1101\_人工智慧資訊擷取技術

HW4 內容說明 張育誠 P77091041

**使用語言：**C#, python

**python 使用套件：**os, glob, pandas, numpy, operator, re, string, json, flask, nltk, gensim, matplotlib, sklearn, random

**流程：**1. 使用者透過 UI 選擇Corpus，以及預先訓練好的 Word2Vector 的模型，然後輸入想要搜尋的 pattern，並決定要使用哪一種TF (replace TF with 1 + log(TF) or not)  
2. 模型先搜尋出相似度最高的50個字，然後把Corpus前100個摘要，進行去除標點符號以及停用字，並且斷句  
3. 去除不用的字後再將剩下的字詞用空白串接起來  
4. 使用sklearn的TfidfVectorizer計算出TF-IDF矩陣，將每一個相似字進行stemming，與TfidfVectorizer切出來的每一個字的stemming結果進行比對，如果有比對到相同的，就將該字詞的Max(TF-IDF) 及對應的句子(1st sentence)抓出來  
5. 秀在 UI 上面

<https://github.com/caubekimo/NCKU_110_IR/tree/master/hw4>