國立臺灣大學資訊管理學系 2018學士班資管專題

自動化財經新聞篩選 - 以原油期貨為例

專題學生：陳欣宜、胡逸凡
專題指導教授：李瑞庭 教授

摘要

投資人每日要收集大量的財經新聞作為交易分析的參考，重要新聞網站如彭博、路透社、CNN、BBC...每日的新聞量加起來超過百則，**該如何讓投資人掌握最即時、重要的新聞**，就是本專題最主要想解決的問題。以下，我們以原油期貨的新聞資料為例，進行文字探勘，從資料收集、處理、分析到最後的建立模型，希望能達到自動化財經新聞篩選的效果。

資料收集

爬蟲程式嘗試不同來源的原油歷史新聞語料，調查更具威信力的新聞來源。最後使用以下三個網站 2017 年 10 月至今的新聞：

- EBSCO Host 共 1308 篇，平均每篇新聞約 3500 字
- OilPrice.com 共 2218 篇，平均每篇新聞約 4000 字
- CNBC news 共 907 篇，平均每篇新聞約 2700 字

數值資料：2017年10月至今的WTI原油期貨每日開盤價、收盤價、交易量

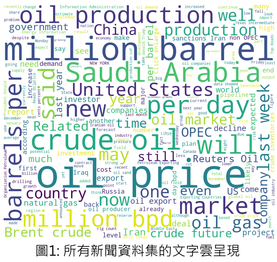

圖1: 所有新聞資料集的文字雲呈現

文字處理

根據我們的觀察以及文獻參考，單篇新聞若字數太多，可能反而會成為noise，降低預測準度，因此我們嘗試用 summarize 與 不 summarize 兩種方式，觀察是否字數會是造成模型好與壞的因素之一。

- 新聞前處理
  - 參數一：是否做 stemming
  - 參數二：是否做 summarization
    - Summarize方法：利用 python gensim 套件裡的 summarize 函式，每篇文章取200字摘要。
- Feature type
  - 文章的詞向量空間：Unigram, Bigram, Bigram with windows rolling
  - 情感分析：positive score, negative score
  - 石油價格歷史資料：過去一個月價格平均、過去一個月標準差、過去3,5,10日價格平均及標準差

重要新聞定義

對於投資人而言，他們最注重的還是金融商品的價格波動，因此我們在定義重要新聞時，也主要以該篇新聞是否會造成「期貨價格顯著波動」為準。我們試用了以下方法進行自動標記：

- 只考慮波動性：以標準差為主要判斷依據
  - 方法一：未來n天的開盤價標準差 > Threshold
  - 方法二：未來n天的開盤價標準差([+1,+n]) > Threshold * 過去m天的開盤價標準差([-m,-1])

圖2：近兩個月的原油價格。上方箭頭指向為利用標準差方法2所標記出來的答案

- 考慮波動性與上漲下跌：
  - 方法一：當未來n天的漲幅 > Threshold
  - 方法二：除了符合Threshold之外，加上文章的情感分析作為標記標準。一天中的新聞正負向參雜，但多數投資人較在乎的，是會準確造成後續價格走向的新聞。在計算 sentiment 的部分，我們使用以下兩種方法：
    - 自定義字典的正負向字加總
    - 使用 nltk 套件中的 SentimentIntensityAnalyzer 計算正負向、以及不確定性分數

建立字典

為了找到有顯著代表性的字詞，我們將目標文章與其他文章字詞的差集視覺化，發現其中的確有顯著差異，且利用 Bigram 可看出這些選出來的字詞很多都是財經市場上會遭成波動的議題，例如：中美貿易戰爭、川普、頁岩油作為字典。

建立模型與預測

利用上述字典所建立的詞向量空間，設計以下幾種模型：

- SVM(Support Vector Machine)
- Naive Bayes
- Random Forest

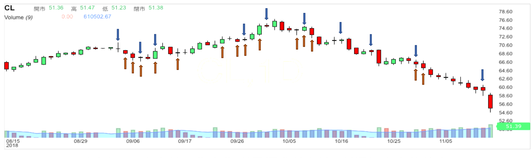

圖3：預測結果。上方箭頭指向為利用標準差方法2所標記出來的答案，下方箭頭指向為預測為重要文章的出現日期

結論

現階段已經可以建構有意義的 bigram 字典，但預測出來的文章結果分布還有很大進步空間。目前選出來的文章，可以看出大多集中在重要新聞應該被標註的日期附近，如果用交易日當作判斷得分基準的話，可以得到比較好的結果，可以理解為一天當中真正影響價格的文章並不多，但也許只要一兩篇就會影響很大，而且我們的資料集不夠大，學習的對象有限，以下幾點構思:

1. 更多元的資料集，不能只使用單一來源，但難處在，越大的資料集裡面垃圾也越多，需要”乾淨的”資料集難度很大
2. 標註的方法需要再試更多種可能性
3. 可以縮小範圍，限制與某主題相關(新聞進來之前先用分群)，例如中美貿易戰，的新聞，再給予預測。
4. 逐漸建立原油相關字典，慢慢完善

國立臺灣大學 資訊管理學系
Dept. of Information Management
National Taiwan University