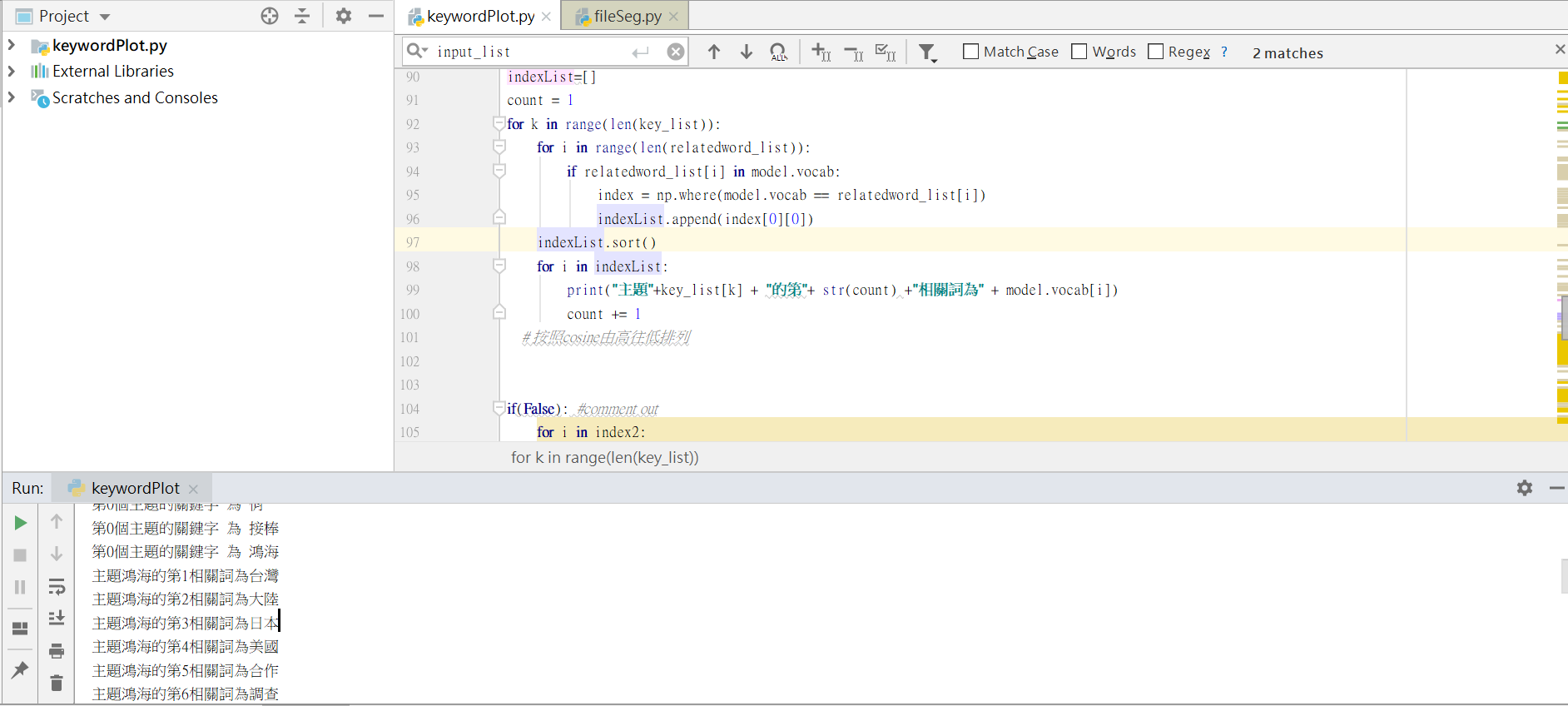
使用Jieba套件對中文資料進行斷詞斷句(目前繁體中文的斷詞斷句效果較簡體差)  
藉由word2vec實行Vector space based方式，將文字資料弄成向量空間，計算彼此之間的餘弦值來比較關聯程度  
這邊已做到變數化，可以輸入至8個主題(因為matplot只有八種顏色)  
fileSeg.py : 已經做完斷詞斷句輸出為hw1Output1.txt，再根據此輸出，弄成word2vec可以讀取的形式hw1Word2Vec.bin  
所以直接執行keywordPlot.py即可，根據上面的指示輸入想要查詢的主題  
舉例：這次作業要求 “鴻海”之關聯詞，輸入”鴻海”為程式所需之主題，接下來再輸入0代表要抓取全部的關鍵字可自由輸入想抓取多少關鍵字)  
程式抓完關鍵字會再從中裡面抓裡面有[老師給的245個關鍵字]，以此作為最終輸出結果  
  
輸入”主題”=>程式找出主題之關鍵字 => 這些關鍵字與老師給的245關鍵字做比較，如果此關鍵字存在才放入關鍵詞List裡面，最後依照cosine距離print出來

輸出結果如下(主題鴻海的第i相關詞為x)(按照cosine距離排序)



雖然往下還有更多的關聯詞(最後一個就是cosine距離越低，但還是有出現在老師詞集上面的關鍵字)，根據作業要求，取前20個相關詞為輸出結果

原因：想試試看實作word2vec,用文字變成空間向量，算文字之間彼此的距離的方式，來找關鍵字或是相關詞，發現效果還算可以。以空間來說，距離越低即cosine越低，直白解釋就是關聯低。

程式碼+資料集之連結：<https://drive.google.com/open?id=1ktgdJSihi7WMtGSU9pxuakdWYKeH0Hib>