IR 程式作業技術報告

第八組 104753034 崔嘉祐, 104753039 張至偉

Packge : python Whoosh 2.7.0

參考SEIRiP課本中P.15，索引建立的過程與要素，並從實作面討論遇到的問題與解決方法。

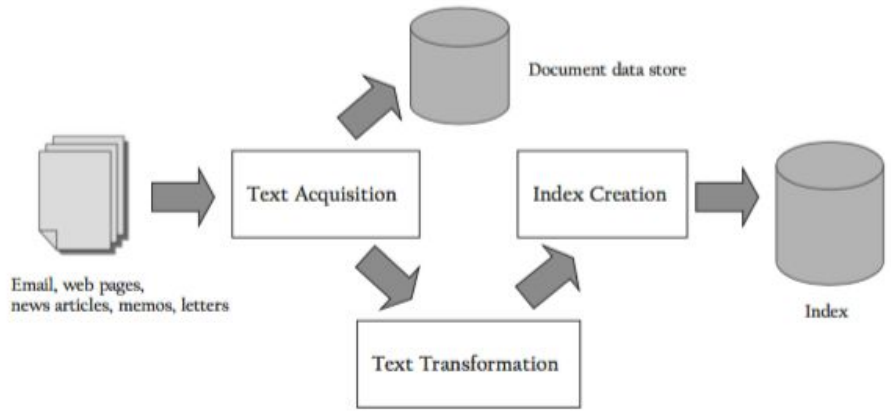
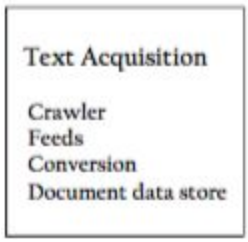
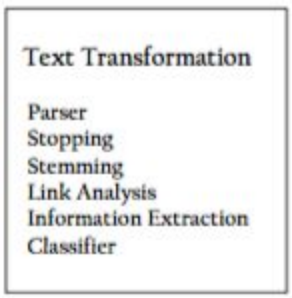
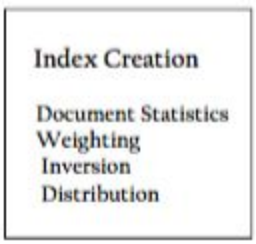


Fig. 2.1. The indexing process

Text Acquisition	Text Transformation	Index Creation
		
資料來源：聯合報“客家”相關報導。 取得方式：同學們**分工下載**。	遇到的問題：因為資料的語言為中文，中文的句子中沒有斷句，而搜尋的query通常都是以詞的形式來輸入，會造成輸入的query無法找到相關的對應文章。 處理方式：對於資料文本進行2-grams或3-grams的斷詞處理，以利於query與文本的搜尋對應。	遇到的問題：建立Index表格時，首先要定義Index的欄位們，好用來決定什麼欄位與query比對，或決定什麼欄位來參與query搜尋結果的排名。 使用工具：Whoosh，以python語言寫成，用於建構Index，並決定對應的Schema來完成此次客家資料的Index表格。

參考SEIRiP課本中P.16，搜尋的處理與系統流程，並從實作面討論遇到的問題與解決方法。

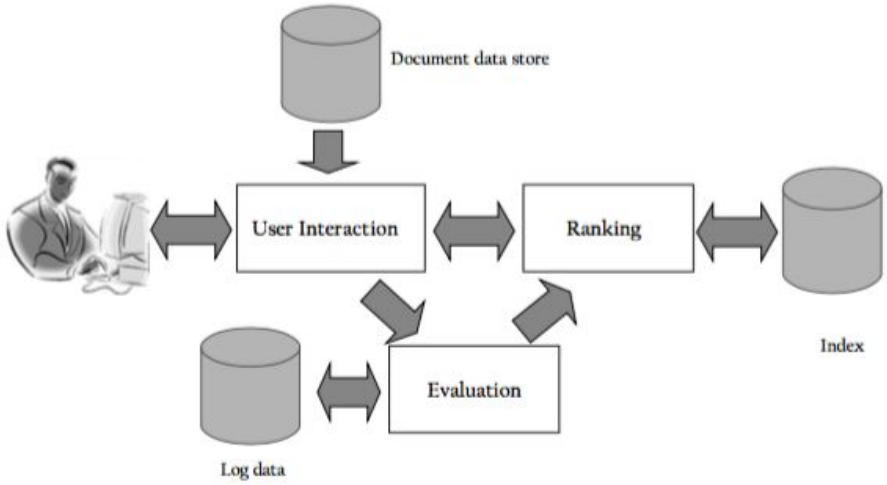
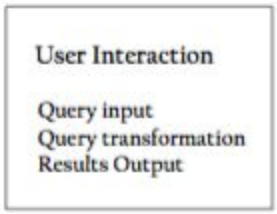
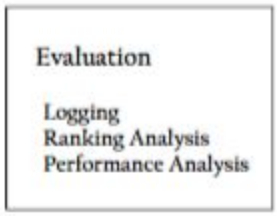


Fig. 2.2. The query process

User Interaction	Ranking	Evaluation
		
遇到的問題：題目要求，要能支援query中的boolean運算。 解決方式： 先使用Whoosh中的Parser功能，先以Boolean運算的關鍵詞為目標，將Sentence切割成個別子句，並完成子句間的Boolean運算後，再以其運算結果進行搜尋。	遇到的問題：如何排序搜尋結果？ 解決方式：利用Woosh中searcher.search()函式中的第二個參數，將Schema中可以排序的欄位來進行排序，排序的依據，為關鍵字詞的出現次數，或是報導時間，或是報導的地點...等等。 Snippets的實作與Highlight相關的關鍵字詞： 使用Whoosh中的 **whoosh.searching.Results** **whoosh.searching.Hit** 來實作Highlight KeyWord。 使用Whoosh中的 **Fragmenters Type** 來實作snippets的顯示。	留在最後有多足夠的時間再來補強我們的Search Engines。

References:

圖片來源：Search Engines Information Retrieval in Practice (©W.B. Cro , D. Metzler, T. Strohman, 2015)
工具來源：Python package Whoosh 2.7
相關資訊來源：https://pythonhosted.org/Whoosh/