Data Mining－Final Project

Group 16: 312554001 戴珮珊、0816077 陳柏安、109550117 李念嘉

1. **Describe how you solve this problem. Details include preprocessing, embeddings, model selection, hyperparameters should be provided.**

    a. Preprocessing:
    首先處理 behaviors.tsv，將 user_id 映射到 integer，接著平衡 impressions 中的數據使0 (no click) 及 1 (click) 配對成一對。使用 explode fuction 展開 impressions 欄位，亦即展開每個 user 的 impressions。

    在 train_news.tsv 我們載入資料並將 category、subcategory、title 以及 abstract 等映射到對應的 integer，確保每筆新聞都能被 model 更好的處理。

    b. Embeddings:
    使用 GloVe 進行 word embedding 的處理，如果當中有對應不到的詞，將進行隨機初始化。這些 embeddings 可以幫助 model 更好的理解資料中的所有訊息，以提高 model 的準確度和穩定性。

    c. Model Selection:
    在實作中我們使用 NRMS（Neural Recommender using Multi-Head Self-Attention）進行 model selection，是一種基於 multi-head attention 衍生出的推薦模型，可以更好的去捕捉到 user 想看的新聞。

    d. Hyparameters:
        i. Learning rate: 0.0001
        ii. Batch size: 128
        iii. Epoch: 2
        iv. Attention Heads: 15

2. **Choose a variable (e.g. different model, different approach) excluding hyperparameters and compare their performance. (You cannot change ONLY the hyperparameters)Explain what causes the difference of performance or why.**

    除了上述的 NRMS 以外，我們也實作了 spec 上提供的 baseline approach，將計算出的 user embedding 和用 BERT 模型計算好的 candidate news embedding 以 dot product 相乘計算相似度，並比較這兩種方法造成的結果差異。結果我們的 baseline approach 一直無法達到 simple baseline。

我們認為 baseline approach 效果不彰有幾個原因。首先，在這個方法中我們只使用了簡單的 Linear NN 還有 GRU 來訓練 embedding，而 NRMS 使用的 multihead attention 機制能更容易捕捉到上下文之間的關係。還有，在 NRMS的 preprocessing 方法中，多做了額外的處理，像是平衡點擊與不點擊的數據，可以有效減少因為數據不平衡導致的結果偏差。

3. **Do some error analysis or case study. Is there anything worth mentioning while checking the mispredicted data? Share with us.**

   Error analysis 可以更了解 model 的不足之處，從中也可以發現改進的方向。假設有一篇新聞被預測錯誤 (user 有 click，預測結果為 no click)：
   - category: health
   - subcategory: nutrition
   - title: 30 Foods That Are Never Worth The Calories
   - abstract: What a waste of stomach space!

   從文本特徵中我們可以檢查 title 中的 Food、Calories 有沒有被模型正確的 weighting，而 abstract 中的 waste of stomach space 表達上較為隨意，亦即沒有前後資訊的情況下難以確定新聞的實際主題，容易降低預測的準確度。

   在 category 以及 subcategory 的部分我們可以觀察其種類下的準確率，如果準確率過低有可能是數據質量不高造成 model 誤判，亦或是數據量不足難以進行種類上的匹配。

   Reference: GitHub - yusanshi/news-recommendation: Implementations of some methods in news recommendation.