1. 執行環境 Mac terminal

2. 程式語言 Python 3.8.5

3. 執行方式

安裝套件：pandas，numpy
執行指令：
python3 pa2.py 1.txt 2.txt
(原始檔案名)

```
lalami@wangpeilindeMacBook-Pro hw2 % python3 pa2.py 1.txt 2.txt
originalFileName:1.txt and 2.txt
creating Dictionary...
Finish Create Dictionary.
cosine Similarity:  0.20076005955064838
lalami@wangpeilindeMacBook-Pro hw2 %
```

結束後會在同個資料夾產生dictionary.txt/ doc1.txt/ doc2.txt 的文件

並印出 cosine similarity

(document 1 與 2 的 cosine similarity: 0.20076005955064838)

4. 作業處理邏輯說明：
以預計算 cosine similarity 的 txt 檔名為參數，包含幾個 function：
(1) toTerms：同 hw1，將文章切詞
(2) createDictionary：對所有文章進行切詞，在每次文章切完詞後存成 dataFrame 形式，透過.groupby('term').size() 方式計算出各文章 term 的 tf，將第一篇的結果放入 docTerms 後，每次 merge 一篇新處理好的文章 tempTerms，計算出 df，並存檔成 dictionart.txt
(3) toUnitVector：依照公式計算 tf-idf unit vector 並存成 docID.txt
(4) readUnitVectorTxt：如果 docID.txt（UnitVector）檔案還不存在，先執行 toUnitVector，讀檔後回傳 dataFrame 形式
(5) cosine：傳入預計算的文章檔案名，讀取文章的 tf-idf unit vector，使用 inner merge 合併，將有數值的維度數值相乘後相加，得出 cosine similarity