

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：

```
<documents><document filename='ntul-od-th14408_001_00_00_1.txt'><corpus>淡新檔案</corpus><title>ntul-od-th14408_001_00_00_1.txt</title><doc_topic1>一四四〇八案</doc_topic1><compilation>144:礦產</compilation><compilation_order>0</compilation_order><compilation_vol>14408</compilation_vol><author></author><docclass></docclass><docclass_aus></docclass_aus><doctype></doctype><doctype_aus></doctype_aus><book_code></book_code><doc_source>台大圖書館</doc_source><era>光緒</era><time_ad_date>unknown</time_ad_date><time_ad_year>清:07</time_ad_year><time_ad_year><time_orig_str>光緒七年六月初三日</time_orig_str><time_norm_year>清光緒七年</time_norm_year><geo_level1></geo_level1><geo_level2></geo_level2><geo_level3></geo_level3><geo_longitude>1</geo_longitude><geo_latitude>1</geo_latitude><doc_content><section id='ntul-od-th14408_001_00_00_1.txt' title='section_1'><div class='CaseItem'><div class='ItemFile'>ntul-od-th14408_001_00_00_1.txt</div>一四四〇八·一【章程】（職員<PersonName>唐培善</PersonName>等集股承辦油井章程）<br /><br />謹將集股承辦油井章程。開列於左<br />憲電<br />計開：<br />一議：招集股商。現擬先集小股一萬多元，專為開闢礦及匠人、小工前往先行探鑽油源，以資費用。一俟果能見效，再收大股之銀。至於先集小股分作一佰股，每股先收<LocName>福州</LocName>收銀七十二元平色銀一百元，他日鑽得油到，事有成效，先出小股之銀，每一百元作為股本二百元，以期騰讓。倘若鑽油不出，辦無成效，用去小股銀若干，作為虧折，不必再補贖讓。<br />一議：經由英商蘭蘭請來礦師，查明每年春之三千兩正，於三月經其領到<LocName>台北</LocName><LocName>聚看油源，據云不可失去機會，油井實為中國富源，至於廢勘油源經已尋獲，若得手續前去，一探便可得其來源。據其所言，當難深信。今經與其查明：一俟鑽得油源，倘之每口出油所照市價可值若干，除去每日匠便費為一前老息外，能有餘利，則照由<LocName>福州</LocName>動身之日起算奉之，若鑽至三、五個月，倘無見效，不能支給辛工，亦作罷論；至於鑽油之時，所有一切費用，係由辦油總局發給。<br />一議：收取股份銀兩，宜有定限也。鑽油見效，應取大股之五萬元，限三月內，一律收結，以備發款買置機器，庶免有阻工程；如有過限，截止不取，切勿觀望自誤。<br />一議：招集股分，宜先招足額數也。無論何人欲附股分若干，須要先行註冊，以招足額數為止。一俟得油見效，預先通知有股分者，得以依限交銀前來，以免臨時招股，恐有不及而難工程。<br />一議：設局立名，盡一事權也。今擬設局在於<LocName>台北</LocName><LocName>油山地方，當日<LocName>台北</LocName>油務總局專為採辦煤油，就地監製，餘外生意，概不統照，以專經營。<br />一議：股分各名，宜開姓氏也。凡入股者，務將姓名、籍貫註明，以便奉遞信息，所有股份銀兩，可就附近口岸交伊代理之人，其銀該合<LocName>福州</LocName>平色為准，以昭畫一。<br />一議：股大任重，准派司事也。在股分做到八千元以上者，准其派一人在局司事，其能當何職，應受薪水若干，須由總理公議酌定；若其人或不稱職、或不奉分任，由總理之人隨時刷退，仍請原人另派，以昭公平，而免誤公。<br />一議：得節局用，限制不逾也。開辦後，隨時覈看局內督工司事人等，均憑本人材幹酌給薪水，按月發給。<br />不得挪移信用外，除飯食油鹽紙等雜用，按照實數開銷公帳，所有各人私家應繳開項，一概不認，以重公平。<br />一議：結帳分紅，明言章程也。擬定每年結帳一項，刊刻結算，分送有股分之人。惟第一年總以見成油之月起計至十二個月為期，即將每年得利息多少，先提一份老本，且俟於在局辦事者，在二成，其餘或仍照股份均派。<br />一議：先收小股銀兩一萬元，宜用有限度也。今暫復行開辦，係信用船政手續前往探鑽油源，並要運置鐵器、木料，需費並不甚巨，所收一萬元以備運置等雜物件，不擬給半工飯食之用。如果鑽至數月又無成效，其銀無論用去多少，即須俟工竣後，再行發還。</div></div></div></documents>
```

- 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

文件檢索

設定

文本來源

分類

排列方式

THDL

環境

日期 (年)

送出

按「送出」後

文本來源: THDL 找到筆數: 13	
為臺道總冊事	同治8年
險石油一種	同治10年
險火山	同治10年
險產煤及茶穀棧屬地方	同治10年
奏報日本兵船已抵臺灣番境現正密籌防範暨詳陳臺灣地利日本詭謀等情	同治13年4月21日
奏報臺灣地方現況摺片	同治13年4月21日
李鶴年又奏、查臺灣一島，周袤三千餘里，孤嶼環瀛，土壤肥沃，禾稻不實而長，物產繁滋，礦、煤、樟腦、水藤、糖、蔗，靡不充餘。	同治13年
閩浙總督李奏	同治13年
臺灣雜錄	光緒1年

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。

奏報日本兵船已抵臺灣番境現正密籌防範暨詳陳臺灣地利日本詭謀等情

同治13年4月21日

作者: 閩浙總督李鶴年

閩浙總督兼福建巡撫李鶴年奏：據臺灣道夏獻綸、為據枋寮巡檢等，稱得三月二十二、二十三等日，有日本火輪船兩號，駛至蘭(王爺)社臺灣港口停泊，人數約有八九百名，先遣洋人二十餘名，至南境番界探勘紮營地勢各情。業由該道轉報前來。臣查日本中將，在廈門呈遞照會後，並不候臣照覆，即行開駛赴臺；又不往噶臺灣鎮道，遽行登岸窺覷紮營，居心殊為叵測，便得志於生番，必將統轄中國、僅以山深澤廣，失利喪師，難保不別生枝節。事關臺灣全局，自宜先事豫籌。現已遵飭候補參將李學祥、遊擊王開俊，督帶健勇屯駐鳳山一帶，以資鎮壓。臺地民情強悍可用，並已遵飭鎮道，號召閩粵聯莊，暨飭地方文武，嚴密防範。一面遴派幹員，馳赴蘭(王爺)社，面見該國兵官，按約理論，阻令回兵。臺灣口岸，原有長勝福星輪船駐泊，茲又前派參將員錦榮，督帶精武兵船，駛泊澎湖一帶，以遏聲息。廈門為臺灣入省咽喉，已派靖遠輪船駐。

彼。並飭水師提督李新燕，召募精勇，鎮調精兵，嚴加防範。又會同船政臣沈葆楨，飛調安瀾商船各輪船來臺，以壯聲勢。惟念邊疆易開不易戢，番地屬地，究有區分，如果倭兵侵入臺灣腹地，自當督飭鎮道該國兵團，合力堵剿；若僅以成線地球難民為名，與生番復仇，惟當按約理論，不遽發難致討，以免冒開自我。臣受任封圻，不敢過事張皇，亦不敢稍存大意，俟該中將接到臣照覆後，如何情形，再行奏報。

李鶴年又奏：查臺灣一島，周袤三千餘里，孤嶼環瀛，土壤肥沃，禾稻不實而長，物產繁滋，礦、煤、樟腦、水藤、糖、蔗，靡不充餘。其生番所居內山，未闢境地，尚什之七。其內材木連山，傳聞五金晶玉之礦，礦油煤油之井，園地皆有，物產頗富，更勝於已闢之地；且內外山地，俱宜栽茶，自西陲各國通商以來，無不覬覦其地，特以歐洲公法，有守單均勢之例，互相鈔制，莫敢先發難端。

日本倭人，在明天啟間，曾窺其地，後為荷蘭所奪，鄭成功又奪於荷蘭；迨康熙中，鄭氏遁而臺灣進入版圖，此日本所以尤為耿耿也。按之明人鄭若晉日本圖纂，倭人入犯中國，必至小琉球分(內附)。小琉球者，即臺灣之小島也。蓋其隸諸摩州及五島，皆與臺灣密連，隔輪舟一日可至，故為入犯必由之路。該國在明代三百年間，屢寇閩廣江浙浙海一帶，大為中國之患。自國朝定鼎以來，始漸傳帖息。海不得波，皆由臺灣隸入版圖屏蔽之力，從前中國與該國互市，惟三船前往，無便船西來；及各國通商，倭人始入內地，乃議和未久，遽發精兵，或者謂有西人說中勾引，恐難保其必無，雖出該國心體定諒，藉口報復生番，蓄圖覬覦，斷非可能。查倭性狡獪，好勇鬥狠，明洪武間命使往諭，南經入貢；按與胡惟庸通謀不軌，永樂朝遽使招諭，又曾先納款，而仍事寇抄。其後屢啟邊釁，史紀昭然。是狙詐狡貪，為其故習，非西洋各國效倭守約之比。

臣近接總理衙門來函，內開上海鈔送長崎電稿，祇云前年人民在臺灣生番地界違風船隻違員查問確情，並有誠恐偽詐之徒，擅行謠言等云。又另鈔英國使臣威妥瑪呈送節略，亦有日本並無出有向中國稱兵明文之語。原該國於與兵內犯之舉，故作隱約之詞，其心尤為險惡。竊該國中將西鄉爾會，於中國救援難民，嚴版還謝，即於生番，亦似有不適用兵之意；然既不同之總理衙門，又不候臣照覆，徑行統軍赴臺；復不往諸臺灣鎮道，直抵蘭(王爺)社岸紮營，或實於番地藉端重山深宵密之說，豫留為將就道兵之計；或為潛相勾結，俾圖占據之謀，均不可測。

總之，臺灣為沿海各省門戶，又且上倚物阜，乘隙伺者不一，即使目前不致成釁，日後之隱憂方大。臣惟有竭盡愚誠，隨時度勢，革舊兵戎，互相為用，務使理屈在彼，不令冒開自我。一面遴練兵勇，購置器械，健備聯援，延攬人材，以期有備無患，仰祈皇上委任封疆之至意。

來源: 近代中國對西方及列強認識資料彙編

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>