Homework3: PLSA

B10632026 吳苡瑄

- 程式執行順序
  - 處理讀檔，把所有資料先進來存成我好用的格式 (dict 之類的)
  - 根據 kaggle discussion 中的建議，依照 word_list - doc_list，把所有 tf 非零的 i, j, tf 存起來，這樣可以有效的節省空間 (億 → 三百多萬)
  - 做 EM 演算法去不停更新和收斂 P(tk | wi, dj)、P(wi | tk)、P(tk | dj)
  - 最後照公式去算結果，我還有先把一些值先算好存起來，ex. P(wi | dj)、P(wi | BG)
  - 輸出結果

- 遇到的困難
  - 對 EM 演算法的理解是最大的問題，寫壞了好幾次才終於理解
  - 電腦的效能不足以應付暴力解決所有問題，必須思考優化方法，也沒辦法說先跑一遍再慢慢優化，因為一遍也跑不完，所以只能直接去思考優化方法，但這樣比較抽象也比較難 debug

- 心得

  好難，一開始很不理解 EM 演算法，寫程式、出問題、回頭看講義、問學長，這樣的流程來回好多遍了，最後才真正理解 EM 是怎麼回事，了解以後因為跑得太慢，需要用 jit 來幫忙，又會遇到一些要傳參、jit 不收的運算的問題，最後可能因為我還是寫得不夠好，跑程式也要跑超久，大概一個小時才會跑完一次，所以調參數變得很麻煩又很珍貴，從開始寫作業到跑出一個正常能交的結果就花了我三天，也沒什麼心思調參數了，記錄我調參數的測試過程，然後謝謝 NLP 實驗室的 Hank 學長，我的 EM 一直搞不懂問了他很多次。。

```
a=0.4 b=0.4 iter=120 topic=32 0.56
a=0.4 b=0.4 iter=120 topic=64 0.55
a=0.4 b=0.6 iter=50 topic=10 0.54
a=0.7 b=0.2 iter=150 topic=64 0.55
a=0.7 b=0.2 iter=120 topic=32 0.57068
a=0.7 b=0.2 iter=150 topic=32 0.57150
a=0.7 b=0.2 iter=200 topic=32 0.57128
a=0.7 b=0.3 iter=150 topic=32 0.55473
a=0.7 b=0.3 iter=200 topic=32 0.55481
a=0.7 b=0.2 iter=150 topic=48 0.55666
a=0.6 b=0.2 iter=150 topic=32 0.5714
a=0.6 b=0.3 iter=150 topic=32 0.5712
```