

【Practice of Social Media Analytics】

HW1 – Link Prediction

<Kaggle> <https://www.kaggle.com/t/ac970ee3647044789ce429e7447f8b23>

說明：

這次作業是要預測node pair (node1, node2)有沒有隱藏關聯(hidden edge)。會提供大家約20000條的edge，讓大家建立出social network以及訓練資料集。資料集是一個「有向網路」，因此每組node pair代表一條具有方向性的邊，例如(361, 981)代表的是一條由節點361指向節點981的邊。

此 social network 中有大約 5000 條 hidden edge，這是你們要預測的關係。可以使用任何的方法預測，直接使用呼叫的 function 也行，**作業要上傳到 Kaggle 評分**，**寫完後要繳交程式檔和說明文件到 Moodle**。說明文件中請簡述演算法流程，並說明如何執行你的程式。

作業成績評分方式：60%是 performance(Kaggle 分數)，40%是 report

`data_train_edge.csv` 是拿來 training 的資料

node1	node2
287	68
63	552
189	20
380	376
370	443
87	427
212	411
407	525
28	128

`predict.csv` 是要預測的node pair (node1, node2)

node1	node2
306	308
917	918
49	989
222	256
270	384
397	397
899	185
954	327
217	162

`ans_example.csv` 是上傳到Kaggle的格式

predict_nodepair_id	ans
0	1
1	1
2	1
3	0
4	0
5	0
6	0
7	1
8	0

`predict_nodepair_id`是單純的index，從0開始依序排到10199，`ans`就是預測結果0或1。(0表示你預測此node pair沒有hidden edge，1表示你預測此node pair有hidden edge)

Description :

This assignment is to predict whether a node pair (node1, node2) has hidden relation (i.e., hidden edge). There are about 20,000 edges provided for you to reconstruct the social network and the training dataset. (This is a directed network, so each node pair represents a directed edge. E.g., (361, 981) represents an edge from node 361 to node 981.)

The social network has about 5,000 hidden edges. These are the relationships you are asked to predict. You can use any prediction method, and you can use any functions/libraries/packages directly. **The result should be uploaded to Kaggle platform for evaluation. You also need to upload the program files and the report document to Moodle.** In your report, please briefly describe the algorithm you use, and provide instructions about how to execute your program.

Homework scoring method: 60% is performance (Kaggle score), 40% is report

data_train_edge.csv is the information used for training

node1	node2
287	68
63	552
189	20
380	376
370	443
87	427
212	411
407	525
28	128

predict.csv is the node pair (node1, node2) to be predicted

node1	node2
306	308
917	918
49	989
222	256
270	384
397	397
899	185
954	327
217	162

ans_example.csv is the format uploaded to Kaggle

predict_nodepair_id	ans
0	1
1	1
2	1
3	0
4	0
5	0
6	0
7	1
8	0

predict_nodepair_id is index, starting from 0 to 10199, *ans* is the prediction result 0 or 1. (0 means there is no hidden edge for this node pair; otherwise the ans is 1.)