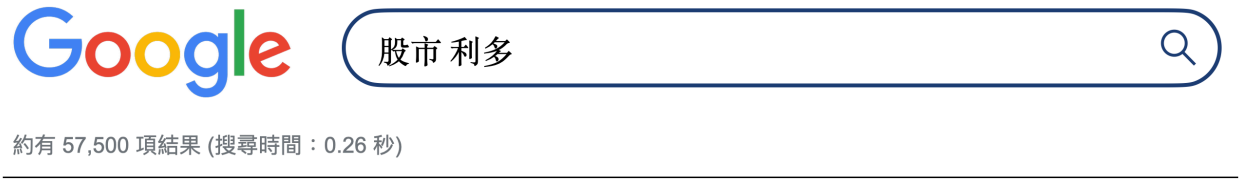
NLP

Summary：

我主要負責新聞情感分析的部分。對於新聞情感分析。我們主要是透過機器學習的方法，來判斷新聞內容的情感。



Process：

首先，我們所爬取的新聞資料來源主要有YAHOO新聞網、鉅亨網、中時電子報與自由時報聯合報以上五個。先透過從GOOGLE收尋引擎輸入關鍵字，再進入各分頁的新聞當中，將其股票代碼、公司、日期、標題、內容和來源爬取下來並做清洗，最後再儲存在MongoDB當中。

接著，我們從MongoDB中匯出正負面新聞各約8000篇，當作訓練資料做模型。將新聞內容進行斷字斷詞，我們所用的套件為Jieba，由於Jieba斷詞的系統主要是簡體字，因此匯入自訂義字典方便新聞斷字斷詞，並將特殊字元及沒意義的單詞，用停用字典給過濾掉。

模型貼標：

將正負面新聞斷字斷詞結果收集成一個word_index字典。
word_index字典的主要用途，是要將新聞內容結果，從文字轉成數值，而所用的特徵。這樣有利於機器學習模型去判讀。
文章斷字斷詞結果只要有出現在word_index字典中設為1，而沒出現設為0。並在正面文章特徵後面用1作貼標，負面文章特徵後面用0作貼標。

Positive

Negative

最後，將訓練資料拿去模型進行訓練，所選用的模型為監督式機器學習模型，有羅吉式回歸、SVM及貝氏分類三種。接著再對測試樣本用模型來預測其結果。
從中可以看出SVM模型對測試樣本來預測其結果的準確較高，因此我們用此模型來當作判斷各股票新聞情感結果的依據。

透過判斷各股票新聞的正負面，將此股票新聞的情感分析結果做個平均，此即為此股票的新聞面的分數，最後再將此分數儲存在Redis當中。

主講人：楊凱欽

https://github.com/s1033500