Hw2-1 report

Leader: 陳偉 R06944043

Teammate: 周思佳 R07922146，陈熙 R07922151

Private Score | Public Score： 0.04657 | 0.04766

I. Preprocessing & negative sampling:
- 對 user food pair 計算出現次數，作為 concurrence矩陣的值，最終結果沒有考慮 user 和 food 的 features.
- 矩陣中 random 15% 的點作為 validation set， 用於調整參數。

II. Model：
- 採用 Warp model， 用 lightfm 實作，將得到的 concurrence 矩陣作為輸入，訓練 model，最終參數：
  - learning_rate = no_components = 29,
  - user_alpha = 0.00048625731451155697,
  - item_alpha = 0.00048625731451155697,
  - max_sampled = 37,
  - epoch = 197

III. Parameter Tuning：
- 用 skopt 的 forest_minimize 來自動調參，其原理是在所規定參數範圍內隨機產生參數，並用此參數 train 模型，用 validation set 評估參數好壞，通過大量的隨機挑選，得到當前最優參數，經驗上來說，隨機選取參數的方法比 grid search 選出來的參數 performance 更高，在我的實驗結果中表現也是如此。這種方式得到的參數比 default 參數的 performance 高 0.003，大概是第 2 名與第 12 名的差距，是很大的提高。

IV. Experiment：

尝试了 1. 使用user 和 food feature, 2. 增加 data sparsity, 3. DeepWalk。
- user 和 food feature，不管是加入全部 feature 還是部分 dense 的 feature，都使 performance 降低。
- 增加 data sparsity, 統計了下 occurrence matrix 的sparsity, 發現 sparsity = 3.6 %，顯然低於老師上課講的 10% 的 threshold，於是想到通過刪除掉 一些不常出現的 food，來增加 sparsity。結果表明 performance 沒有上升，猜想助教有對 data 做過處理，剩下的 food 都是頻率較高的，所以刪除會損失 performance。
- DeepWalk，屬於突發奇想，隨便亂試，考慮到 deep walk 學到使相連的 node 之間的 embedding similarity 越高，因此想到把 user 和 food 都當作 node，user 有吃過 food 就作為中間有 path，去訓練，結果performance 很差 (Private | Public：0.00082 | 0.00076)，推測理由是，1. 資料中 user 和 food 都是互相高頻出現的，因此任何 user 和 food 都兩兩有 path，導致產生的 embedding 沒有區分度，計畫未來將 path 設計得有差異性； 2. Matrix 有其結構性，而 graph 沒有，正因為其缺少結構性，才折衷使用 random walk，用非 optimal 的方法去訓練結構性的資料，沒有好好利用其結構的特徵，自然 performance 會相較更差。