Agenda

1. 選股依據
2. 資料前處理
3. 關鍵字萃取
4. 分類模型

Agenda

1. 選股依據
2. 資料前處理
3. 關鍵字萃取
4. 分類模型

本組報告選用長榮海運作為期中報告之公司，係因討論聲量大及其為全球指標性的海運公司

長榮海運
EVERGREEN

本組選擇 2603 長榮
作為本次期中報告分析公司

討論聲量大

長榮海運 2022 年豪發 45 個月年終獎金 引起大量民眾熱烈討論與關注

全球指標性海運公司

長榮海運在 2022 Q1 全球貨櫃運力中排名第 6，而過去也都保持在前十名內

Agenda

1 選股依據

2 **資料前處理**

3 關鍵字萃取

4 分類模型

資料前處理｜先將文章集以及股價資料篩選出關於長榮海運之資料以及股價

時間區間：2022 年 1 月 1 日至 2023 年 3 月 24 日

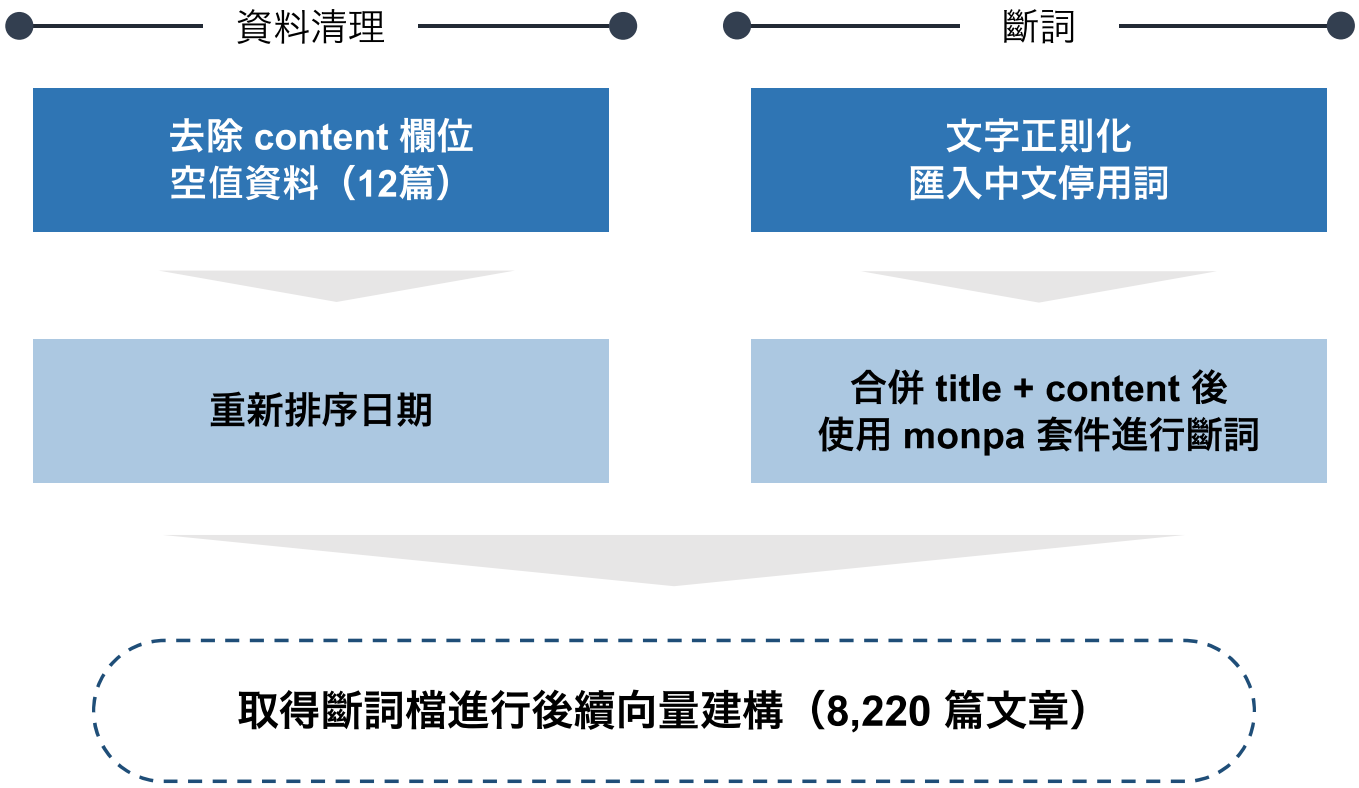
文章集 （11142 篇文章）	股價資料
合併網路公開之 新聞、論壇、BBS 文章	股市價量交易資訊
選取文章 title, content 中 含「長榮」的文章	選取 長榮 (2603) 股價資訊

資料前處理｜根據文章發布日期後一天之股價變化標註於該篇文章中

漲	3,964 篇	• 當日漲幅 > 0 %，將**前一日文章**標記為「漲」
跌	4,267篇	• 當日漲幅 < 0 %，將**前一日文章**標記為「跌」

以標籤為漲、跌之文章集作為原始的資料集（共 8,231 篇文章）

資料前處理｜去除缺失值以及將文章內容進行斷詞以進行後續向量空間之建構



Agenda

1 選股依據

2 資料前處理

3 **關鍵字萃取**

4 分類模型

本組採用三種關鍵詞萃取方式，其中以 Chi-Square 表現最佳，因此將結果建構向量投入模型

TF-IDF
max_features = 5000

Chi-Square
K = 5000

Chi-Square
K = 1000

本組嘗試三種關鍵字萃取方式選取較具有鑑別力的關鍵詞
最後選擇以 Chi-Square 結果建構向量投入模型判斷準確率

Agenda

1 選股依據

2 資料前處理

3 關鍵字萃取

4 **分類模型**

將 2022 年的文章作為訓練資料，而 2023 年至今之資料作為測試資料以進行模型準確判斷

訓練資料

2022/1/1 – 2022/12/31
共 7,301 篇文章

測試資料

2023/1/1 – 2023/3/21
共 919 篇文章

將測試資料斷詞結果映射至訓練集向量空間中

接著我們採用六種不同的預測模型進行訓練與預測，其中以 Decision Tree 之準確率表現最佳

Decision Tree

準確率：54%

	真實為漲	真實為跌
預測為漲	274	180
預測為跌	255	210

Random Forest

準確率：53%

	真實為漲	真實為跌
預測為漲	300	280
預測為跌	154	185

KNN

準確率：51%

	真實為漲	真實為跌
預測為漲	229	224
預測為跌	225	241

Naive Bayes

準確率：50%

	真實為漲	真實為跌
預測為漲	244	221
預測為跌	237	217

Logistic Regression

準確率：49%

	真實為漲	真實為跌
預測為漲	272	291
預測為跌	182	174

SVM

準確率：47%

	真實為漲	真實為跌
預測為漲	347	378
預測為跌	107	87

後續調整方向

1 向量空間重新調整

可針對現有斷詞表現進行重新調整，以產出模型配適更加的預測效果

2 資料搜集更加全面

現有資料可能僅涵蓋文章類型相關資料，但影響股價的因素不僅只是輿情分析，更多的是包含多面向之市場趨勢現況、公司前景，甚至是公司之股利政策相關要素。因此如果要建構出預測力更佳的股價預測模型，應考慮更加全面影響股價之要素

影片解說檔案

影片連結	https://drive.google.com/file/d/15nC2UV66QA_zHIZPHVCYtakzAvP7P6tO/view?usp=share_link