Data Mining 2022 Project 3 Grading Policy v3

Project 3: Link Analysis
實作三種演算法：PageRank, HITS, SimRank
可以開放使用packages，唯演算法部分需要自行實作，
例如**可以使用networkx套件，但不可以呼叫套件中的 pagerank 函數或是其他兩個演算法的函數**。
如果對於所使用的套件函數有疑慮，可以先來信詢問。

TA Email: nckudm@gmail.com (IKM Lab)

- Dataset:

請下載 moodle 上這兩個檔案，將會看到Graph 1-6與IBM檔案，即這次作業所使用的資料（皆為directed）。

```
[2022] Project3 Dataset檔案
[2022] 2021 project1 使用的 IBM dataset檔案
```

12/15 Update
ibm 檔案解讀方式：
ibm 檔案格式如下，請將 (1, 307) 看成一個 directed edge, (1, 443) 為另一個 directed edge…。以此類推。

```
1          1        307
1          1        443
```

輸出時請先按照 node 編號順序 sort 輸出，假設sort完後 node 編號由小到大為 [1,3,23,40,307, ...]，則輸出的 array 其對應的 index 0 為node 1的數值、index 1 為 node 3 的數值、index 2 為node 23 的數值…。matrix[1,2] 為 node 3 對 node 23的數值…。以此類推。

- PageRank, HITS, SimRank 評分標準：
  - Enforceability [10pts]
    12/15 Update
    - damping_factor = 0.1
    - decay_factor = 0.7
    - iteration = 30
    - 在使用以上參數時，Graph 1~6 和 IBM 檔都須可以執行。
      12/18 Update
      由於runtime問題，SimRank 不用執行 Graph 6與IBM檔。
    - PageRank 演算法的公式請以老師投影片p.37實作，以上damping factor即為該頁公式中你應代入的d。
  - 初始值：
    - HITS 中 hub 與 authority 初始為1。
    - PageRank 中 pagerank 初始為1/N。
  - Correctness of answer [10pts]
    請將檔案寫成如下的檔案結構並且一併放入繳交資料夾內。

```
/results
├── graph_1
│   ├── graph_1_HITS_authority.txt
│   ├── graph_1_HITS_hub.txt
│   ├── graph_1_PageRank.txt
│   └── graph_1_SimRank.txt
├── graph_2
│   ├── graph_2_HITS_authority.txt
│   ├── graph_2_HITS_hub.txt
│   ├── graph_2_PageRank.txt
│   └── graph_2_SimRank.txt
├── graph_3
│   ├── graph_3_HITS_authority.txt
│   ├── graph_3_HITS_hub.txt
│   ├── graph_3_PageRank.txt
│   └── graph_3_SimRank.txt
├── graph_4
│   ├── graph_4_HITS_authority.txt
│   ├── graph_4_HITS_hub.txt
│   ├── graph_4_PageRank.txt
│   └── graph_4_SimRank.txt
├── graph_5
│   ├── graph_5_HITS_authority.txt
│   ├── graph_5_HITS_hub.txt
│   ├── graph_5_PageRank.txt
│   └── graph_5_SimRank.txt
└── graph_6
    ├── graph_6_HITS_authority.txt
    ├── graph_6_HITS_hub.txt
    ├── graph_6_PageRank.txt
    └── graph_6_SimRank.txt  # optional
└── ibm-5000
    ├── ibm-5000_HITS_authority.txt
    ├── ibm-5000_HITS_hub.txt
    ├── ibm-5000_PageRank.txt
    └── ibm-5000_SimRank.txt # optional
```

  - 三個演算法共 20 * 3 = 60pts
  - 寫檔案要求請見2021年的spec(已放上moodle)
  - 輸出的array shape 為 (1, N)，輸出的 matrix shape 為 (N,N)
    (N: 該圖 Node 數量) 。
  -

12/18 Update
輸出precision：小數點後第三位（ `np.savetxt(...,fmt ='%f3',...)` 。
https://moodle.ncku.edu.tw/pluginfile.php/1111104/mod_resource/content/0/Data Mining Project3 作業注意事項.pdf

- Report 評分標準：
  - Find a way [10pts] **12/15 Update**
    Find a way (e.g., add/delete some links) to increase hub, authority, and PageRank of Node 1 in first 3 graphs respectively
    需要附上圖 (至少說明Graph 1~3)
    需要貼上調整後 hub, authority, pagerank 的數值。
  - Algorithm description [10pts]
    解釋每個演算法步驟流程，只貼 code 不說明則不給分。
  - Result analysis and discussion [10pts]
    說明每個圖 (至少說明Graph 1~3) 的結果並討論。
    針對不同 damping factor 和 decay factor 進行討論。
  - Effectiveness analysis [10pts]
    須列出三個演算法在每張圖上的執行時間，並討論其原因。

- Submission
  - **Deadline: 12/27(Tues) 9:00**
  - Language：python3
    可以直接寫成notebook（.ipynb）的形式，或著寫成python class再import至notebook中執行，並**保留正確執行留下的cell outputs**，該notebook請取名 `output.ipynb` 。
  - 請包成 `zip檔` ，檔案名稱 {學號}_DMProject3.zip（eg. P76111111_DMProject3.zip），解壓縮後要呈現以下的結構。
  - src檔案夾內放你的原始程式碼與 `output.ipynb` 。我們會看code但不會執行，除非你的程式碼長得和別人非常相似（抄襲疑慮） or 你的程式碼看起來不可能跑出正確結果，對於前者將會給0分，對於後者將會請你至實驗室使用繳交的程式碼demo。

```
P76111111_DMProject3
├── results # 上面提到的
│   ├── graph_1
│   ├── graph_2
│   ├── graph_3
│   ├── graph_4
│   ├── graph_5
│   ├── graph_6
│   └── ibm-5000
└── src   # your source code directory, need to include the notebook with outputs
└── report.pdf
```