文字探勘初論 期末專案計畫書

第 4 組

專案名稱

產業新聞分類與關鍵字檢索

組員名單

R09725015 郭太元
B11705034 蔡逸芃
B10705054 劉知微
B11705027 陳承妤

計畫概述

- **資料來源：**網路爬蟲，取得有興趣的產業之新聞，轉換為 TF-IDF 向量並標記產業 (如爬取 semiconductor news，標記 semiconductor 類)。若文章品質不佳則使用現成資料集。
- **產業新聞分類：**使用監督式學習 (如 NB、KNN、SVM 等模型)，將新聞以產業別分類。
- **評量分類成效：**使用兩種測試方式，第一種為「在訓練類別之中的產業新聞」，直接將第一點的資料分割為訓練與測試資料 (如 semiconductor news 以 9:1 切分)。第二種為「不指定產業之產業新聞」，額外爬取不指定產業的 industry news，轉換為 TF-IDF 做為測試資料；需要手動標記類別，並可能會有落於分類之外的文章。
- **關鍵字檢索：**新聞文章進行分類後，使用 LSA 對各類別的文章提取關鍵字。若能取得文章發布時間，還可以針對不同時間範圍的文章檢索關鍵字，觀察不同時期關鍵字的變化。