106 學年度大四工工專題摘要

第 19 組	以文字探勘技術為基礎的文本校正方法
指導教授	侯建良 教授
參與學生	103034017　蔡季翰 103034021　賴宣宇 103034053　蔡孟桓 103034060　謝瑞璟

摘　要

隨著中文輸入法不斷的進步，智慧選字功能為中文輸入帶來許多方便，有效加快了輸入速度，但卻也衍生出容易出錯、漏字的問題。

本研究即是想以文字探勘技術，找出判斷文本中有出錯或漏字之處，並提醒使用者，也提供校正建議。例如：如果想輸入「總而言之」，可能使用者不小心打成「總言」，本研究希望能判斷出「總言」不為一個詞(需校正詞)，可能是漏字原因造成，能提醒使用者，提供校正建議「總而言之」。我們想試著利用統計及文字分析的概念來，做出一套演算法辦到這件事。當然，「總言」也可能是「總統所言」，所以我們也要建一套規則去判斷他，當「總言」被判斷出來，系統便會比對已有的資料庫文件，找到含有「總言」兩字的字詞及計算頻率，並依文字過去出現頻率，使用者輸入法會進行候選字之評分，並將建議選擇列出給使用者挑選。

要寫出此演算法，我們首先先建立兩個詞庫以判斷是否為待校正並給出校正選項：一個為由 python 的 jieba 套件及微軟舊注音兩個內建詞庫構成的基本詞庫，另一個是隨時代變遷或不含在基本詞庫內之詞的新興詞庫。新興詞庫的建造法是先蒐集大量的資料庫文件，並將詞語按規則依序拆開同時計算頻率，頻率高者我們便把他判斷為詞，可加入新興詞庫。

當待校正文本被讀入並利用 python 的 jieba 套件斷詞後，每個斷詞會先被檢查是否在基本詞庫內，若不在便會被判斷成需校正詞，系統會再將待校正詞每個字拆開，分別去兩個詞庫找含有這些拆開字的詞，並依照頻率依序排列給使用者選取校正選項。例如：上例待校正文本中「總言」被判斷為需校正詞，系統「總言」將拆開成「總」及「言」，並在基本詞庫找到「總而言之」、在新興詞庫找到「總統所言」，系統再依頻率排列建議給使用者修改。此外，系統也會加入使用者經驗選項，被使用者選取次數較多的會調高詞頻，使用者也可設定系統判斷的需校正詞判斷錯誤，此詞便會被加入新興詞庫。