

Project #2 for the Biomedical Information Retrieval Course

1. Project Overview – word embedding

詞嵌入 (Word Embedding) 是自然語言處理的重要技術，透過神經網路將詞彙轉換為低維度的稠密向量，使語意或上下文相近的詞彙有機會在向量空間中呈現較高的相似度。本專案要求學生使用約 1,000 篇英文科學論文摘要，實作並比較兩種經典詞嵌入架構：

CBOW (Continuous Bag-of-Words)：利用目標詞周圍的上下文，預測目標詞。

Skip-Gram：利用目標詞，預測其周圍的上下文詞彙。

完成模型訓練後，學生須使用餘弦相似度分析詞彙關係，並利用如主成分分析 (PCA) 及 t-分布鄰域嵌入 (t-SNE)，將高維詞向量降至二維，進行詞嵌入視覺化與語意分析。

Word embeddings provide a way to represent words as dense vectors in a continuous vector space.

Instead of representing a word simply as an independent symbol:

"cancer" "tumor" "patient" "treatment"

we represent each word as a vector: $[\{R\}^{\{d\}}]$

where (d) is the embedding dimension.

The key idea is that words occurring in similar contexts tend to have similar representations.

In this project, you will build your own small-scale word embedding system using approximately 1,000 scientific abstracts.

You will implement and compare two classical neural language-model architectures:

Continuous Bag-of-Words (CBOW) and Skip-Gram

You will then investigate the learned semantic space using: cosine similarity, nearest-neighbor analysis, PCA, t-SNE, 2-dimensional visualization

The central question of this project is:

Can a relatively small collection of scientific abstracts learn a meaningful semantic representation of words?

2. Learning Objectives

本專案應完成下列工作：

文字前處理：執行斷詞、英文小寫化、標點處理，並決定是否移除停用詞及低頻詞。

建立詞彙表：統計文件數、總詞數、詞彙數及高頻詞。

產生訓練樣本：設定上下文視窗 (Context Window)，分別建立 CBOW 與 Skip-Gram 所需的訓練資料。

訓練詞嵌入模型：分別訓練 CBOW 與 Skip-Gram，建議向量維度為 50 或 100，並記錄主要超參數及訓練損失。

詞彙相似度分析：使用餘弦相似度，針對至少 10 個指定或自選詞彙，找出各模型最相近的前 5 個詞，並比較結果。

詞嵌入視覺化：使用 PCA 與 t-SNE 將詞向量降至二維，繪製散佈圖並標示詞彙名稱，觀察語意相近詞是否形成群聚。

Required Analysis 請回答以下問題

以實驗結果回答下列問題：

RQ1：CBOW 與 Skip-Gram 有何差異？

比較兩種模型的訓練方式、訓練損失、近鄰詞彙及對低頻詞的表現。

RQ2：詞嵌入是否具有語意意義？

選擇至少 5 組相關詞彙，例如 cancer、tumor、treatment、diagnosis，分析其餘弦相似度及近鄰詞彙，判斷模型是否學到合理的語意關係。

RQ3：PCA 與 t-SNE 的視覺化結果有何不同？

觀察語意相近詞是否形成群聚，並說明兩種降維方法的特性與限制。注意：視覺上形成群聚不代表模型一定具有良好的語意表示能力。

RQ4：模型參數會如何影響結果？

至少選擇一項參數進行比較，例如上下文視窗大小（2 與 5）、向量維度（50 與 100），或最低詞頻門檻，分析其對近鄰詞彙及視覺化結果的影響。

RQ5：小型語料的限制為何？

討論 1,000 篇摘要是否足以學習穩定的詞嵌入，以及低頻詞、專業術語、語料領域和資料量可能造成的影響。

Background:

CBOW vs Skip-Gram

Create a comparison table:

Property	CBOW	Skip-Gram
Input		
Prediction target		
Training examples		
Training speed		
Rare words		
Expected behavior		

Then answer:

Why might Skip-Gram be particularly useful for rare or domain-specific scientific terms?