機器學習期末報告

組別：**NTU_r05922156_LetMePass**

資訊所黃子瑋 r05922156、資訊所張馭荃 r05922052、電子所謝孟洋 r06943148

題目：Conversations In TV Shows

一、Preprocessing/Feature Engineering

我們將五個 training file 的內容依序讀入，並且將三行句子串聯成一行。接著，再利用 jieba 套件將句子斷詞，我們採用的字典為 dict.txt.big，即助教在 ML HW6 中提供的字典。我們的 feature 就是以斷詞後的句子當作 input，丟進 gensim 套件中進行 training，以獲得詞向量。最後，我們將 testing data 讀入，分別對題目以及六個候選選項利用 jieba 套件將句子斷詞，然後將斷詞後的問題以及選項利用 gensim 中的 model 將其轉換成詞向量，並且把詞向量加總以獲得句向量。我們將問題的句向量分別對六個候選選項求得 cosine 相似度(內積並除以各自的長度)，最後選擇 cosine 相似度最大的選項為我們的 output。

二、Model Description (At least two different models)

我們針對 gensim 套件中的 model 設計兩種模型。

模型一、使用停用詞 + CBOW model

```
model = word2vec.Word2Vec(X_train, min_count=0, size=100, sg=0, window=7)
```

由於在句子中會有一些沒有意義的詞語，例如：而且、還有、即使、的、了。因此我們在斷詞之後就會先將這些詞語篩掉，這樣才能從句子中抽選出較具有代表性的詞語(關鍵字)，然後再丟進 gensim 的 model 進行 training，我們將詞向量的維度設成 100，並且使用 CBOW model，將 window 參數設成 7。

模型二、不使用停用詞 + Skip-Gram model

```
model = word2vec.Word2Vec(X_train, min_count=0, size=128, sg=1, window=50)
```

另外一個想法是，句子中的停用詞雖然無意義，但是對於句子的構成也是非常重要，因此我們又重新把停用詞放回去。然後在 gensim 訓練 model 的部分，將詞向量的維度設成 128，並且改用 Skip-Gram model，以及將 window 參數設成 50。

共通部分：Handling OOV

若有句子不能適當的分詞，我們就會將該句子的詞向量設成 0 向量。

三、Experiments and Discussion

在 Kaggle 上的 Evaluation Function 為正確率，也就是說，給定 N 筆 testing data，若有 m 筆資料是正確的，正確率則為 m/N。

模型一、使用停用詞 + CBOW model

Kaggle: 0.44940

模型二、不使用停用詞 + Skip-Gram model
Kaggle: 0.49565

由 Kaggle 上的結果我們可以知道，模型二的正確率比模型一還來得高，我們推測可能是因為停用詞雖說無意義，但是在該句子中卻扮演著重要的角色，因此並不需要在模型中設定停用詞。另外，由於 CBOW model 是從前後詞的詞向量來預測出可能的目標詞，然而 Skip-Gram model 則是輸入目標詞詞向量，預測出可能的前後詞。CBOW model 比較像是本次作業會需要用到的方法，因為本次作業就是要用前文來推測後文，但是 Skip-Gram model 的效果卻會比較好。我們認為可能是因為訓練資料中的詞彙較少，反而應該是輸入目標詞，輸出前後詞比較恰當，因此選擇 Skip-Gram model 較佳。