

## HW5 Articles Clustering & Classification

Data pre-processing:

1. 使用 Pandas 讀取檔案
2. 使用 for 迴圈讀取 dataframe 中每個 row 的值（每篇文章的內容），使用 jieba 斷詞（使用範例 stopwords，取前 30 關鍵詞），存成 list
3. 再使用一個 for 迴圈將前步驟的 list 一個一個詞讀出，用空白字元間隔，存成一個 string
4. 跳出前步驟迴圈後 將該 string 存到另一個 list (shape 為(data 筆數,1))
5. 使用 scikit-learn 的 TfidfVectorizer 將前步驟存 string 的 list 做為輸入，計算 tf-idf 值，回傳一 sparse matrix

Experiment Result:

- K-Means Clustering (k = 5)

```
K-Means Clustering Result:  
[4 3 0 0 4 3 0 0 0 4 0 3 0 3 0 0 2 4 3 0 0 4 0 0 0 2 1 0 0 2 3 0 0 4 0 0 4  
0 2 0 1 3 4 0 4 2 3 0 4 3 0 2 0 0 3 4 4 2 0 0 1 0 3 0 0 4 0 4 1 1 2 0 1 0  
0 2 2 2 3 1 4 0 0 2 4 1 3 0 0 1 4 3 1 0 0 1 1 0 1 2 4 4 1 2 0 0 4 3 3 3 2  
0 1 2 1 4 0 1 0 2 0 2 3 0 1 3 3 0 0 2 4 0 3 0 3 0 0 0 3 3]
```

- Decision Tree Classification

```
Decision Tree Classification Result:  
Accuracy: 71.4285714286 %
```