HITS

Implementation detail



HIT、PageRank、SimRank 三個的開頭大致上都相同

1.讀取graph\_X.txt檔案

2.先存進list (a)裡面

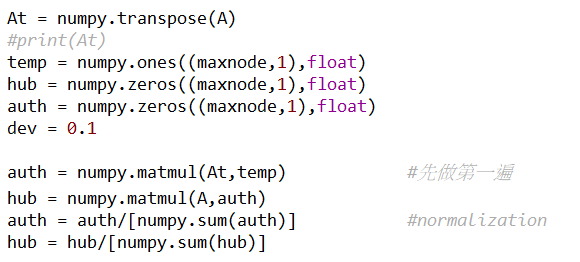
3.處理每一個line中的逗號跟換行(\n)

4.紀錄總共用多少node (maxnode)

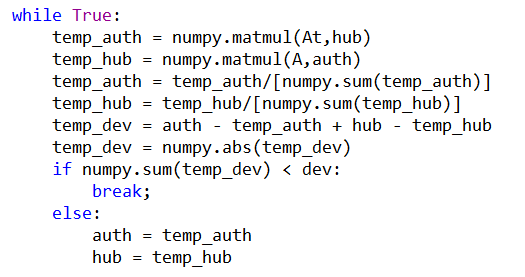
5.用maxnode宣告numpy array (A)

6.根據a中data把A中index對應令為1

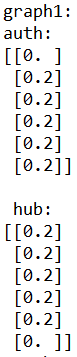
7.若有需要再做除sum 讓每一row總和變為1

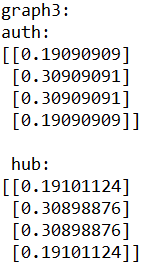


8.hub,authority的初值設定，先做第一遍還有正規化



9.while loop直到差值小於dev

Result analysis and discussion

用Adjacency matrix來表示

連接關係為A 轉置為At

有1->2 則A[1][2] = 1

同時At[2][1]=1 代表node2 有被node1指向

若要表示Authority的傳遞

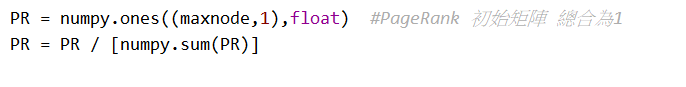
At\*[1,1,..1,]t在正規化可以獲得Authority的比例

我沒有做project3 dataset以外的兩張graphs

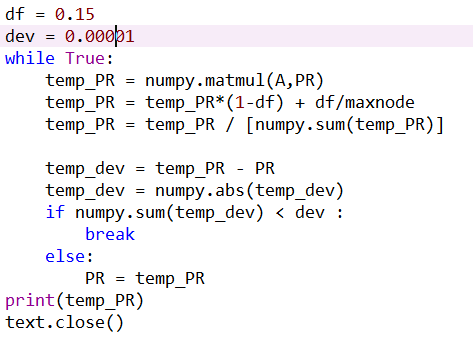
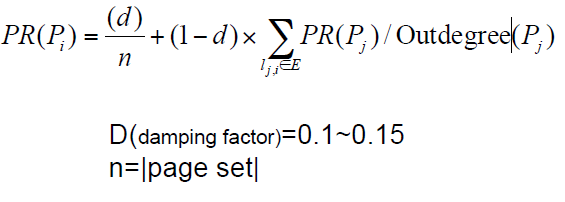
PageRank

Implementation detail

前1~7跟HIT差不多



8.令Pagerank一開始的矩陣



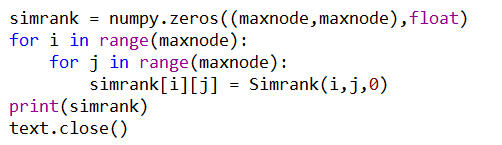
9.temp\_PR = temp\_PR\*(1-df) + df / maxnode

df為damping factor

maxnode為node的總數

SimRank

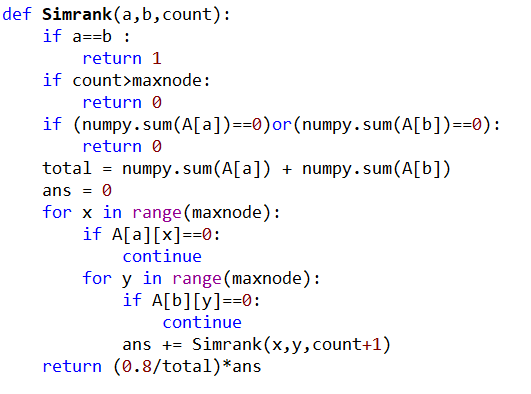
Implementation detail



8.另一個page size大小的矩陣來表示答案S[i][j]表示i跟j的相似度

(i.e.會是個對稱的矩陣)

9.



除了當(1) a==b return 1

(2) 當a或b沒有parent node,retrun 0

我額外增加了一個遞迴超過page size時也會return 0

因為我有遇到迴圈的情況卻不知道怎麼解決，所以用這個方式來中斷

C我令0.8

graph1~3都有結果跑出來

但是graph4卡住

Discussion