Hw1-2 report

陳偉 R06944043

Private Score | Public Score： 0.50470 | 0.50383

Hw1-2 和 hw1-3 使用基本相同的 parameter 和 code。少了 time 這個 feature，以及 classify 換用了 regression 。看排名變動非常大，原因大概就是，prediction 不是很 powerful 的情況下(接近 random)，之前在前面的同學去 overfit public score，結果導致 private 比較慘。經驗就是，不要 overfit public。

I. Preprocessing & negative sampling:
   - validation 資料來自 train 10%，validation 部分的 vertex 不出現在 training 中
   - negative labeled data 取 random walk step = 2 不構成邊的點
   - Graph 中 vertexes 的 degree

Final Features：

- Tf-idf score； tf-idf 算出的 cosine similarity 作為 feature，不用 tf-idf 直接作為 feature 是因為 維度太多，training 時間長，引入 noise 且易 overfitting。
- Degree; 计算出 training 中 mean degree，unseen vertex 的 degree 设置为 mean degree，degree feature 有 degree， degree 相加。没有 degree 相减是因为 unseen vertex random 出现在 source 和 target，会导致 degree 相减会变成 noise。但相加不会受到位置影响。

Classifier：

Random forest regression

- 最好的上傳結果來自 單一 random forest regression 取 regression，沒有加 word embedding 的 feature，因為訓練時間隨 feature 數 linear 增加。
- 因為 prediction 都集中在 random，0 占比又稍多，prediction 中輸出更多 0。將 regression 的 threshold 從 0.5 提高到 0.7，相當于 與 all zero 的 ensemble。