Assignment #1 - Link Prediction

a. 說明如何執行程式(並附上程式碼檔案)

將程式碼與 'data_train_edge.csv', 'predict.csv', 'ans500_ground_truth.csv' 置於同資料夾

python social_netwok_hw1_3_gt.py

python data_revised.py

b.簡介你所使用的程式架構及演算法流程(如果有進行前處理也請解釋原因)

一、資料前處理:

1) 把 ans500_ground_truth.csv 加入 training data

2) 取 negative sample(不能取到 predict.csv 中第 500 row 以後的 data)

二、定義 feature function

1) Similarity measure
    1) Jaccard Distance(for followees and follower)
    2) Cosine distance(for followees and follower)
2) Ranking Measures
3) Page Ranking(of source/dest)
4) Other Graph Features
    1) Shortest path
    2) Checking for same community
    3) Adamic/Adar Index
    4) Is persion was following back
    5) Katz Centrality(of source/dest)
    6) Hits Score(of source/dest)
    7) SVD
    8) num_followers_s
    9) num_followees_s
    10) num_followers_d
    11) num_followees_d
    12) inter_followers
    13) inter_followee
    14) Weight Features
        1) weight of incoming edges
        2) weight of outgoing edges
        3) weight of incoming edges + weight of outgoing edges
        4) weight of incoming edges * weight of outgoing edges
        5) 2*weight of incoming edges + weight of outgoing edges
        6) weight of incoming edges + 2*weight of outgoing edges

三、計算 feature

將原始 dataframe 的資料 apply 以上的 function 算出 feature,加入新的以此 feature 為名的 column

四、Train

將所有在步驟三中的 feature 使用 RandomForestClassifier/GradientBoostingClassifier 進行訓練,並使用 RandomizedSearchCV 找出最佳參數(best estimator)

五、Predict

將訓練完的 predict.csv 給入訓練完的 classifier 進行預測，並依規定儲存結果。

六、Data revised

藉由觀察資料,將在 training edge 出現的 edge 以及連到自己的 edge 設成 1,並補上'ans500_ground_truth.csv' 的答案。

c.結果分析

RandomForestClassifier 中影響因素最高之前三名為'cosine_followers','inter_followers','adar_index'

,故最後以這三者為 feature 下去訓練,結果比取全部 feature 好。GradientBoostingClassifier 中影響因素以'adar_index' 為壓倒性的高,所以最後只取'adar_index' 下去訓練，結果比取全部 feature 好。

GradientBoostingClassifier 結果又比 RandomForestClassifier 結果好。Kaggle public score 最高為 0.95457(revised)。