期中報告

計畫名稱：
輿情分析與監督技術之研發開發計畫

報告人：

執行單位名稱：

目　錄

壹、計畫推動架構與目標

貳、計畫基本資料 & 專案組織架構

參、執行成果與目標達成情形

一、計畫執行進度

二、現階段進度展示及應用情境

三、遭遇困難及解決方案

四、期末成果展望

五、專利申請/獲得、技術移轉、論文發表、研發人才培育

肆、檢討與建議

伍、附件

壹、計畫推動架構與目標

- **目標**

  旨在透過應用文字探勘與視覺化設計的方法，為公司進行社會之輿情分析與監督機制的建立

- **推動架構**

貳、計畫基本資料 & 專案組織架構

- **計畫基本資料**
- 計畫名稱：「輿情分析與監督技術之研發」—產學合作計畫。
- 計畫合作單位：國立中山大學xx系。
- 計畫總主持人：國立中山大學xx系黃xx教授。
- 核心計畫團隊：

  國立中山大學資管系xxx博士生。

  國立中山大學資管系碩士生三人

貳、計畫基本資料 & 專案組織架構

- 專案組織架構

負責人	人數	職稱
計畫主持人	1	教授
計畫共同主持人	1	助理教授
博士研究生	1	博士生
碩士研究生	3	碩士生

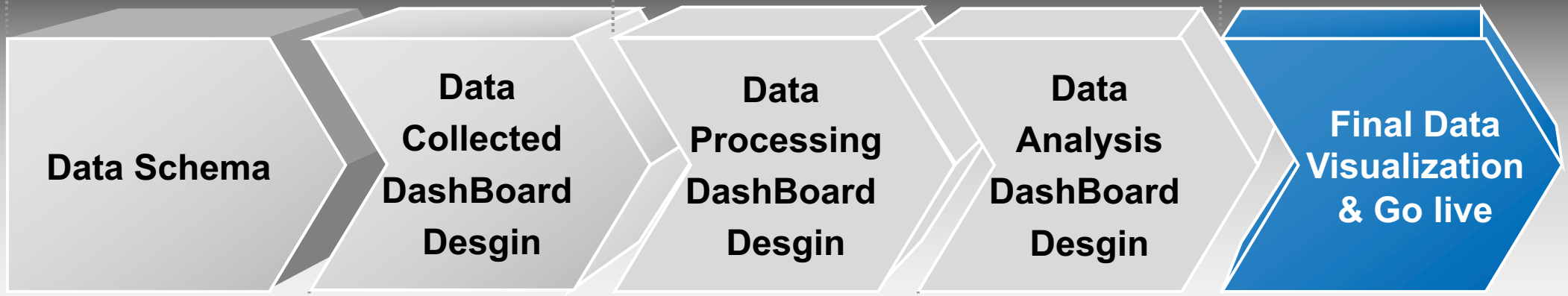
參、執行成果與目標達成情形- 計畫執行之進度

Social Listening & commander center Project Design

Data source and table Schema on MongoDB
due on 4/23/2015

Data Cleaning,
Data parsing.
DashBoard on MongoDB
Design the auto-scrape mechnisam
Due on 6/25/15

Dashboard design and data visualization due on 8/15/2015



1.Data Scraping
2.Server Migration
due on 5/21/2015

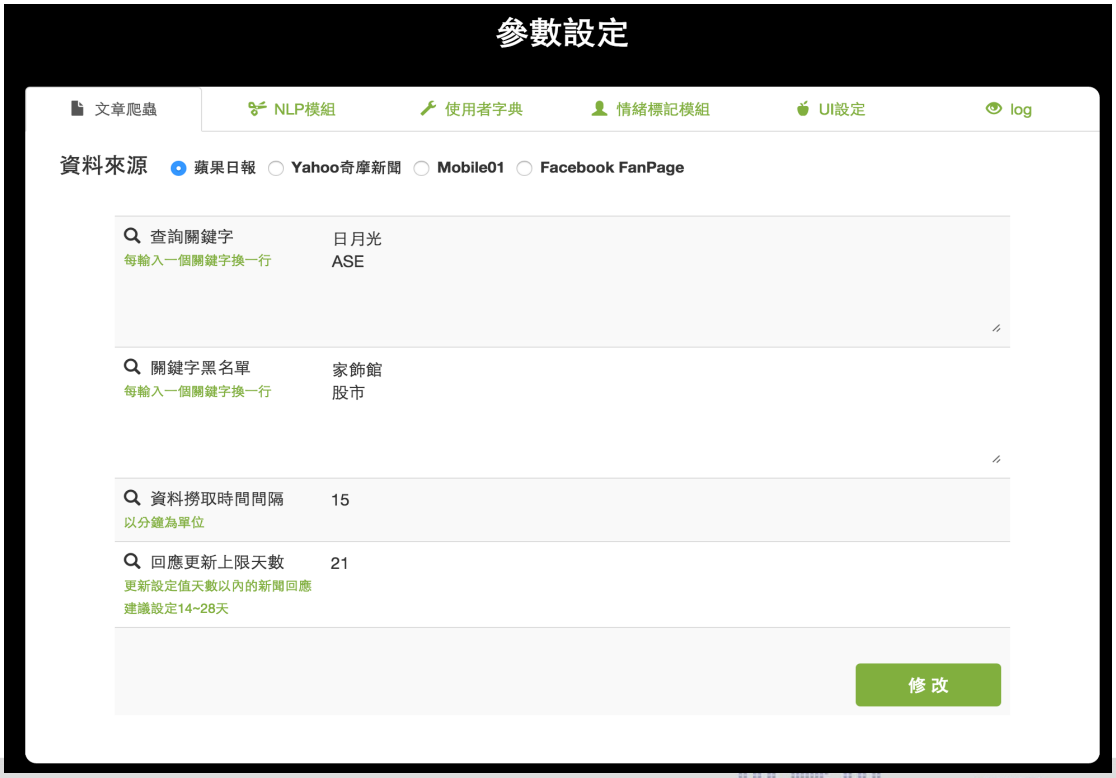
Topic Modeling,
Sentiment analysis,
Social network anlaysis
Visualization on all Analysis
Due on 7/31/2015

參、執行成果與目標達成情形

- 完成新聞子系統之進度
- 支援系統
  - 網路爬蟲程式之關鍵字設定
  - 自然語言處理(NLP)程式，包括斷詞和詞性標記
  - 自動擴充情緒字典
  - 自動排程網路新聞資料定時擷取程式
- 前端系統
  - Dashboard/戰情室看板之設計
  - 整合性視覺化圖表
- 網路資料來源
  - 新聞 (奇摩新聞，蘋果日報)

二、現階段進度展示及應用情境

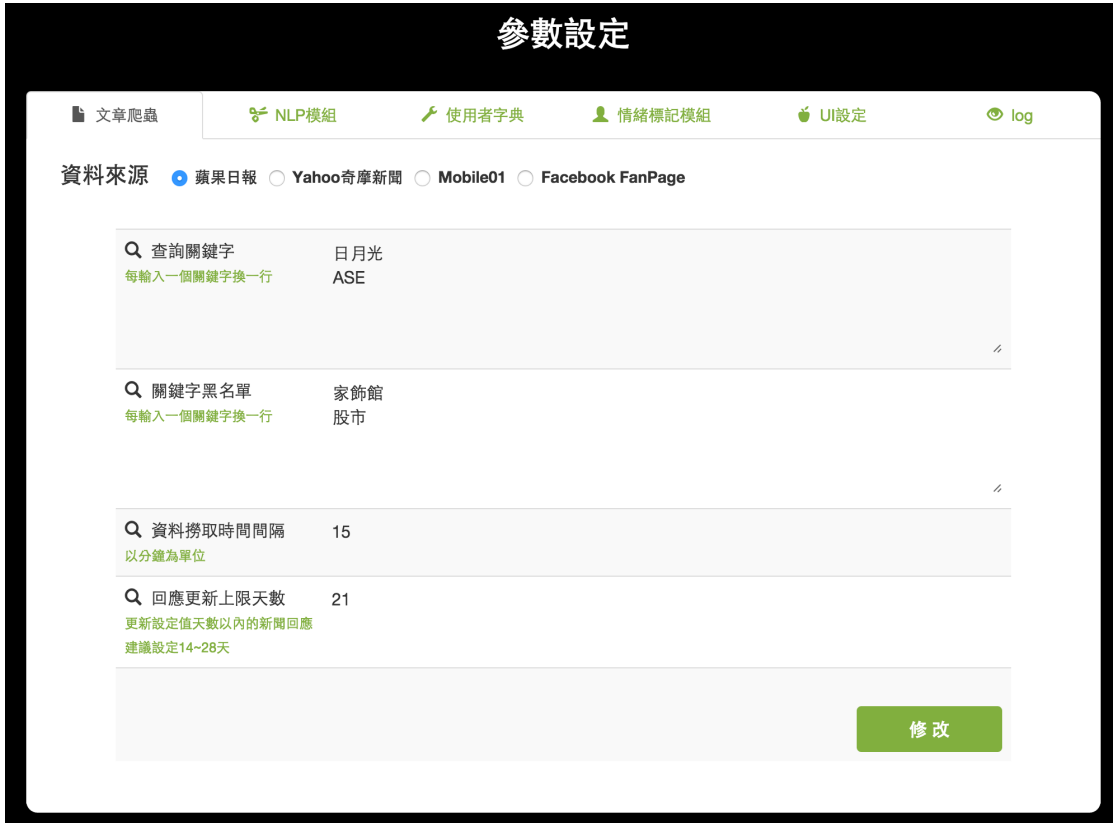
- 1. 支援系統
- 1.1.系統參數介面設定: 此系統用來設定所有系統模組之參數，增加系統彈性與自動化之設計介面。
- 1).文章爬蟲-關鍵字設定:用以設定爬蟲資料擷取之查詢關鍵字與去除關鍵字黑名單，以避免擷取到不相關之資訊



- 2). NLP模組：可設定文章句子斷句之標點符號與資料清除時停止詞

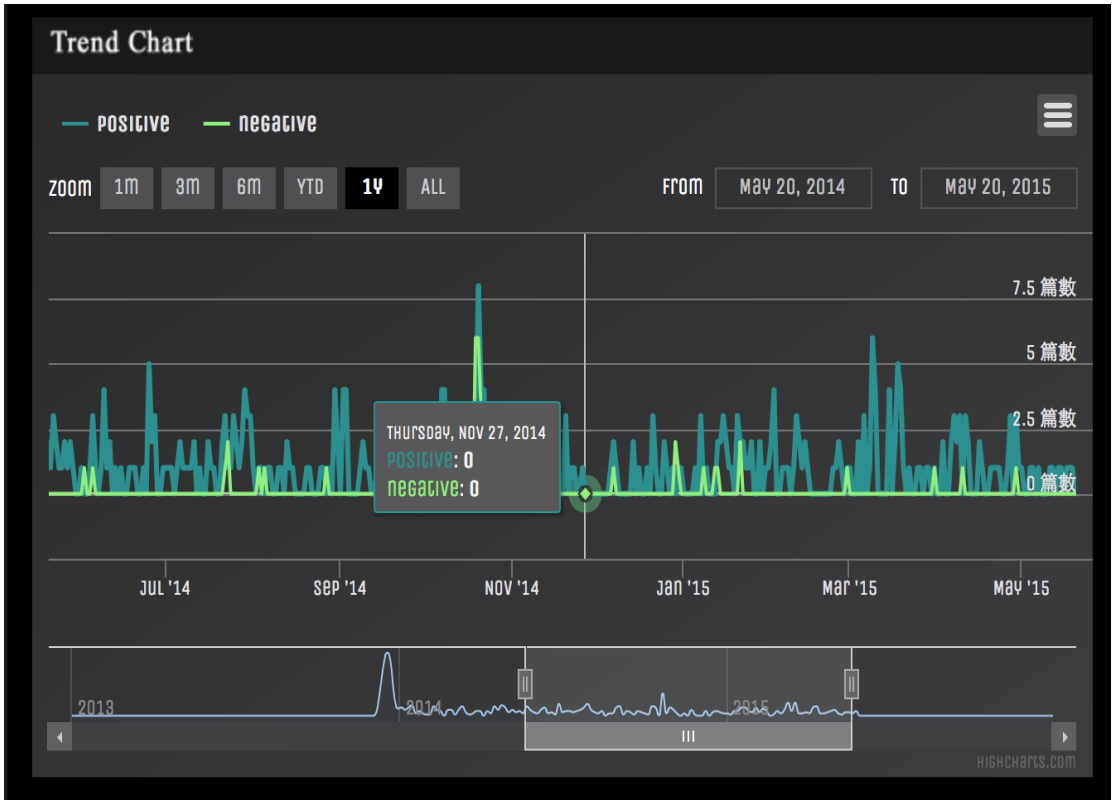
- 3). 使用者字典：為斷詞之需要，可新增相關新的詞彙以利斷詞之正確性

- 1.2.自動排程網路新聞資料定時擷取：可設定網路資料擷取之相關間隔之間與抓取有效天數

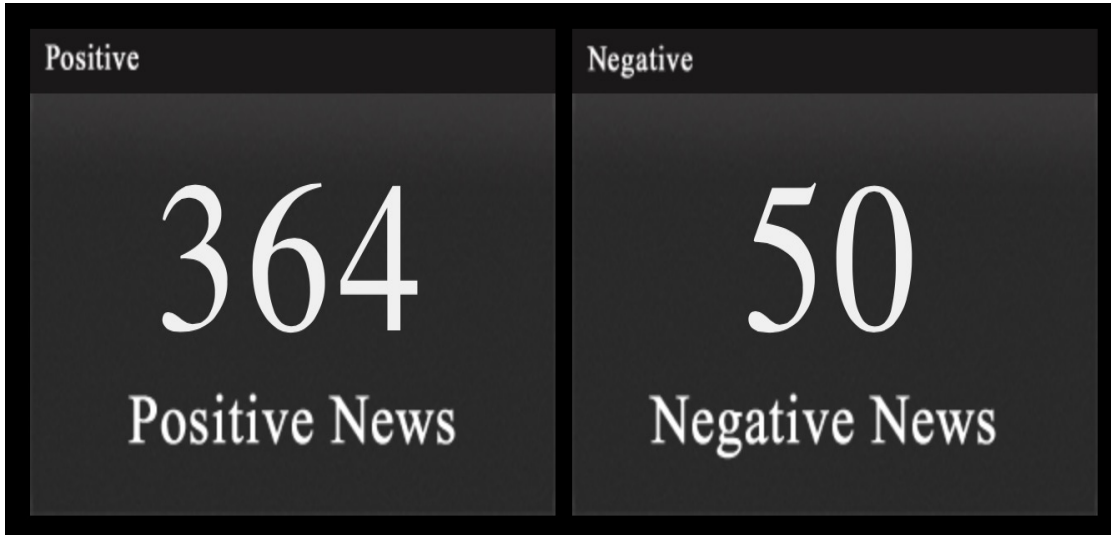


- 2.前端系統：利用highchart的工具來實作戰情室的看板設計，整合式的圖表設計讓資料分析更淺顯易懂。
- 2.1.Dashboard 戰情室看板分析

- **2.2.Trend Chart：以時間軸的方式來呈現新聞之情緒分析**



- **2.3. Trends Aggregate Sentiment for News：彙整在時間軸內出現的新聞數之正負面的評價**

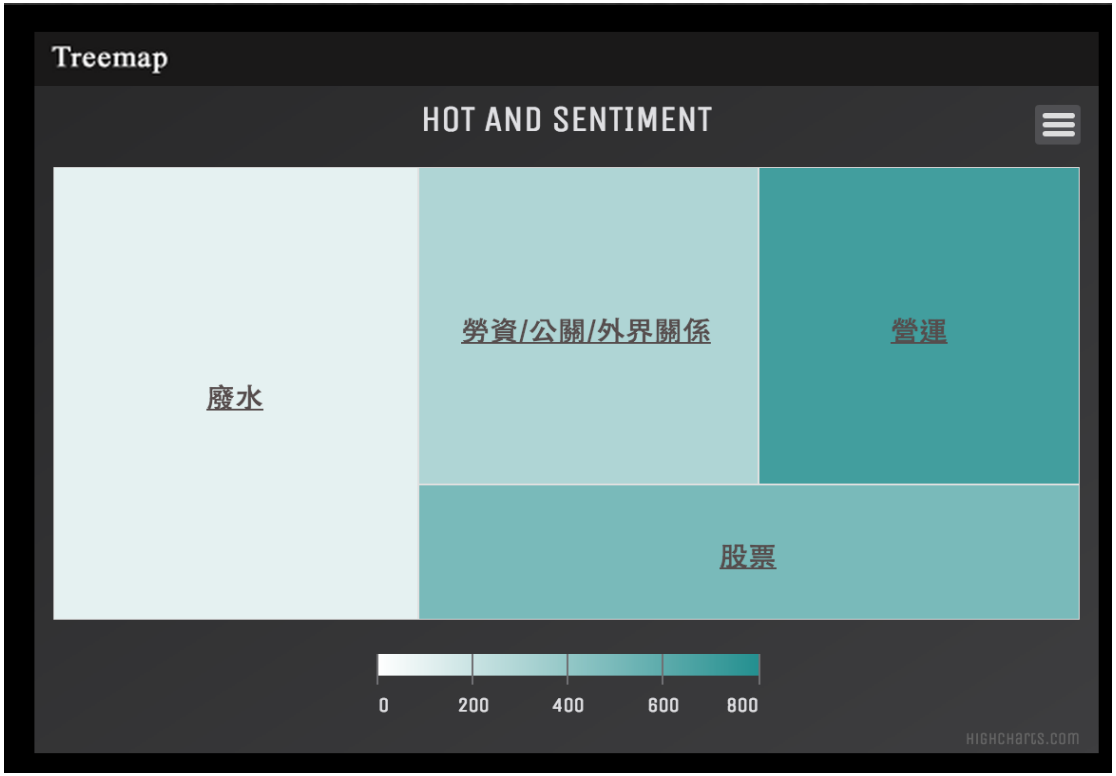


- **2.4. Opinion Leaders：依新聞回應之FB的按讚數與回應人數來計算意見領袖的排名**

Opinion Leader

leader	Like(%)	Cmt(%)	Cmt	Senti	Media	FB
陳村上	0	0.0833	1	5566	appleDaily	
叫什名	0	0.0833	1	5566	appleDaily	
林凡菱	0	0.0833	1	5566	appleDaily	
陳細安	0	0.0833	1	5566	appleDaily	
謝拓先	0	0.0833	1	5566	appleDaily	
葉俊麟	0	0.0833	1	5566	appleDaily	

2.5.Tree Maps:針對各項議題做熱度與情緒分析



- **2.6.WordClounds：針對目前新聞所有名詞出現頻率以文字雲呈現**

三、遭遇困難及解決方案

模組項目	遭遇問題	解決方案
網路資料擷取	關鍵字查詢需要排除不相關的資訊	參數設定關鍵字，排除不相關的資訊
NLP-斷詞	中文斷詞結果不理想	比較 CKIP & Jieba 之後採用 Jeiba 可擴充字典方式來斷詞可大幅提升斷詞準確率
NLP-詞性標記	文章句子標詞性與相依關係標註準確性不理想	經由 CKIP & Stanford Parser 比較結果後，採用 Stanford Parser 進行 POS 分析
情緒字典設定	情緒字典自動擴充功能為本系統之核心功能，但因詞性標記的不精確和中文特性，文獻上提到的一些情緒字典擴充方法效果都有限	目前使用 unsupervised learning 的方式進行此，目前的演算法是採用使用 pagerank 的方式來進行情緒字典的擴充，並採用多種方式來進行成果比較，以找出最佳之演算方式來進行情緒辭典之擴充
視覺化設計 Trend Chart	Trend Chart 在連動時間抓取資料時，時間過長，而且連動時間區段無法有效呈現	重新利用 highchart 的設計來存取資料與利用 trigger event 遞延時間的方式，以解決此問題
視覺化設計 Opinion Leaders	使用 php 跑迴圈抓取計算會使效率非常差，耗時過長需要改善	重新利用 mongoDB 特性來抓取資料，省下迴圈之計算，效能大幅提升，圖表整合流暢。

四、期末成果展望

- 期末成果包含：
- 完成可設定關鍵字與時程的網路資料擷取的技術：資料包含針對國內主要新聞媒體（蘋果日報、YAHOO奇摩新聞）。
- 抓取社群媒體與線上論壇如(PPT、FaceBook之公開社團、Yahoo奇摩知識、Mobile01)等網站資料並分析，
- 完成文字資料清理之設定，斷詞與標詞性的API介面，亦將技術移轉給公司MIS人員。
- 完成文字資料分析：設定議題與追蹤分析、社會輿情分析、社會網路分析等三大方法
- 完成戰情室看板之圖形化網站：多維度的資料分析呈現，重要議題追蹤與社會輿情分析之圖表與相關細項資料查詢之介面。
- 相關技術移轉給公司MIS人員。
- 完成計劃期末報告一份。

專利申請/獲得、技術移轉、論文發表、研發人才培育

項目	績效指標		預估數	實際數	備註
專利申請/獲得	國內		0 件	0 件	
	國外		0 件	0 件	
技術移轉			0 項	0 項	
論文著作	國內	期刊論文	0 件	0 件	
		研討會論文	0 件	0 件	
		SCI論文	0 件	0 件	
		專書	0 件	0 件	
		技術報告	0 件	0 件	
	國外	期刊論文	0 件	0 件	
		學術論文	0 件	0 件	
		研討會論文	0 件	0 件	
		SCI/SSCI論文	0 件	0 件	
		專書	0 件	0 件	
		技術報告	0 件	0 件	
人才培育	博士生		0 人	0 人	畢業任職於業界：0 人
	碩士生		0 人	0 人	畢業任職於業界：0 人
	其他		0 人	0 人	畢業任職於業界：0 人

肆、檢討與建議

- 輿情分析為本計劃之核心模組，但因中文情緒分析在目前的研究領域上尚未非常成熟，所以本團隊嘗試各種不同之演算法來改善其準確率，因此必須花費較長的時間，才能調整至最佳的成效。
- 視覺化設計的部分，也因為需要結合使用mongoDB與highchart的新技術，加上時間軸連動的方式，為求最完美的呈現方式，因而投了許多時間去重新設計查詢與整合畫面的流暢與清晰度。
- 本計劃每週都有定期開會討論，使得計劃之進度皆符合預期規劃，且相關同仁也積極參與，並共同探討新技術與演算法之應用方式，另外每週之會議紀錄與相關開會資料皆有完整資料可備查

伍、附件

文件清單	說明
開發環境 **Tool List**	所有功能的開發環境工具清單
硬體設備需求	**CPU/Memory/HD** 最低硬體設備需求
Coding& **Algorithm Introduction by Module**	所有功能的 **Coding** 與 **Algorithm** 簡介
Deploy SOP	**Server** 與程式安裝的 **SOP for ASE** 環境安裝

Thank You

www.aseglobal.com