IR: Homework 3

Name: 莊富成 Student ID: P77101016

Library: RE, NLTK, string, glob, pandas, matplotlib, genism, sklearn

1. Source documents:
   1. Homework2的csv檔案，剔除掉有非英語系或者是abstract長度小於10個字的文章，收集1000/5000篇
   2. 利用pubmed API收集influenza abstract 5000篇，考慮跟COVID-19時間錯開，只收集2019年之前的文章
2. CBOW/Skip-Gram利用gensim:word2vec來training
   1. stop-word考慮句子大多不長並不剔除
   2. 跑1000/5000篇COVID-19的model，比較CBOW/Skip-Gram不同篇數的model差異點
      1. CBOW無論何種都比Skip-Gram 訓練時間快2-3倍
      2. 因無有效的評估方法，用vaccine這字評估
         1. 1000篇: 同樣條件下skip-gram可以跑出類似的疫苗，CBOW則跑出同義類似字
         2. 5000篇: CBOW也會有類似的成果，在epoch算過多次下，skip-gram似乎變成overfitting導致most similarity words越來越奇怪，反而epoch少的時候skip-gram效果不錯
      3. 用TSNE降維度後，分佈圖的確可以看出一區一區的類似字
      4. 不同類比vector相減類比似乎無法類似像google word2vec的效果，應該跟文件內容有關，此外influenza的文件和COVID-19文件的類似關係不一定可以找出，也應該跟收集的文件內容有關。