Hw1-3 report

陳偉 R06944043

Private Score | Public Score： 0.73742 | 0.73181

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
- Times 只取 year
- Graph 中 vertexes 的 degree
- word embedding 來自 fastText

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。
- Time; prediction 與時間差不一定是 linear 的關係，time feature 除了有 source 和 target 的 time、time 相加、相減，還加入相減的 one-hot。
- 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。
- Tf-idf weighted word embedding，比 mean word embedding 的 performance 有一點點提升，使用 random forest 時沒有加入這一 feature。

Classifier：

Random forest， logistic regression

- 用這兩個是因為它們有最快的訓練速度，最好的上傳結果來自 7個 random forest 的 ensemble，但沒有加 word embedding 的 feature，因為訓練時間隨 feature 數 linear 增加。
- 訓練的時候有統計 testing 的 distribution，classifier 用 random forest 的時候，testing prediction 1/0 distribution 偏離 50% 相較 logistic regression 嚴重，logistic regression 是 50% 左右，猜測是 random forest 更容易 overfitting 的原因。