**310706034 資管碩一 資料探勘研究與實務 HW2 - sentiment analysis**

**1.a資料前處理**



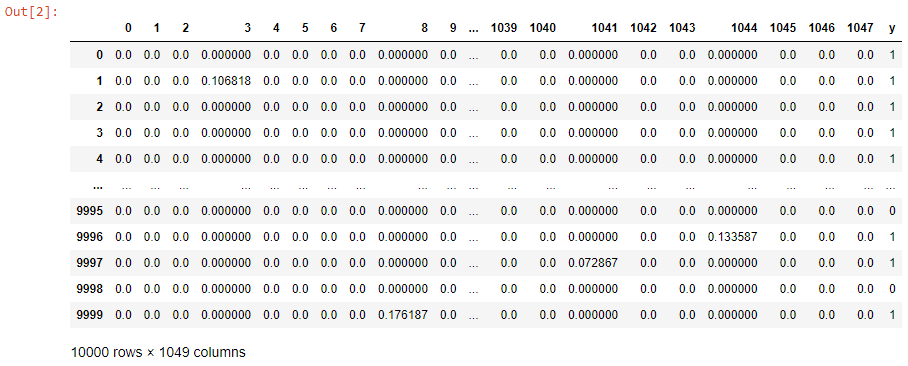
**1.b去除停頓詞**



**1.c 生成tf-idf文字向量**



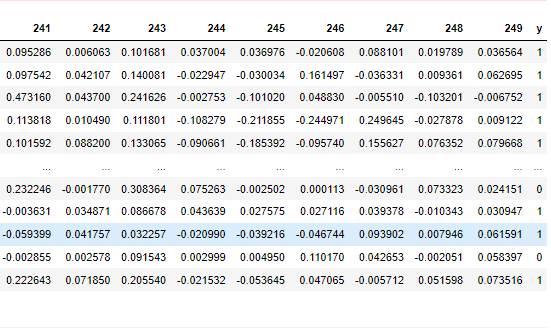
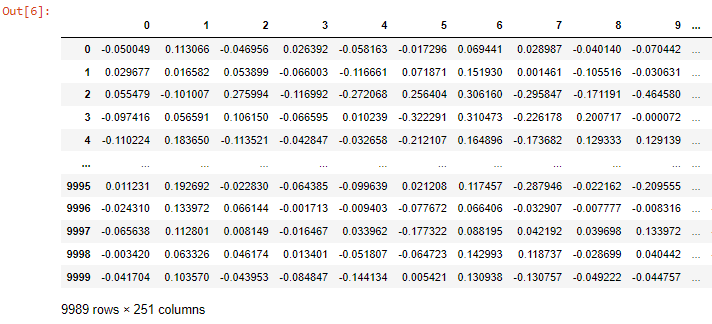
tf-idf文字向量如下，min\_df設為0.01，得到1047個字詞作為解釋變數



**1.c 生成Word2vec文字向量**



Word2vec文字向量如下，vector size設為250，得到250維的解釋變數:

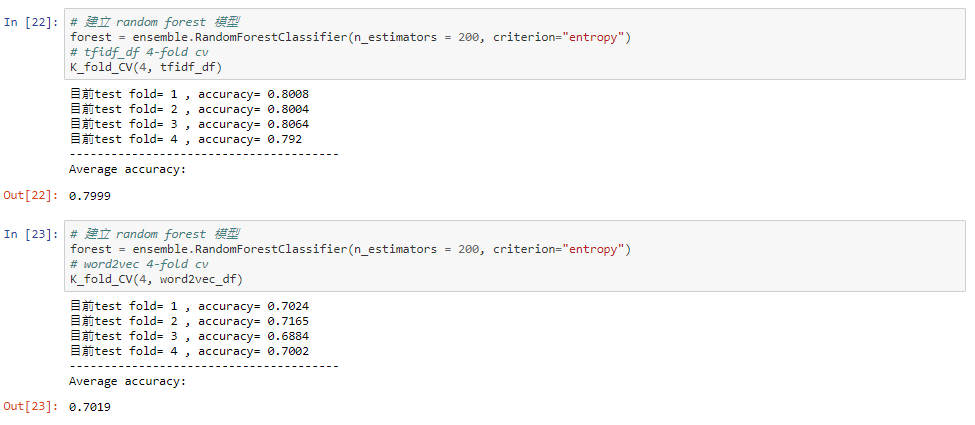


**2.K-fold CV Function & random forest**



**3.預測結果**

皆使用兩百棵樹作為隨機森林參數



tf-idf 的平均準確率為0.7999、Word2vec的平均準確率為0.7019，可能原因是Word2vec有一些字詞並沒有word vector，導致資訊量較tf-idf低，造成平均準確率較低。