數位方法：基於關鍵字抽取的文本分類

前言與動機

THDL與淡新檔案這兩個資料庫具有許多與清代石油開發的史料，並且在沒有任何數位協助的情況下，閱讀和分析與我們組的報告主題相關的文本將需要耗費大量時間。因此，由於上述原因並且由於本課程也是資工系開的課，我們認為有必要納入資工領域以協助文字分析 (或任何其他方面)。

經過了討論和熟思，我們決定嘗試實行文本比較工具與文本分類。很遺憾，由於缺乏可用的資料集，並且時間與人力的原因，只能完成基於關鍵字抽取的文本分類系統。

雖然本系統以及使用的演算法可被認為樸素，不過實現的文本分類仍然有意義的，並且（希望）可以協助組員更便利的查找與報告主題相關文本。

目標與作品

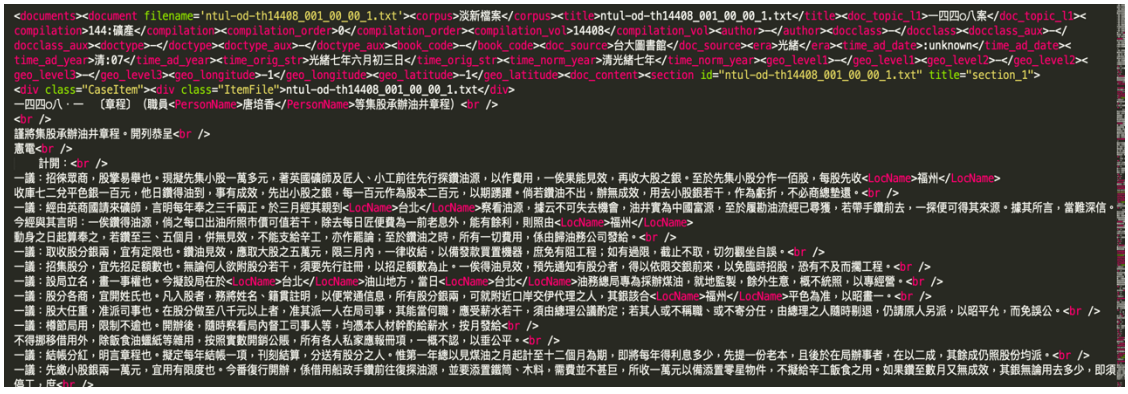
目標：從給定的相關文本集合中，可以根據文本內容將文本分類為有組織的類別。
作品：

1. csv檔：包含所選文本及其元素，分類與額外有用的資訊的表格。
2. 網頁：一個文本檢索系統，根據來源和分類檢索文本。

詳細步驟

1. 首先，從資料庫中查找並下載與報告主題相關文本為xml檔。

2. 接著，用python程式將xml檔轉換為csv格式。
    a. 要為相應的csv檔選適當的欄位名稱，必須首先查看xml檔，以了解它們的標記格式 (tagging format)。
    例如：



b. 由於從THDL與淡新檔案下載的xml檔的標記格式有點不同，因此它們各自的csv檔的欄位名稱也有些差別 (CSV檔中的欄位名稱的命名約定遵循相應的xml標記)：
    i. 從「THDL」下載的：[n, title, author, time_orig_str, time_norm_year, ad_year, 內容]
    ii. 從「淡新檔案」下載的：[n, Abstract, author, doc_source, time_orig_str, time_norm_year, ad_year, 內容, ContentLocation]

c. 使用的python模組/套件:

```
1  import xml.dom.minidom
2  import xml.etree.ElementTree as ET
3  from bs4 import BeautifulSoup
4  import csv
```

3. 基於文本的內容，將表格中的每個文本做分類，接著適當地將取得的資訊加上在CSV檔中。
    a. 首先，必須界定類別單位及其各自的關鍵字。不幸的是，這只能手動完成。
        i. 有4個類別：「環境」、「戰略」、「商人」、「官辦」。
        ii. 每個類的關鍵字（每個類別5個）：
            1. 環境：'山','採','產','礦油','林木'
            2. 戰略：'海防','奏','軍','洋','艦隊'
            3. 商人：'商','墾','租','口供','業務'
            4. 官辦：'辦','油山','機器','飭','新竹'
    b. 對於每個文本，讀取「內容」部分，接著逐句處理它。
    c. 對於所有關鍵字，如果它出現在句子中，將它的出現記錄到其相應的變數(在列表中)。
    d. 文本分類將基於關鍵字的出現次數。如在文本中關鍵字的出現次數不滿足閾值，則將文本分類為“N/A”。

4. CSV檔更新後，將這些文件轉換為json格式，以編寫個網站。

5. 為了提升易用性與文本可讀性，創新一個網站。
    a. 網址：http://www.b06902100proj.byethost7.com/fp2/DRH.html
    b. 使用的程式語言/構架：HTML, CSS, JS (jQuery), Bootstrap

網站介面

初始顯示

按「送出」後

按「送出」後，設定：

- 文本來源：THDL
- 分類：戰略
- 排列方式：日期 (年)

按 年份旁邊的藍色按鈕後，文本內容將顯示。關鍵字用顏色標記。

分類成果與分析

1. THDL: 總共有43文本，10文本歸類為「環境」、22文本為「戰略」、1文本為「商人」、4文本「管辦」、3文本為「N/A」。
2. 淡新檔案：總共有142文本，1文本歸類為「環境」、1文本為「戰略」、10文本為「商人」、126文本「管辦」、4文本為「N/A」。

符合預期，結果並不完美，肯定可以改善。對於「淡新檔案」的部分尤其如此，既然幾乎所有的文本都被歸為同一類。正如口頭報告中亦提到，這系統有些缺點：

1. 許多組關鍵詞混雜，造成分類的精準度下降.
2. 未設定最少的關鍵詞數 (閾值太低)，導致關係甚低的內容混入其中
3. 淡新檔案的案關係緊密，導致官辦得分類太過集中

不過，這並不代表該系統無用。它仍減少閱讀與使用所需的時間 (但不是很顯著地)，而且文本也可以按年份或字數排序，這對讀者該也會有幫助。

可行的改進與總結

倘若倘若時間不成問題，則有許多方法可以改善系統，如：

1. 在網站的設定裡，加上另一個設置選項：「最少的關鍵詞數」，以更好地過濾文本。
2. 加上更多類別，添加和調整關鍵字。在幾乎所有情況下，文本分類都需要大量數據（和參數）才能達到很高的準確性。
3. 使用百分比格式，而不是將文本分類為一個類別。例如：

   文本A：25% 環境，50%戰略，10%商人，5%管辦，10%其他
4. 有些文本很長，可能對每個段落進行分類會比較好與準確。

以上的所有方法都可行的，並且很合理地會提升系統的實用性，但也必須記住，文本分類就像文本比較一樣，它們都需要很複雜的演算法與足夠大的資料集才能達到理想的準確性水平。

總而言之，在我們處理大量文本的情況下，文本分類將非常有益的。然而，建立這個系統，尤其是要做理想的，會需要大量的時間，數據與氣力。

*CSV檔可以在這裡下載 ：https://filebin.net/ipc7r1axjnn7x3ze