外，基於「文檔」的相似度評分因為包含超過句子內部的資訊，具有更好的效果。