**自然語言處理**

**Natural Language Processing**

**Assignment 2: Arithmetic as a language**

|  |  |
| --- | --- |
| 開設學校： | 國立清華大學 |
| 就讀校名： | 國立北科技大學 |
| 就讀班級： | 資工碩一 |
| 修課學號： | 113598043 |
| 修課學生： | 張育丞 |
| 授課老師： | 高宏宇 教授 |

Oct 24, 2024

**目錄**

[**1.** **前言** 0](#_Toc179217339)

[**2.** **程式實作** 0](#_Toc179217340)

[2-1. Part I: Data Pre-Processing 0](#_Toc179217341)

[2-2. Part II: Use Pre-Trained Word embeddings 1](#_Toc179217342)

[2-3. Part III: Train your own word embeddings 2](#_Toc179217343)

[**3.** **進階探討** 6](#_Toc179217344)

[3-1. 向量維度 (vector\_size) 6](#_Toc179217345)

[3-2. 視窗大小 (window) 6](#_Toc179217346)

[3-3. 最小詞頻 (min\_Count) 6](#_Toc179217347)

[3-4. 工作者數量 (workers) 6](#_Toc179217348)

[3-5. 參數調整前後對比結果 7](#_Toc179217349)

[**4.** **環境建置資訊** 8](#_Toc179217350)

[**5.** **實驗心得** 9](#_Toc179217351)

**圖目錄**

[圖1 (a) 資料讀取，(b) 提取前十筆資料，(c) 提取結果 0](#_Toc179217410)

[圖2 (a) 創建資料框架與資料寫入，(b) 寫入結果 1](#_Toc179217411)

[圖3 詞彙比較處理 2](#_Toc179217412)

[圖4 詞彙比較處理結果 2](#_Toc179217413)

[圖5 (a) 下載.gz壓縮檔案，(b)解壓縮與合併並刪除.txt和.gz檔案 3](#_Toc179217414)

[圖6 維基百科文章取樣20% 3](#_Toc179217415)

[圖7 Word2Vec模型訓練與處理 5](#_Toc179217416)

[圖8 Word2Vec模型的詞彙比較 6](#_Toc179217417)

[圖9詞彙比較處理結果 6](#_Toc179217418)

[圖10 參數調整前 7](#_Toc179217419)

[圖11 參數調整後 7](#_Toc179217420)

**表目錄**

[表1 開發平台與工具 8](#_Toc179217465)

[表2 套件名稱與版本 9](#_Toc179217466)

# **前言**

# **程式實作**

# **境建置資訊**

本報告使用到良好的設備(表1)與幾個套件工具進行實驗(表2) ，這些工具可以使得整個實驗更加地有效率開發讓整體更加完善。

表1 開發平台與工具

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Running environment | System: Windows 11 Pro,  CPU: Intel Core i9-12900K,  RAM: 16GB \* 2 (32GB)  GPU: NVIDIA RTX 3090 \* 1 |
| Visual Studio Code | Sep 2024 (version 1.94) |
| Environment management | Conda 24.9.0 |
| Python version | Python 3.11.9 |

# **實驗心得**