質化研究與量化研究的融合

文字計數新方法實驗

指導教授： 郭豐州 老師

專題成員：

96156214徐微婷　　96156224林玉清

96156229鄭婉伶　　96156237尤雅亭

96156238陳介文　　96156246詹凱婷

社會科學研究方法的兩大主流

- 質化研究法
  - 傳統的研究方式(B.C-1960)
  - 以訪談、觀察、文獻探討分析推論
  - 缺點：無法抽離自己的價值觀 ，研究結論較為主觀
- 量化研究法
  - 20世紀的研究方法
  - 以數學統計的方式，例如：用小樣本去推測母體的意見，選舉民調、出口民調、滿意度調查等均屬之
  - 缺點：雖然較科學，但是忽略文獻中原來就有的背景脈絡關係與邏輯性

質化 v.s. 量化

1960年代的學者

質化研究太主觀，
很不科學耶！

我們是不是太依賴量化研究
賣弄數據了？質化研究還是
有相當的價值啊！

2000年代的學者

質化&量化

- 於是，我們試圖在質化研究過程當中加入量化方法，讓彼此支援，產生交集
- 以一種研究方法為主，另一種方法為輔，來補足彼此不足的地方

文字計數

- 英國政治學者群長期閱讀政黨的文宣去歸納各政黨各種政策（例如：勞工、社會福利）在左派與右派間的變遷情形。但是此工作耗時耗力
- 2004年英國政治學者Michael Laver提出文字計數方法，使用電腦對文章內容進行文字計數，藉由文字出現頻率去推測此篇文本某一概念的強弱
- 他的演算法不僅在英文，應用在歐洲其他語言文字也都成功，對歐盟組織運作也有實質幫助。因此啟發了我們「 Michael Laver文字計數應用在中文」的研究動機

實驗過程

單字

- 依照Laver演算法，以一個中文字為一單位

斷詞

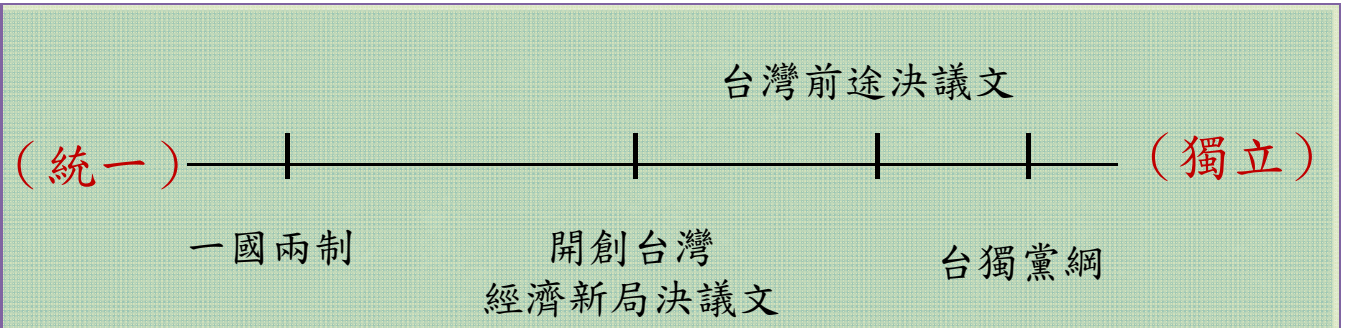
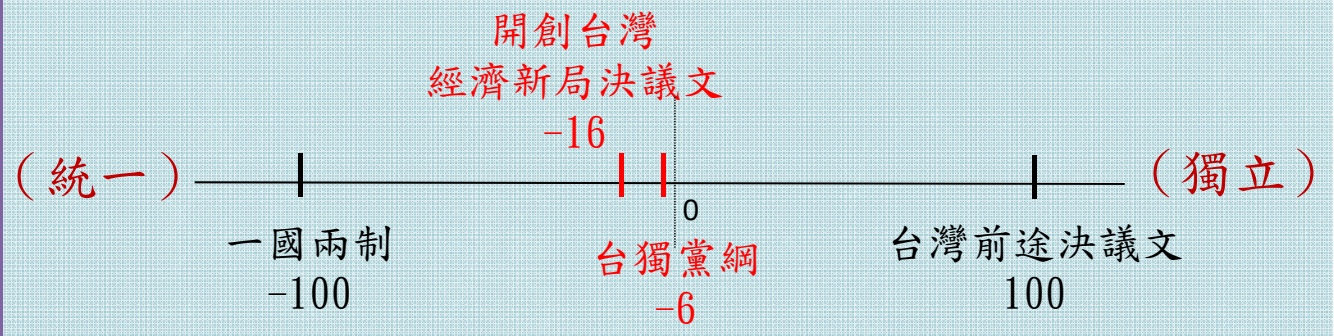
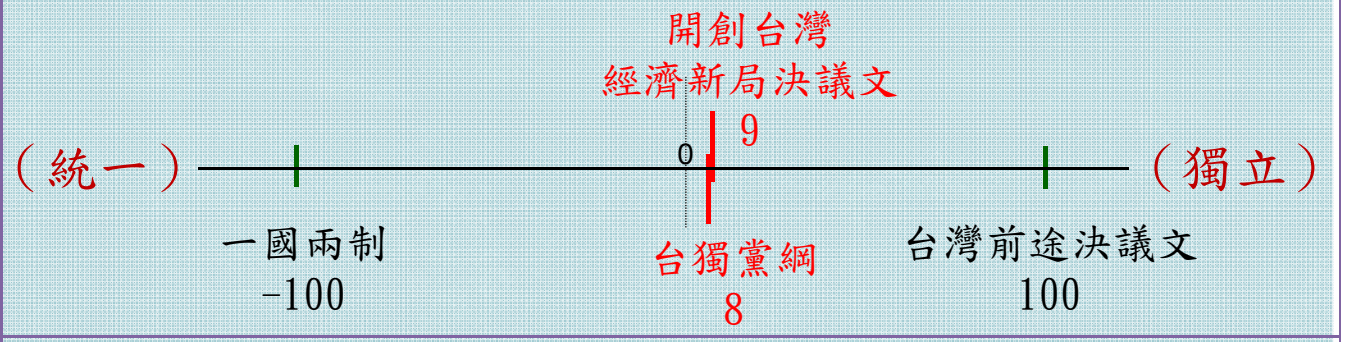
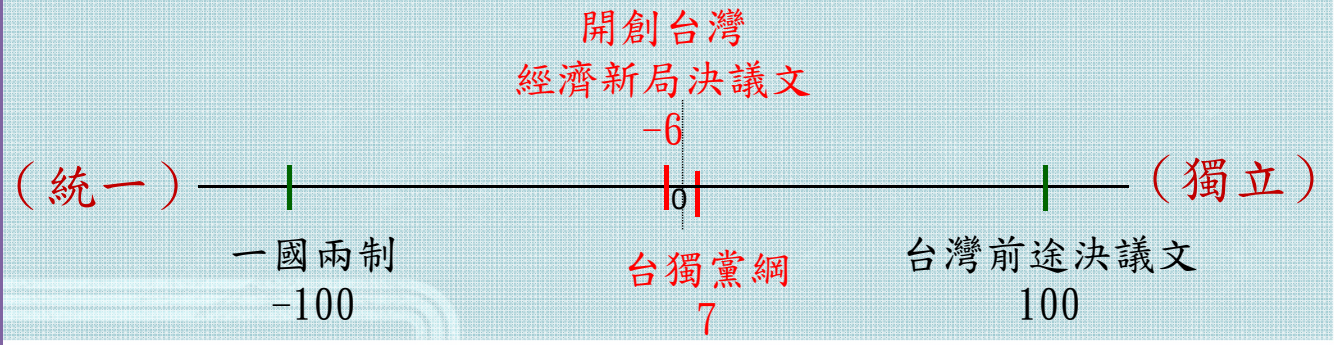
- 修改Laver演算法，使用中研院的中文斷詞系統，以一個斷詞當作一個單位

組合詞性

- 一個斷詞當作一個單位，並把一些否定詞與其後面的動詞、副詞...等組合成一個詞

測試文本介紹

- 台獨黨綱(1991民進黨)
- 台灣前途決議文(1999民進黨)
- 開創台灣經濟新局決議文(2001民進黨)
- 一國兩制(1982鄧小平)

預期結果	
Laver 演算法結果 （以單字為計算單位）	
斷詞結果 （以一個詞為計算單位）	
組合詞性結果 （以一個詞為計算單位）	

小結論

- Michael Laver文字計數應用在中文無論以單字或詞為計算單位結果都不對。中文和西方文字的意涵和文章的結構都有顯著的差異之故
- 於是，我們另尋方法，以拿來做中文自動作文評分系統的貝氏分類法來計算原始文件和測試文件的距離

貝氏分類法

運用貝氏定理，來做研究分析

$$P(H_i|d)=\frac{P(d|H_i)P(H_i)}{P(d)}=\frac{P(H_i\cap d)}{P(d)}$$

- P(d)= 該文本出現關鍵詞的個數
- H_1 = 第一類別所有關鍵詞的個數
- H_2 = 第二類別所有關鍵詞的個數

類別可以不只兩種，目前我們先取兩種可表現出對立關係的文本來測試

貝氏分類法應用步驟

- 步驟一
  取當中兩篇概念極端對立的文本當作機器訓練用文本
- 步驟二
  從兩篇原始文本的「字詞」取差集，得到相對關鍵詞
- 步驟三
  為了凸顯中文文章結構的特性，依文字出現在文章的位置給予加權

貝氏分類法測試文本一

- 以原先測試Michael Laver文字計數演算法時的四篇文本作測試

原始文本：

台獨黨綱(1991民進黨)（極端獨立）

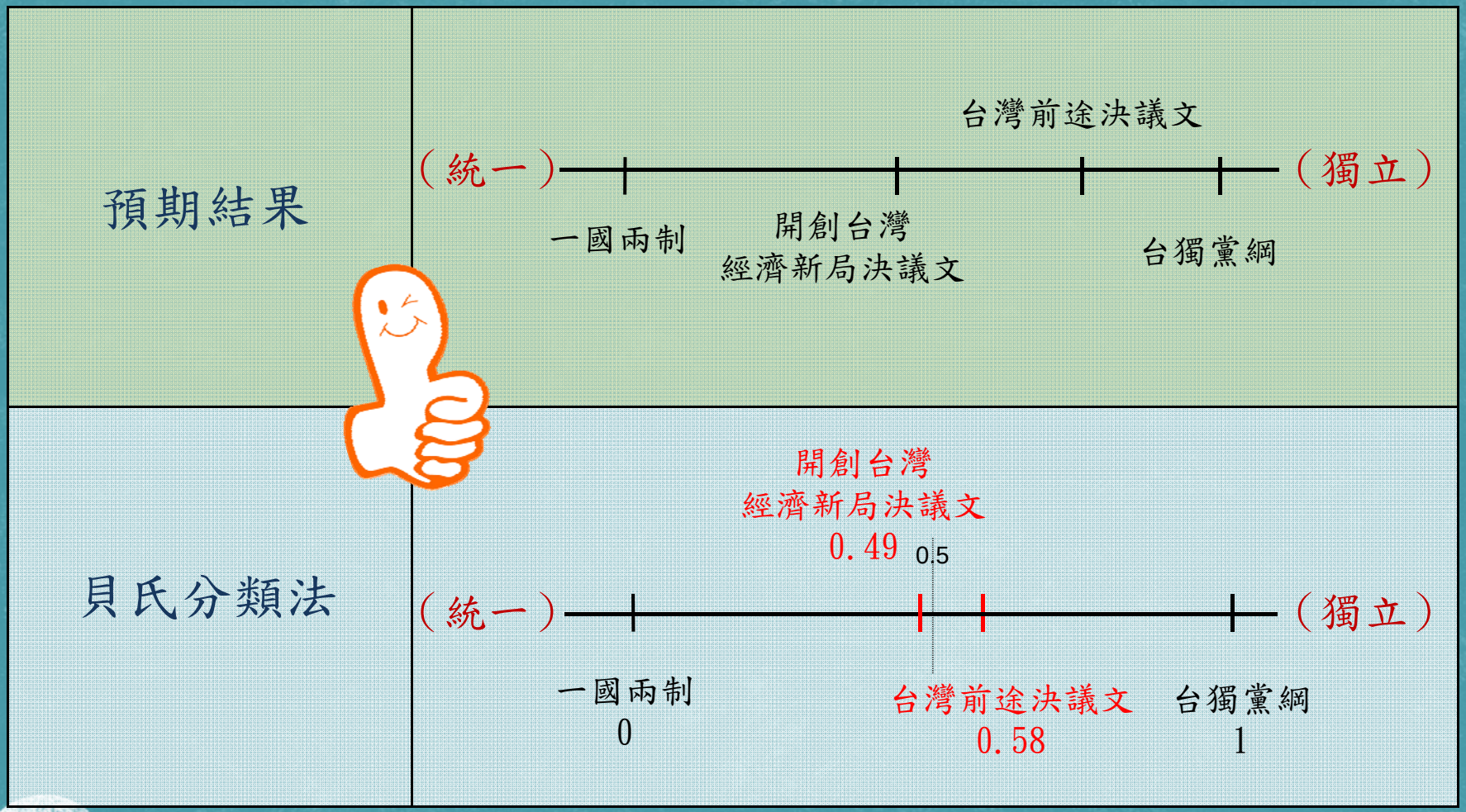
一國兩制(1982鄧小平)（極端統一）

測試文本：

台灣前途決議文(1999民進黨)

開創台灣經濟新局決議文(2001民進黨)

貝氏分類法測試結果


預期結果
台灣前途決議文
（統一）
（獨立）
一國兩制
開創台灣
經濟新局決議文
台獨黨綱
貝氏分類法
開創台灣
經濟新局決議文
0.49
0.5
（統一）
（獨立）
一國兩制
0
台灣前途決議文
0.58
台獨黨綱
1

貝氏分類法測試文本二

◆以近期的ECFA社論文本作測試

原始文本：

自投羅網只會讓台灣全盤皆輸(自由)**(極端反對)**

不要再用ECFA來撕裂台灣(聯合)**(極端贊成)**

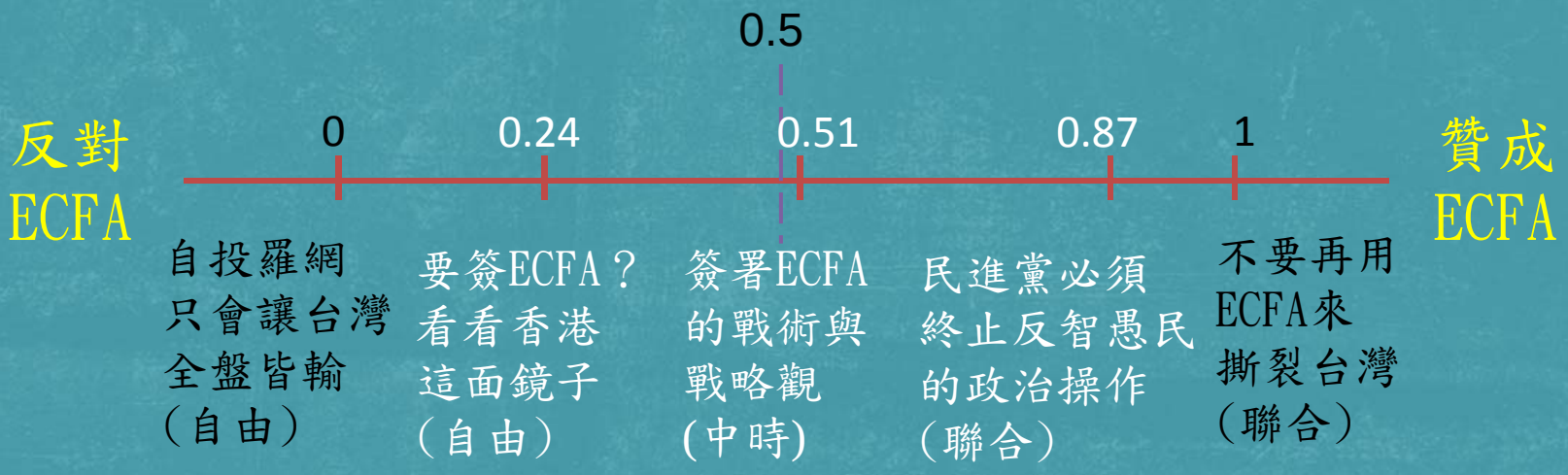
測試文本：

要簽ECFA？看看香港這面鏡子(自由)

簽署ECFA的戰術與戰略觀(中時)

民進黨必須終止反智愚民的政治操作(聯合)

ECFA社論測試結果

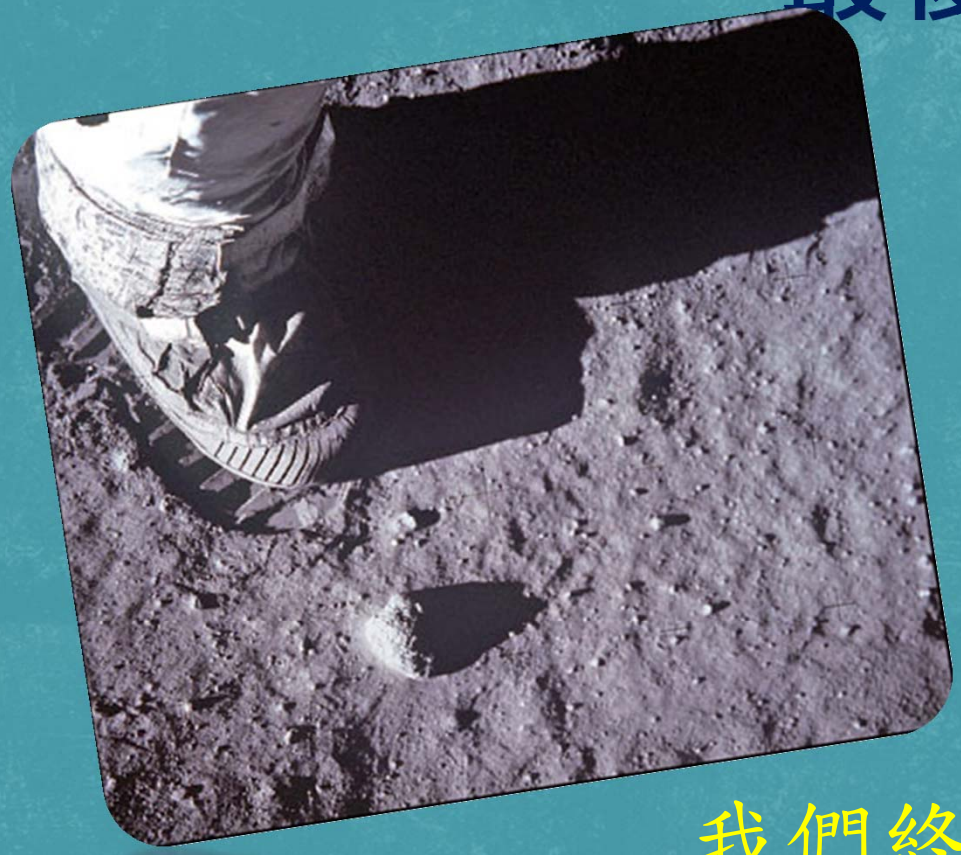


黑色字為原始文本
白色字為測試文本

結論與研究方向

- 貝氏分類法實驗結果跟我們預期結果相近，顯示「以詞為計算單位，用貝氏分類法加上關鍵詞在文章位置加權」的作法，可以運用於中文文章文字計數
- 未來方向
  - 利用字詞以外的文章資訊（例如：概念數、名詞數量、文章總字數、平均段落字數等）去突破原始文本必須是概念對立的文本限制
  - 深入分析文章內容根據關鍵詞的詞性（例如：名詞、動詞）找到更多關鍵特徵值線索

最後



THE END

我們終於...
為量化與質化研究的融合
踏出一小步！