Information Retrieval and Extraction

Student Name: 黃正一
Student ID: R93921092

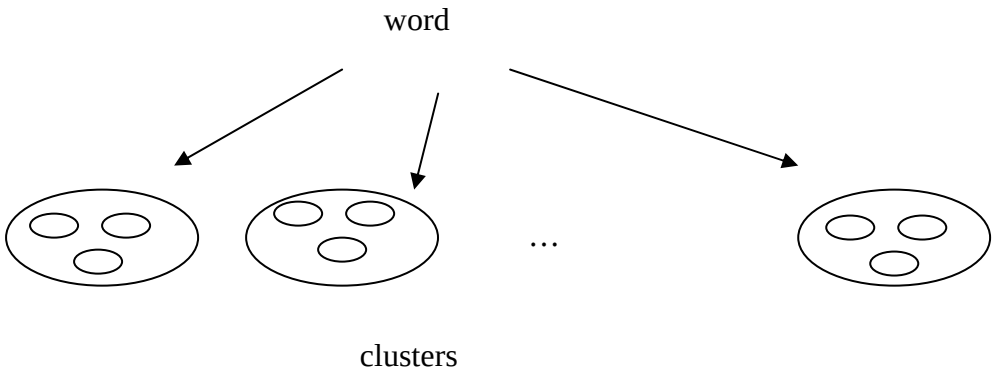
Introduction

對文件做 index，並以 incremental 的方式來做 indexing，並對 indexing 後的文件做 search。

Indexing

Index 的步驟如下:

1. 先對文件去除 function words :。
2. 對文件，做 stemming 的動作: 以查詢 WordNet 來輔助做 word 的 stemming，。
3. 開始做 indexing : 爲了避免用來存放 index word 的 hash table 佔據太多的記憶體，及在做 incremental 時，能夠載入少量的資訊即能做 incremental，這邊對於存放字的 inverted file 在做 cluster。每個 word 會屬於兩類的 clusters，稱 outer cluster，及 inner cluster。Outer cluster 爲根據 word 的第一個字母來 clustering。而當 word 被 clustered 至 outer cluster 後，在根據 word 的第二字母分配至 inner cluster，而每個 outer cluster 中，有三個 inner cluster，此即對 26 個字母分三等份，以 word 的第二個字來判斷，此 word 是該分配至哪個 inner cluster。此種 clustering 的方法，可以減少不必要的資訊被載入記憶體中，並可加入 incremental 及 search 的速度。



Searching

Search 的步驟如下:

1.對 topics 做 function words 的去除及 stemming。

2 先對 topics 做 indexing。
3 根據 indexing 後的文件，及 index 後的 topic 做比對。以計算每個 document 所得到的分數。而這邊計算分數的方式非常簡單，比較 indexing 後的 topics 中和 indexing 後文件的字有多少個相同的，當相同的字數越多，此 document 得到的分數越多，此即此 document 越能代表此 topics。此種 search 的方法的優點爲容易 implement 且速度快，但它的缺點爲 recall 及 precision 低。

實驗數據

Avg P	P at R(10%)	P at 15 docs	F Idx T	Inc T	S T
0.0046	0.0152	0.222	740S	861S	193s

結論 ：
在此 project 中，學到了如何對文件做 indexing，並做 incremental 的動作。
根據 indexing 後的文件內容，加以做 search。由於時間的不足，所以無法在 searching 的 precision 及 recall 上，做太大個改進，這是比較遺憾的事。