PA2：作業格式與要求

游佳民 (YU, CHIA-MIN)

所有班別

同學們好：

以下將詳細說明作業二注意事項：

Dataset download
密碼是「IRTM2024」，共八碼，英文是大寫。

作業內容
參考slide上的作業說明及要求。
可延續PA1。PA1 是讓大家做 tokenize，PA2 是要讓大家練習計算 token 的 tf、df，所以 tokenize 部分大家可以直接拿 PA1 來用，當然，如果覺得之前寫的不好想再修改也是都可以的。
評分標準：沒有明顯 stopword 殘留，符合要求的輸出格式，基本上就是滿分，作業二也是沒有標準答案的，因為大家切出的 token 不同，後續計算出的 tf、df 當然也就不同了。
套件限制：計算 tf、df 這部分請一定要自己寫，不要用現成的套件，e.g., sklearn，但是計算 cosine 用一些向量計算的套件是 ok 的，e.g., numpy、scipy。

檔案繳交格式
Report內容須包含：
標註執行環境 (e.g., Jupyter Notebook, Pycharm, etc.)
程式語言 (請標明版本)
限定使用Python 3以上
執行方式 (重要!!!)
如何編譯或是執行你的程式，並且確認按照所提供的步驟，能夠正確的 output 出結果 (以截圖搭配文字說明)
如果有使用非原生套件請說明需要pip install什麼套件
作業處理邏輯說明
Report繳交格式為PDF檔
輸出檔案位置：為了批改方便，請同學將程式中輸出檔案的位址設為 "./output/"資料夾！
程式應要能輸出所有document的vector file，非只有document 1。但繳交的zip檔中只需放入document 1的vector file。
繳交時請不要包含 data 檔案，而是使用相對路徑 "./data/*.txt" 讀取檔案(與程式檔同一層)。
最後上傳的壓縮檔（學號.zip）須包含程式碼(pa2.*)、Report(report.pdf)、兩個輸出結果(dictionary.txt、1.txt)、以及兩個空資料夾(output/、data/)，架構如下：

R11725042
├── report.pdf
├── 1.txt
├── dictionary.txt
├── pa2.*
├── output/
├── data/

請確認檔名與上述範例相同！

常見問題

檔案：如果有程式執行所需的檔案請記得一併附上（e.g., stopwords.txt）。
程式語言：限定使用Python 3以上。
繳交方式：請繳交到 NTU COOL 的作業區中。
執行方式與作業邏輯說明的差異：執行方式是要怎麼跑這個程式，像是要在 cmd 內輸入怎樣的指令或是需要用 jupyter notebook 跑 ipynb 檔，邏輯說明則是你的程式內部是怎麼處理資料的，像是先移除 stopwords 再 stemming 之類的。
如果有其他問題歡迎在討論區詢問 ~

TA
此公告已關閉評論