Text Mining : Home Work 1

Student : 610721204　陳克威

GitHub Repository URL:

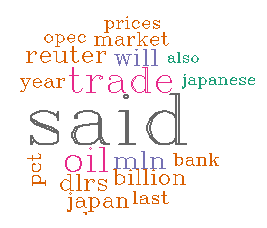
<https://github.com/D1034181036/Reuters_21578>

1. 資料集Data Set：

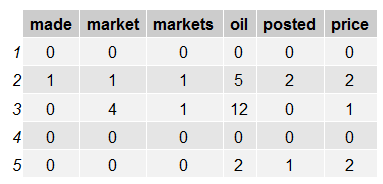
此資料集為一新聞通訊社Reuters在1987年發佈的新聞資料，Steve Finch與David D. Lewis在1996年清理了資料，清理過後總共有21578份新聞文件，因此被稱為Reuters-21578，另外，此資料集也被收錄在UCI Machine Learning Repository，其預設目標為分類(Classification)問題。

Fabrizio Sebastiani將其依照常出現的主題分類成8類與52類，被稱為R8與R52，其中R8的訓練資料有5485份文件、測試資料有2189份文件，總共為7674份文件，本研究使用R8做為資料集。

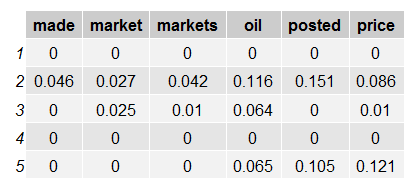
1. 資料前處理Data preprocessing：
2. 將訓練資料5485筆與測試資料2189筆合併（為了建立語料庫）。
3. 因為計算量較大，這裡取其中的三個類別作為子集，分別為[trade, crude, money-fx]，取完子集後的資料為710筆訓練資料，與283筆測試資料，總共993筆資料。
4. 將類別的階級大小(Levels)去除。
5. 將資料建立成語料庫(Corpus)，共有9243個Term。
6. 語料庫：去除標點符號。
7. 語料庫：去除數字。
8. 語料庫：去除多餘的空白符號。
9. 語料庫：去除英文停詞。
10. 語料庫：將所有英文轉換為小寫。
11. 文字雲呈現：最常出現的20個文字。



1. 將語料庫轉換成Document Term Matrix的形式。



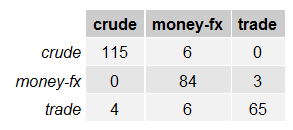
1. 計算此Document Term Matrix的TF-IDF權重值，將其轉換為 Weight-TF-IDF版本的Document Term Matrix。



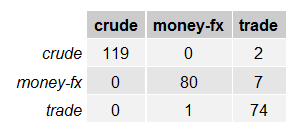
1. 將Document Term Matrix拆分回訓練資料與測試資料。
2. 建立分類器(Classifier)模型 Modeling：

本研究總共建立與比較了四種分類器，分別為：

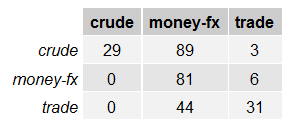
* 1. Decision Tree (CART)



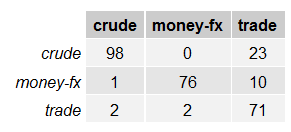
* 1. Support Vector Machine



* 1. Naive Bayes

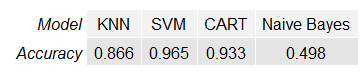


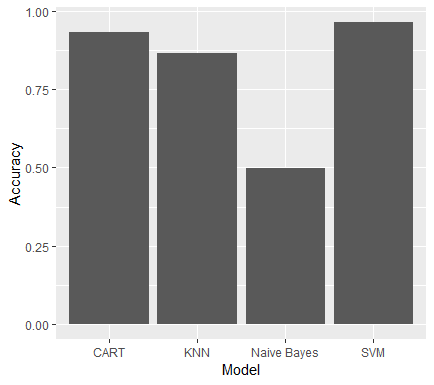
* 1. K Nearest Neighbor



1. 實驗成果 Experiment

以下為本次實驗結果，SVM與Decision Tree的表現較為出色，九成以上的文件都能夠被正確分類，而Naive Bayes的表現較不盡人意。





在本次實驗中，我們使用三個子集做為樣本集做為測試，大約為1000份新聞資料，若在沒有分為子集的情況下，使用一般個人電腦訓練可能會遇到Document Term Matrix過大導致RAM不夠用的問題，這時可能會需要做分批訓練或是平行運算，希望未來能夠進一步使用大型的資料集進行實驗與實際應用。

1. 參考資料

1. Lewis, D.D. (1997). Reuters-21578 Text Categorization Test Collection, Distribution 1.0.

2. Ian Kloo(August 2015) Textmining: Clustering, Topic Modeling, and Classification : <http://data-analytics.net/cep/Schedule_files/Textmining%20%20Clustering,%20Topic%20Modeling,%20and%20Classification.htm> (last access:2019/06/11)

3. Bryan Cole(August 14, 2016) Classifying Documents in the Reuters-21578 R8 Dataset <https://rpubs.com/bmcole/reuters-text-categorization>

(last access:2019/06/11)

4. UCI Machine Learning Repository - Reuters-21578 Text Categorization Collection Data Set : <https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection> (last access: 2019/06/11)