HW4-Report

資管碩二 R10725046 江雨柔

作業四統整

本次作業主要實作一個 HAC。

1. 環境

執行環境 : VScode

2. 程式語言：python 3.8.2

3. 執行方式

```
# 需要下載的環境
pip install gensim
pip install nltk


# 下載 nltk.stopwords
!python -m nltk.downloader stopwords
!python3 -m nltk.downloader stopwords

# 執行方式
# 可以透過 VS code 一格一格跑（run below）
# 或是直接按 Run 全部一起執行
```

4. 作業處理邏輯說明

第一部分 - 製作所需計算之 dictionary

1. 先將每個文黨都做成 term，利用作業一的模式製作每個文檔的 terms
    a. 將標點符號換成空白格、數字直接丟掉
    b. 將文字全部轉成小寫
    c. 利用空白格分割字串，變成一個個token
    d. stemming using Porter’s algorithm，將換行符號或是長度為1的字丟掉
    e. 使用 nltk 提供的stopwords list 去除 stopwords

2. 使用 tf_dict , df_dict 來記錄每個文件的terms的 tf 與 df
3. 製作 tf_idf 的 dictionary 存放所有文章與其token的 tf-idf

第二部分 - HAC

- `C`: 初始設定就是將文件倆倆配對後計算其cosine similarity (與作業二相同的算法)
- `get_max_cos`：每次都會去取最遠的距離
- 將兩文件合併之後，要去計算與其他文件之關係更新`C`
- 不停合併，直到剩下 20 組 13組 8組