本研究訓練Word2Vec之程式碼參考：

https://github.com/Alex-CHUN-YU/Word2vec

https://radimrehurek.com/gensim/models/word2vec.html

詳細操作步驟及改動地方如下說明：

1. 先下載訓練資料，本研究使用wiki公開資料訓練Word2Vec，download wiki data:

https://zh.wikipedia.org/wiki/Wikipedia:%E6%95%B0%E6%8D%AE%E5%BA%93%E4%B8%8B%E8%BD%BD

1. 安裝需要用到的python套件：

pip install gensim

pip install jieba

pip install hanziconv

1. 接著進入 Word2Vec 資料夾
2. 執行

python wiki\_to\_txt.py zhwiki-latest-pages-articles.xml.bz2 （你下載的wiki公開資料的絕對路徑）

此步驟將wiki xml轉換成text，在當前資料夾下會產生wiki\_text.txt，方便程式後續做處理。

1. 開啟segmentation.py
2. 執行 python segmentation.py

簡體轉繁體，再進行斷詞並同步過濾停用詞，在當前資料夾下會產生segmentation.txt，由於檔案較大故斷詞較久

本研究僅訓練英文model，因此在簡體轉繁體的程式碼部分已註解，另外在停用詞的部分本研究使用了NLTK工具箱的停用詞，詳細請參考程式碼。

1. 調整訓練的參數，執行 python train.py，訓練並產生 model，時間上會比較久

參數調整可參考：

https://www.kaggle.com/jerrykuo7727/word2vec

https://radimrehurek.com/gensim/models/word2vec.html

1. 執行 python main.py測試本次訓練的model