國立政治大學資訊科學系

2015 年資訊檢索程式作業

工作目標：自己實作一個簡單的資訊檢索系統並且作為資訊分析的基礎

基本規則

1. 程式作業必須獨力完成，不當的類似程式將被視為抄襲，整個程式作業將以零分計算。情節嚴重者將呈報學校以作弊規則處理。

資料來源與分工

1 進入政治大學圖書館網站，透過“資料庫”，進入“聯合知識庫”
2 這一次的作業每一位同學都要下載以“客家”為關鍵詞所找到的一些文章
3 進行檢索的時候，請限制在“聯合報”的範圍。不要經濟日報、聯合晚報、upaper。
4 每一位同學都要透過 Moodle 課程網頁登記自己負責下載的時段。每位同學分攤不重疊的六個月的時間。
5 每一個人要下載所分攤的時段的每一個月份的八篇文章。
6 以所負責的時段作為檔案名稱。每一位同學把自己準備好的 48 篇文字放在以自己負責的時段為名稱的資料夾。並且把資料夾壓縮成成一份 RAR 檔案。如果你負責的是 2013 年的上半年，則壓縮檔案就叫作 201301.rar，裡面有一個叫作 201301 的資料夾；如果你負責的是下半年，則壓縮檔案就叫作 201302.rar。
7 在時限之前上傳到 Moodle 課程網站，這樣全班同學可以共同使用這一批資料。
8 繳交文字檔案時，請說明你所負責的時段中，聯合報有多少篇提到“客家”的文章。並且報告你所提供的 48 份檔案中“客家”一共出現多少次？(這樣我可以用我的程式檢查你是否真的處理了你自己的資料)

儲存格式

1. 請以沒有 BOM 的文字檔案儲存，中文內碼請用 UTF-8。
2. 儲存的檔案名稱必須符合規則。每一檔案的檔名都以“聯”字開始，並且串上所選擇的報導的日期。一篇在 2015 年 5 月 5 日在聯合報刊出的文章的檔名要用“聯 20150505.txt”； 一篇在 2015 年 11 月 15 日在聯合報刊出的文章的檔名要用“聯 20151115.txt”。請參看 Moodle 課程網站上的範例。
3. 聯合報的文章標題區域，必然有文章主標題，有時在主標題之前會有主題區，在主標題之後會有副標題。主題的文字通常前後有標點符號《和》作為特別標示。
4. 儲存文字檔案的時候，請忽略主題。把主標題放在文字檔案的第一列，副標題放在第二列。聯 20151115.pdf 是一篇沒有主題、有主標題和副標題的文章。所以，聯 20151115.txt 的第一和第二列分別記錄了主標題和副標題。
5. 文字檔案的第三列和第四列，請留白。
6. 大部分的聯合報文章有作者的資訊，大都放在文章開頭、並且以特殊標點符號【】之中。請把作者姓名放在文字檔案的第五列。
7. 複製文章內容的時候，請不要包含【】這一列作者資訊。
8. 文字檔案的第六列留白。
9. 把文章區域複製貼到文字檔案中，然後儲存。

軟體工具

1. 請使用 Lucene 或者 Solr 來協助你建構索引和提供基本的檢索功能。
2. 你也可以利用例如 Python 的 Jieba 來作中文斷詞。

功能要求

1. 為了測試檢索功能，你需要建立一個類似 Google 網頁的檢索介面，讓使用者輸入所要檢索的關鍵詞。
2. 查到文章的時候，你必須表列文章的標題，並且可以直接點選該標題，看到文章的內容。
3. 在表列檢索到的文章時，你還必須提供各篇文章的 snippet。而 snippet 也應該包含使用者所檢索的關鍵詞。
4. 單詞彙檢索時，必須顯示該詞彙在所檢索的語料庫中出現的總次數。
5. 使用者可以檢索多個詞彙，這時你的程式必須能夠做到基本的 ranking，例如包含愈多檢索關鍵詞的文章應該排在比較前面。
6. 多詞彙檢索時，可以用英文表示 and、or、not 的關係
7. 可以進行最基本的 proximity search，例如“客家 /2 小吃”。
8. 撰寫兩頁技術報告，說明各組完成上述功能時，所採用的工具和所應用的技術。
9. 分析你自己的 48 份檔案，看看裡面一共出現過多少次“客家”。(文字資料上傳之時就要完成)
10. 記錄你所負責的時段中，聯合報有多少篇提到“客家”的文章。(文字資料上傳之時就要完成)

分組規則

1. 可以一人獨立寫這一程式，也可以至多兩人一組。但是功能的需求不會因為人數多少而異。
2. 期末驗收的時候，每一成員都必須同時出席接受詢問；沒有出席之個人，程式作業以零分計算。
3. 任一組員必須合理分攤程式的設計，在系統展示的時候，則必須了解該組所繳交的所有程式，並且回答相關問題。

計分規則

1. 合理介面 10%；基本單詞檢索 20%；多詞基本檢索 15%；proximity search 15%；邏輯關係 15%；ranking 15%。技術報告 10%。
2. 沒有上傳文字檔案者，程式作業將以零分計算。
3. 沒有出席參與期末演示和問答者，程式作業將以零分計算。
4. 期末演示時無法適當回答問題者，程式作業分數將會被依照情況打折扣。

資料品質問題

資料錯誤經常是分工下載之後，導致其他同學不能處理語料的主要障礙。為了大家的工作效率。請務必依照前述說明備妥你的資料。如果發現錯誤，將會有負面作用。每一次發現問題，就會扣所得分數的 5%。

預定工作時程

1. 十一月廿四日午夜 23:59 前完成文章分工的時段登記。否則將被任意分配。
2. 十二月一日午夜 23:59 前完成個人文字檔案上傳。
3. 十二月一日午夜 23:59 前完成程式分組，並且在 Moodle 課程網站完成登記，否則視為個人獨自一組。
4. 十二月卅一日午夜 23:59 前寄交各組技術報告。
5. 一月十二日上課時間各組測試與問答。

其他相關問題

1. 在 Moodle 網站上討論