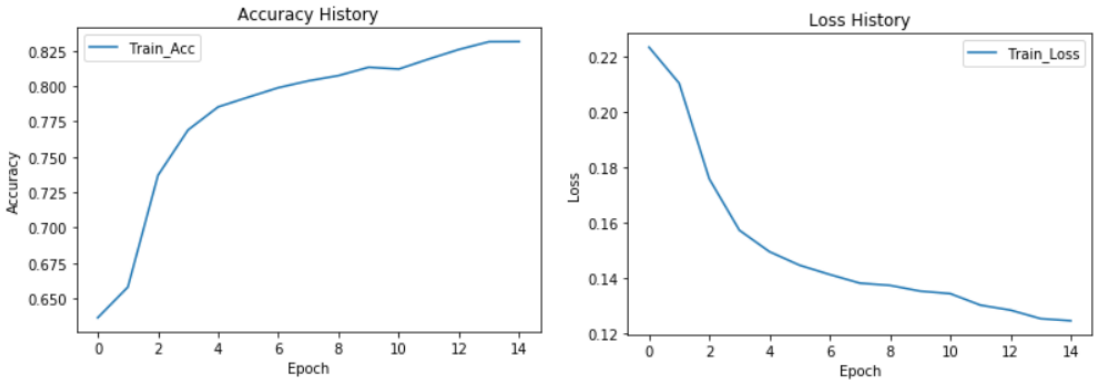
Machine Learning HW5 Report

學號：R07725019　系級：資管碩二　姓名：鄒雅雯

1. (1%) 請說明你實作之 RNN 模型架構及使用的 word embedding 方法，回報模型的正確率並繪出訓練曲線*

   Word Embedding：使用 training data 和 testing data 的 comment 訓練 word2vec 模型，每句 comment 皆會將標點符號去除，並且將 comment 中的@user 也去除，接著使用 spacy(en_core_web_lg)斷字，並進行訓練。
   word2vec 參數：embedding size=500, window=5, min_count=20

   RNN：使用 word2vec 訓練出的 pretrained embedding，LSTM 架 3 層，每層有 800 個神經元，dropout=0.2，LSTM 為雙向的，最後加一層隱藏層，input 為 LSTM 每個時點 output 的平均，輸出為一個神經元，若分數>=0.5 為 1，<0.5 為 0

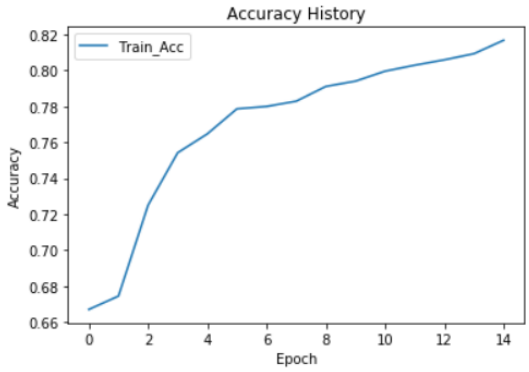


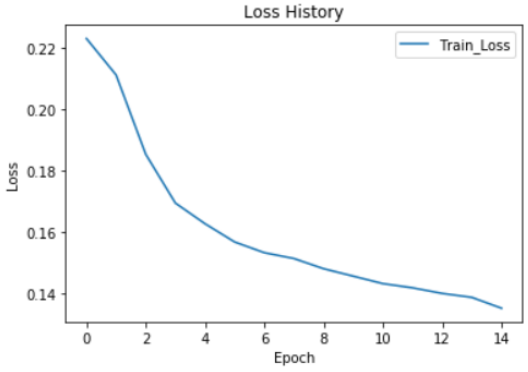
   Kaggle public score：0.79534 / private score：0.82325

2. (1%) 請實作 BOW+DNN 模型，敘述你的模型架構，回報模型的正確率並繪出訓練曲線*。

   Word Embedding 的方式同 1

   DNN：輸入直接將所有 token 的向量平均，接 3 層隱藏層，每層神經元皆為 800，最後輸出層神經元為 1，若分數>=0.5 為 1，<0.5 為 0，每層皆會接活化函數 ReLU()





   Kaggle public score：0.76511 / private score：0.79302

3. (1%) 請敘述你如何 improve performance（preprocess, embedding, 架構等），並解釋為何這些做法可以使模型進步。

   Preprocess 的部分將標點符號以及@user 去除，能將大量的雜訊濾掉，讓 word2vec 學得更好。

   Embedding 的部分將 min_count 參數設置較高(20)，代表只有出現至少 20 次的字會被訓練到，如此一來這些字因為出現的頻率較高，因而能有較好的學習，而出現頻率較低的字就用特殊 token 取代，才不會因為沒有訓練好而影響模型表現。

   LSTM 的部分，因為模型架構問題，所以一個 batch 裡的 comment 一定要一樣長，如此導致很多較短的 comment 需要 padding 特殊 token，但事實上加入了很多原本沒有的資訊，因此使用 torch 中的 rnn_utils.pack_padded_sequence，將每個 batch 拆成 minibatch，使 LSTM 能夠只吃進原有的 comment 長度。

4. (1%) 請比較不做斷詞 (e.g.,用空白分開) 與有做斷詞，兩種方法實作出來的效果差異，並解釋為何有此差別。

   模型皆用 RNN
   空白分開：Kaggle public score：0.76046 / private score：0.77906
   Spacy 斷詞(en_core_web_lg)：Kaggle public score：0.79534 / private score：0.82325
   →不論是 public 還是 private score，用 Spacy 斷詞的結果都較佳，word2vec 訓練出的結果，若用空白斷詞，訓練出的 token 有 1539 個，若用 Spacy 斷詞，訓練出有 1571 個 token，比空白斷詞多，因 Spacy 斷詞是用大量文章訓練出來的，因此斷得較精確，訓練的結果也會較佳。

5. (1%) 請比較 RNN 與 BOW 兩種不同 model 對於 "Today is hot, but I am happy."與"I am happy, but today is hot." 這兩句話的分數（model output），並討論造成差異的原因。

   RNN
   "Today is hot, but I am happy."：-0.0179
   "I am happy, but today is hot."：0.1110

   BOW+RNN
   "Today is hot, but I am happy."：0.
   "I am happy, but today is hot."：0.

   →不論是 RNN 還是 BOW 兩種模型下，兩句話的分數皆接近 0，皆為非惡意留言，但在 RNN 中兩句話的分數有些微差距，因為 RNN 是一個字一個字丟進模型中，因此句子即使所有字相同，但排列順序不同，RNN 產出的結果也就不同，因其將每個字的上下文也學進去了，而反觀 BOW 是將句子中所有字的向量平均最為輸入，因此即使文字順序不同，但所有字是相同的，所以產出的分數也就相同。

*LSTM

t=1 $x=[0\ 1\ 0\ 3]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}0\\1\\0\\3\end{bmatrix}+0=3$, $z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix}+110=10$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix}-10=90$, $z_o=[0\ 0\ 100\ 0]\begin{bmatrix}0\\1\\0\\3\end{bmatrix}-10=-10$

$f(z_i)=1$, $f(z_f)\fallingdotseq 1$, $f(z_o)\fallingdotseq 0$

$\rightarrow c'=1\cdot 3+0\cdot 1=3$, $y=0\cdot 3=0$

t=2 $x=[1\ 0\ 1\ -2]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}1\\0\\1\\-2\end{bmatrix}+0=-2$ $z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}1\\0\\1\\-2\end{bmatrix}+110=10$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}1\\0\\1\\-2\end{bmatrix}-10=90$ $z_o=[0\ 0\ 100\ 0]\begin{bmatrix}1\\0\\1\\-2\end{bmatrix}-10=90$

$f(z_i)=1$, $f(z_f)\fallingdotseq 1$, $f(z_o)=1$

$\rightarrow c'=1\cdot(-2)+3\cdot 1=1$, $y=1\cdot 1=1$

t=3 $x=[1\ 1\ 1\ 4]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}1\\1\\1\\4\end{bmatrix}+0=4$ $z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}1\\1\\1\\4\end{bmatrix}+110=-90$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}1\\1\\1\\4\end{bmatrix}-10=190$ $z_o=[0\ 0\ 100\ 0]\begin{bmatrix}1\\1\\1\\4\end{bmatrix}-10=90$

$f(z_i)=1$, $f(z_f)=0$, $f(z_o)=1$

$\rightarrow c'=1\cdot 4+1\cdot 0=4$, $y=1\cdot 4=4$

t=4 $x=[0\ 1\ 1\ 0]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}0\\1\\1\\0\end{bmatrix}+0=0$ $z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}0\\1\\1\\0\end{bmatrix}+110=10$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}0\\1\\1\\0\end{bmatrix}-10=90$ $z_o=[0\ 0\ 100\ 0]\begin{bmatrix}0\\1\\1\\0\end{bmatrix}-10=90$

$f(z_i)=1$, $f(z_f)=1$, $f(z_o)=1$

$\rightarrow c'=1\cdot 0+4\cdot 1=4$, $y=1\cdot 4=4$

t=5 $x=[0\ 1\ 0\ 2]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}0\\1\\0\\2\end{bmatrix}=2$ $\quad z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}0\\1\\0\\2\end{bmatrix}+110=10$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}0\\1\\0\\2\end{bmatrix}-10=90$ $\quad z_0=[0\ 0\ 100\ 0]\begin{bmatrix}0\\1\\0\\2\end{bmatrix}-10=-10$

$f(z_i)=1,\ f(z_f)=1,\ f(z_0)=0$

$\rightarrow c'=1\cdot 2+4\cdot 1=6,\ y=0\cdot 6=0$

t=6 $x=[0\ 0\ 1\ -4]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}0\\0\\1\\-4\end{bmatrix}=-4$ $\quad z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}0\\0\\1\\-4\end{bmatrix}+110=110$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}0\\0\\1\\-4\end{bmatrix}-10=-10$ $\quad z_0=[0\ 0\ 100\ 0]\begin{bmatrix}0\\0\\1\\-4\end{bmatrix}-10=90$

$f(z_i)=0,\ f(z_f)=1,\ f(z_0)=1$

$\rightarrow c'=0\cdot -4+6\cdot 1=6,\ y=1\cdot 6=6$

t=7 $x=[1\ 1\ 1\ 1]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}1\\1\\1\\1\end{bmatrix}=1$ $\quad z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}1\\1\\1\\1\end{bmatrix}+110=-90$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}1\\1\\1\\1\end{bmatrix}-10=190$ $\quad z_0=[0\ 0\ 100\ 0]\begin{bmatrix}1\\1\\1\\1\end{bmatrix}-10=90$

$f(z_i)=1,\ f(z_f)=0,\ f(z_0)=1$

$\rightarrow c'=1\cdot 1+6\cdot 0=1,\ y=1\cdot 1=1$

t=8 $x=[1\ 0\ 1\ 2]$

$z=[0\ 0\ 0\ 1]\begin{bmatrix}1\\0\\1\\2\end{bmatrix}=2$ $\quad z_f=[-100\ -100\ 0\ 0]\begin{bmatrix}1\\0\\1\\2\end{bmatrix}+110=10$

$z_i=[100\ 100\ 0\ 0]\begin{bmatrix}1\\0\\1\\2\end{bmatrix}-10=90$ $\quad z_0=[0\ 0\ 100\ 0]\begin{bmatrix}1\\0\\1\\2\end{bmatrix}-10=90$

$f(z_i)=1,\ f(z_f)=1,\ f(z_0)=1$

$\rightarrow c'=1\cdot 2+1\cdot 1=3,\ y=1\cdot 3=3$

Ans: $y_t=[0\ 1\ 4\ 4\ 0\ 6\ 1\ 3]$

Word Embedding

$$L = -\log \prod_{c=1}^{C} \frac{\exp(u_c)}{\sum_{i=1}^{V} \exp(u_{c,i})} = -\sum_{c=1}^{C} \log \frac{\exp(u_c)}{\sum_{i=1}^{V} \exp(u_{c,i})} = -\sum_{c=1}^{C} u_c + \sum_{c=1}^{C} \log \sum_{i=1}^{V} \exp(u_{c,i})$$

$$\frac{\partial L}{\partial W_{ij}^{T'}} = \sum_{k=1}^{V} \sum_{c=1}^{C} \frac{\partial L}{\partial u_{c,k}} \frac{\partial u_{c,k}}{\partial W_{ij}^{T'}} \rightarrow \frac{\partial L}{\partial u_{c,j}} = -\delta_{jc} + y_{c,j}, \quad \frac{\partial u_{c,j}}{\partial W_{ij}^{T'}} = \sum_{k=1}^{V} W_{ik}^{T} x_k$$

$$\rightarrow \frac{\partial L}{\partial W_{ij}^{T'}} = \sum_{k=1}^{V} \sum_{c=1}^{C} \frac{\partial L}{\partial u_{c,k}} \frac{\partial u_{c,k}}{\partial W_{ij}^{T'}} = \sum_{c=1}^{C} \frac{\partial L}{\partial u_{c,j}} \frac{\partial u_{c,j}}{\partial W_{ij}^{T'}} = \sum_{c=1}^{C} (-\delta_{jc} + y_{c,j}) \left(\sum_{k=1}^{V} W_{ik}^{T} x_k \right) \#$$

$$\frac{\partial L}{\partial W_{ij}^{T}} = \sum_{k=1}^{V} \sum_{c=1}^{C} \frac{\partial L}{\partial u_{c,k}} \frac{\partial u_{c,k}}{\partial W_{ij}}, \quad u_{c,k} = \sum_{m=1}^{N} \sum_{l=1}^{V} W'_{mk} W_{lm} x_l$$

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases}$$

$$\rightarrow \frac{\partial L}{\partial W_{ij}^{T}} = \sum_{k=1}^{V} \sum_{c=1}^{C} \frac{\partial L}{\partial u_{c,k}} W'_{jk} x_i = \sum_{k=1}^{V} \sum_{c=1}^{C} (-\delta_{kc} + y_{c,k}) W'_{jk} x_i \#$$