**1. 執行環境**

Jupyter Notebook

**2. 程式語言 (請標明版本)**

Python 3.7

**3. 執行方式 (重要!!!!!!!)**

以Jupyter Notebook執行

**4. 作業處理邏輯說明**

沿用作業二，有改t\_index的部分讓dictionary.txt跟Docid.txt可以對應；

C[n][i]用來記錄文件id=n跟文件id=i的相似性

I用來記錄目前還可以merge的文件

用single-link

argmax函數輸入C跟I，回傳

1. i: merge進去的文件id
2. m: 被merge掉的文件id
3. 更新後的C跟I

函數中MAX[0, 0, 0]用來記錄目前找到最大的文件相似性, i跟m

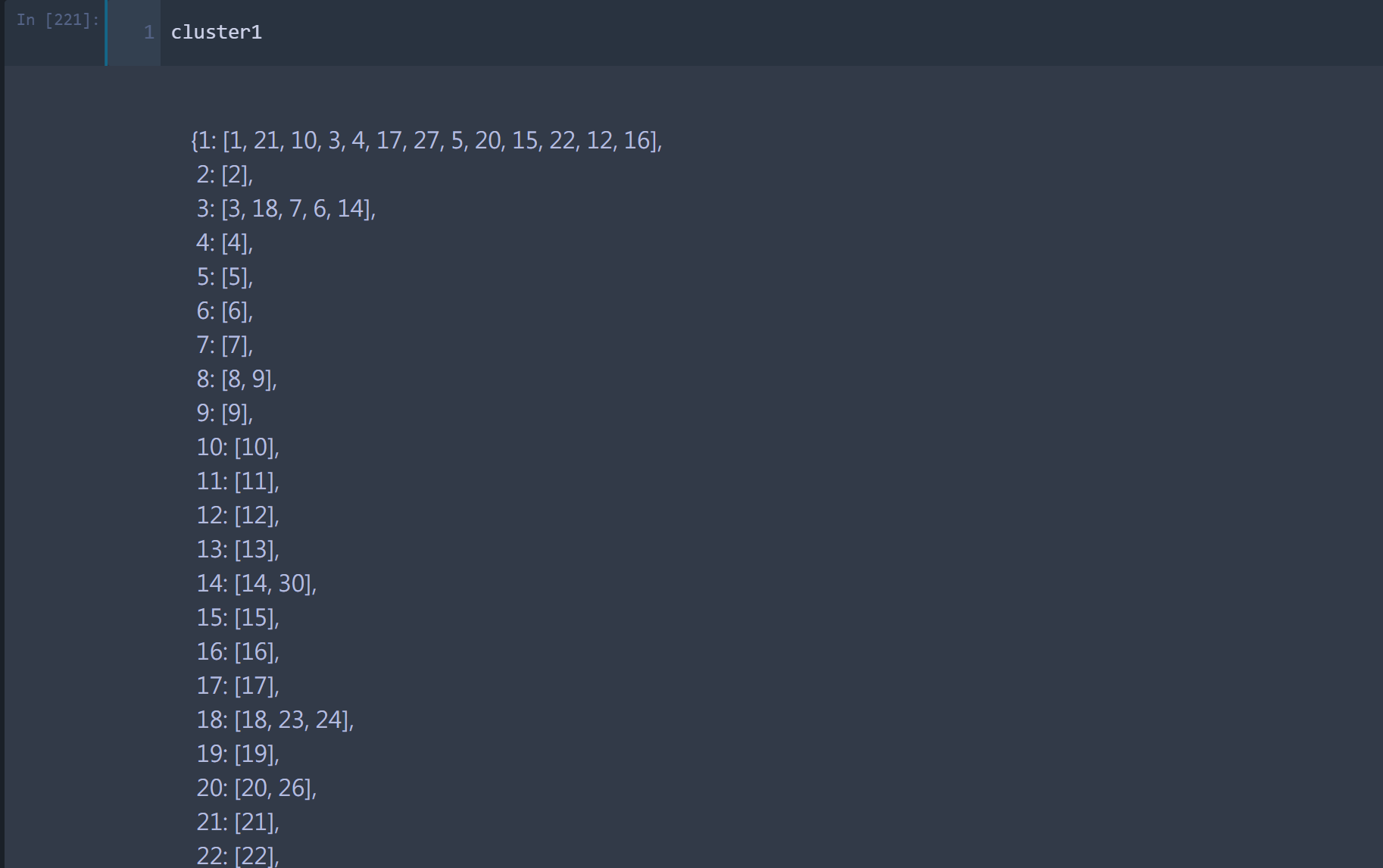
知道了i跟m後，把原先**其他文件**跟i和m的文件相似性更新為較大的那個(可能是i也可能是m)

把m這個文件紀錄為已經merge，I[m] = 0

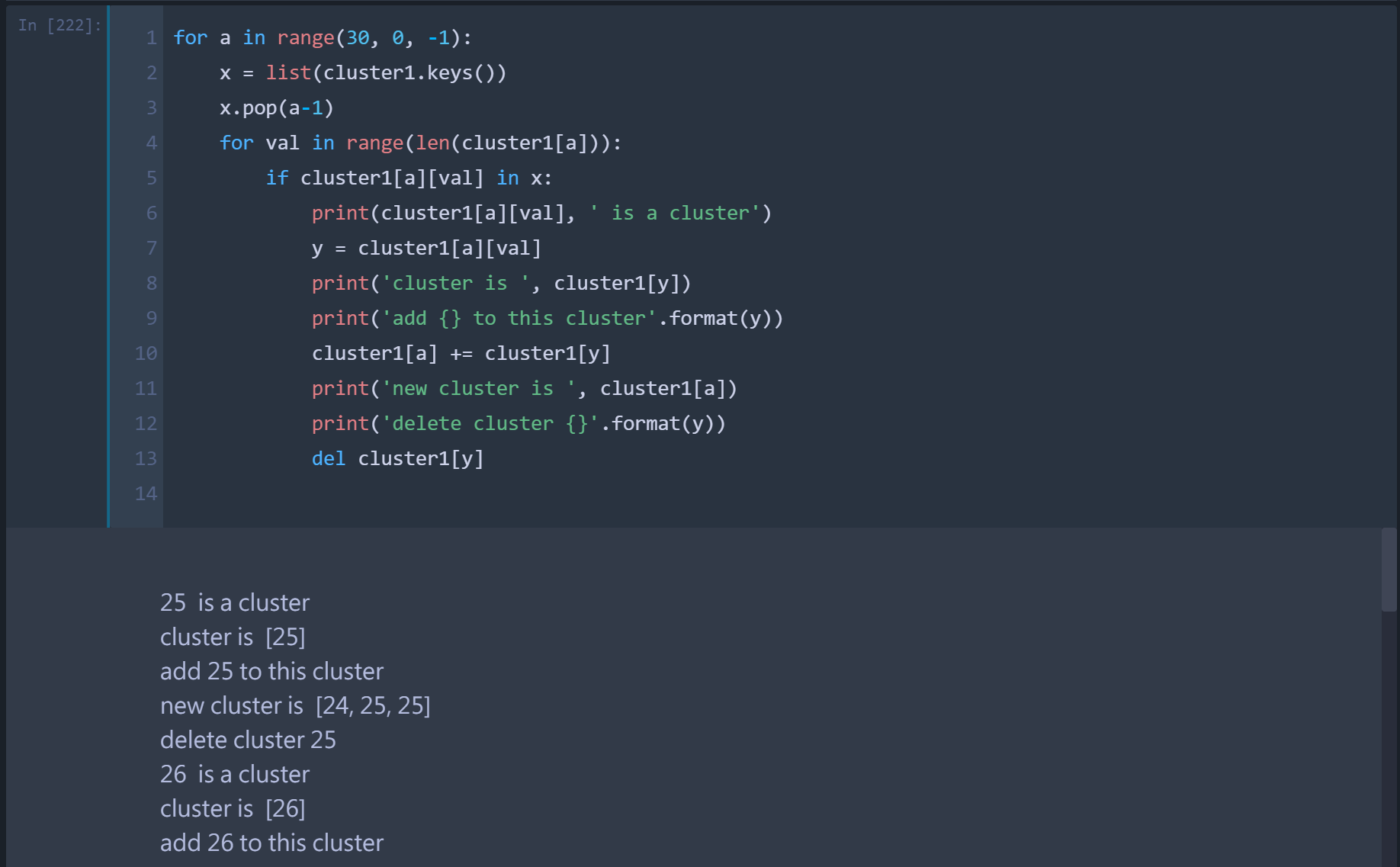
接下來cluster1先把所有文件當作是一個cluster，key跟value都是文件id，其實也避免沒有被merge的文件自己一個cluster沒被算到。

重複執行argmax函數，把m加到i裡

I總和為目前還可merge的cluster數，I==8, 13, 20就停



cluster1類似上圖，但每個cluster中的list中的id，也有可能是已經存在的cluster



所以要用這一步把小cluster加進大cluster中，並移除掉小cluster

但這樣裡面會有重複，所以新用cluster紀錄並移除重複list(set(…))

寫文件:

每換到一個新的cluster就要多換一次行

最後沒來得及跑完，不清楚如果跑了1個多小時還是沒有跑完是否正常，試過只看文件1到100結果是成功的

