Data Structure Programming HW1 Readme

學號：B06901045 系級：電機四 姓名：曹林熹

使用語言：Python

套件：os, numpy, pandas, sys

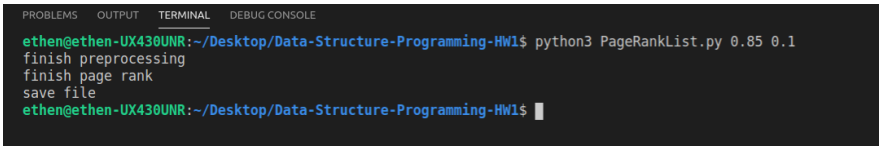
使用方式：

1. PageRankList

   在 terminal 輸入

   python3 PageRankList.py <d_value> <DIFF_value>

   執行完程式後，可以在當前資料夾得到 pr_xx_yyy.txt

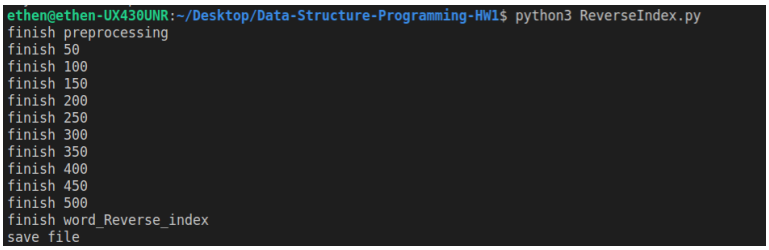


2. ReverseIndex

   在 terminal 輸入

   python3 ReverseIndex.py

   執行完程式後，可以在當前資料夾得到 reverseindex.txt

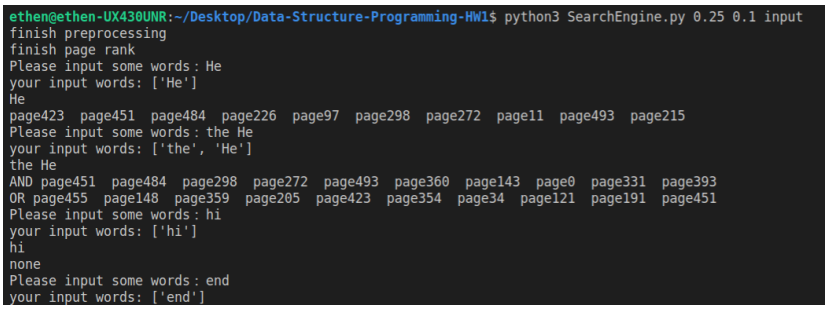


3. SearchEngine

   在 terminal 輸入

   python3 SearchEngine.py <d_value> <DIFF_value> input

   執行完程式後，可以與程式互動。輸入一段文字得到由大至小的前十個 pages，直到使用者輸入 ‘end’ 則停止程式。



在 terminal 輸入

python3 SearchEngine.py <d_value> <DIFF_value> <non_input>

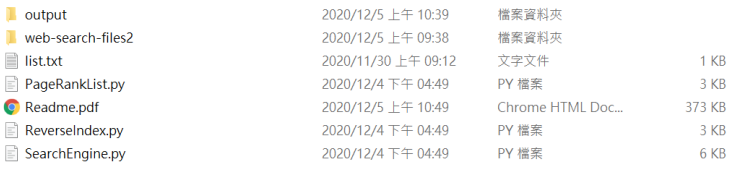
<non_input> 代表 'input' 以外的任何指令，執行完程式後，可以在當前資料夾得到 result_xx_yyy.txt。

```
ethen@ethen-UX430UNR:~/Desktop/Data-Structure-Programming-HW1$ python3 SearchEngine.py 0.25 0.1 x
finish preprocessing
finish page rank
save file
```

資料結構與演算法操作：

1. 首先我的 preprocess 先將資料按照名字都存進一個 numpy array 裡面，因為資料有 500 筆且每筆資料含有不相等的 outbranch page 數量，因此在這邊我跑了兩個 for loop，time complexity 與 space complexity 皆為 $\theta(n^2)$。
2. 接著跑 page rank 演算法，使用 while loop 讓 diff 小於 DIFF 即可以停止程式。起初的 pr vector 是每一項皆為 $\frac{1}{501}$ 的 [501, 1] np array，因為我們有一個空的 page500 也會影響 page rank 演算法。計算完成後將 pr 用 DataFrame 的方式存下方便讀取，最後寫入 txt 檔，整個過程加上預處理花費 $\theta(n^2)$。
3. 跑 reverseindex.txt 花費時間較久，因為我必須將每個 page 的 words 做比較，每個 page 有 20 個 words，總共互相比較會花費 $\theta(20n * 20n) = \theta(n^2)$ 的時間，space 一樣存在一個 DataFrame 內再寫入 txt，也是花費 $\theta(n^2)$ 空間。
4. 跑 SearchEngine 有分是不是互動式的 mode，不過演算法差不多。這邊每當輸入一個新字時，我們會對這個字去每一個 page 找有沒有存在裡面，而我需要順便拿先前建立的 pr 這個 DataFrame 拿出對應 page 的 page rank 做排序，這樣才可以顯示正確的順序。整個演算法花了 $\theta(n^2)$ 的時間，space 花費 $\theta(n)$ 空間，因為我們的 inputs 是 $\theta(n)$ 空間而已。

*上傳到 CEBIA 的 zip 包含以下檔案，其中 output 有 pr_xx_yyy.txt、reverseindex.txt、result_xx_yyy.txt



output	2020/12/5 上午 10:39	檔案資料夾	
web-search-files2	2020/12/5 上午 09:38	檔案資料夾	
list.txt	2020/11/30 上午 09:12	文字文件	1 KB
PageRankList.py	2020/12/4 下午 04:49	PY 檔案	3 KB
Readme.pdf	2020/12/5 上午 10:49	Chrome HTML Doc...	373 KB
ReverseIndex.py	2020/12/4 下午 04:49	PY 檔案	3 KB
SearchEngine.py	2020/12/4 下午 04:49	PY 檔案	6 KB