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。

2. 資料前處理 Data preprocessing：

(1)將訓練資料 5485 筆與測試資料 2189 筆合併（為了建立語料庫）。
(2)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。
(3)將類別的階級大小(Levels)去除。
(4)將資料建立成語料庫(Corpus)，共有 9243 個 Term。
(5)語料庫：去除標點符號。
(6)語料庫：去除數字。
(7)語料庫：去除多餘的空白符號。
(8)語料庫：去除英文停詞。
(9)語料庫：將所有英文轉換為小寫。

(10) 文字雲呈現：最常出現的 20 個文字。

(11) 將語料庫轉換成 Document Term Matrix 的形式。

	made	market	markets	oil	posted	price
1	0	0	0	0	0	0
2	1	1	1	5	2	2
3	0	4	1	12	0	1
4	0	0	0	0	0	0
5	0	0	0	2	1	2

(12) 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。

	made	market	markets	oil	posted	price
1	0	0	0	0	0	0
2	0.046	0.027	0.042	0.116	0.151	0.086
3	0	0.025	0.01	0.064	0	0.01
4	0	0	0	0	0	0
5	0	0	0	0.065	0.105	0.121

(13) 將 Document Term Matrix 拆分回訓練資料與測試資料。

3. 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)

	crude	money-fx	trade
crude	115	6	0
money-fx	0	84	3
trade	4	6	65

(2)Support Vector Machine

	crude	money-fx	trade
crude	119	0	2
money-fx	0	80	7
trade	0	1	74

(3)Naive Bayes

	crude	money-fx	trade
crude	29	89	3
money-fx	0	81	6
trade	0	44	31

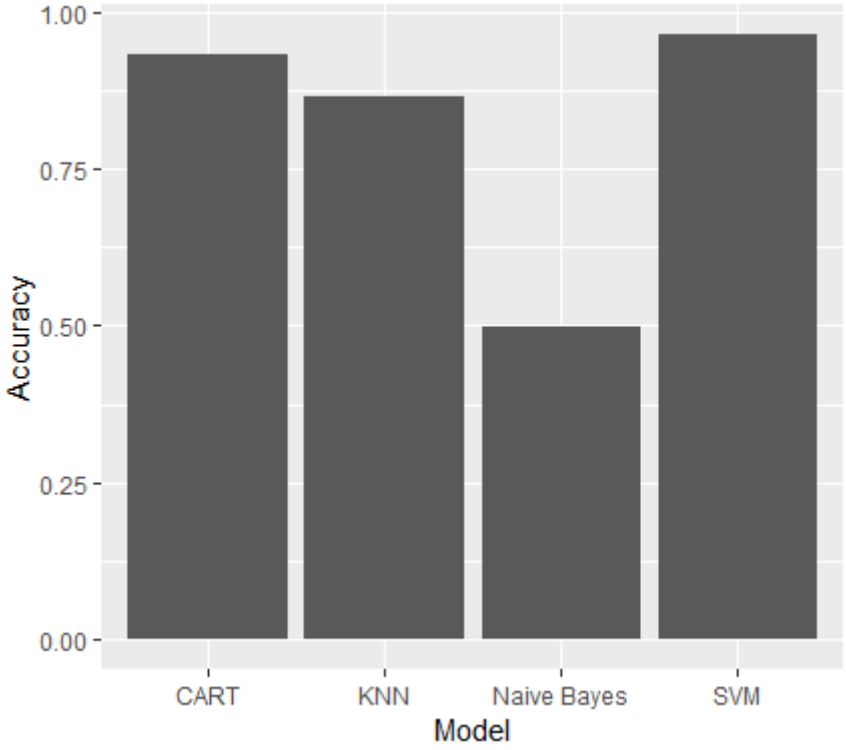
(4)K Nearest Neighbor

	crude	money-fx	trade
crude	98	0	23
money-fx	1	76	10
trade	2	2	71

4. 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。

Model	KNN	SVM	CART	Naive Bayes
Accuracy	0.866	0.965	0.933	0.498



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。

5. 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)