**電影評論**

**說明文件**

目錄

[一、 需安裝套件 2](#_Toc40398238)

[1. 需要的所有套件， 2](#_Toc40398239)

[二、 資料準備 2](#_Toc40398240)

[1. 訓練資料 2](#_Toc40398241)

[2. 測試資料(模型已經訓練結束後) 2](#_Toc40398242)

[三、 模型選擇 2](#_Toc40398243)

[1. LSTM 2](#_Toc40398244)

[四、 作法 3](#_Toc40398245)

[1. 訓練資料預處理 3](#_Toc40398246)

[2. 文字資料數據化 3](#_Toc40398247)

[3. Training 4](#_Toc40398248)

[4. 訓練結果 5](#_Toc40398249)

[五、 結果 5](#_Toc40398250)

[1. 模型預測評論 5](#_Toc40398251)

# 需安裝套件

## 需要的所有套件，

keras🡪 pip install keras

tensorflow🡪pip install tensorflow

yaml🡪pip install yaml

# 資料準備

## 訓練資料

正面評論 : 6250則 🡪 取 2成(1250則)用來訓練過程中驗證模型

負面評論 : 6250則 🡪 取 2成(1250則)用來訓練過程中驗證模型

model.fit(x\_train, y\_train, batch\_size = batch\_size, epochs = epochs, verbose = 2, validation\_split = 0.2)

資料內容範例 :

Although this was obviously a low-budget production, the performances and the songs in this movie are worth seeing. One of Walken's few musical roles to date. (he is a marvelous dancer and singer and he demonstrates his acrobatic skills as well - watch for the cartwheel!) Also starring Jason Connery. A great children's story and very likable characters.

## 測試資料(模型已經訓練結束後)

正面評論 : 6250則

負面評論 : 6250則

# 模型選擇

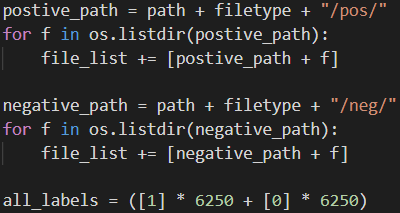
## LSTM

LSTM，是目前RNN(Recurrent Neural Network)中最常使用的模型，考慮到一般的斷詞，透過單詞轉化為向量，只考慮各別單個詞訓練出來的模型，對於一則評論的準確度並不夠精準，因此採用LSTM，透過長短型記憶，來應該每個單詞前後詞出現的情況來影響單詞的意義與特徵，簡單來說就是每個片詞都會互相影響到，考慮電影評論不會因為某幾個字決定整篇文章的正負面，因此採用LSTM此較深度的模型。

# 作法

## 訓練資料預處理

先將正負向資料輸入到同一個List(file\_list)中，然後創立另一個List(all\_labels)將對應的正向標記為1，負向標記為0，如下圖 :

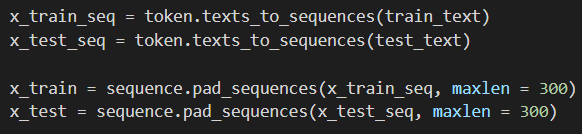


## 文字資料數據化

由於模型吃的資料必須為數字，所以要先將我們的訓練文字轉化為數據，首先要建立數據化的參考字典，因此使用Keras提供的Tokenizer方法，放入全部的訓練資料，並只保留前1000出現次數最多的詞(避免一些語贅詞)，這樣就建立好我們的數據化字典。



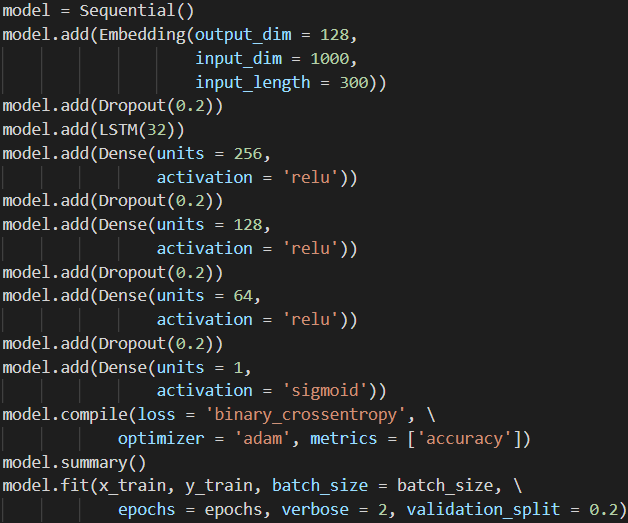
建立好字典之後，就可以把我們的訓練資料轉為數據化，並指定長度為300，超過將以0補滿。

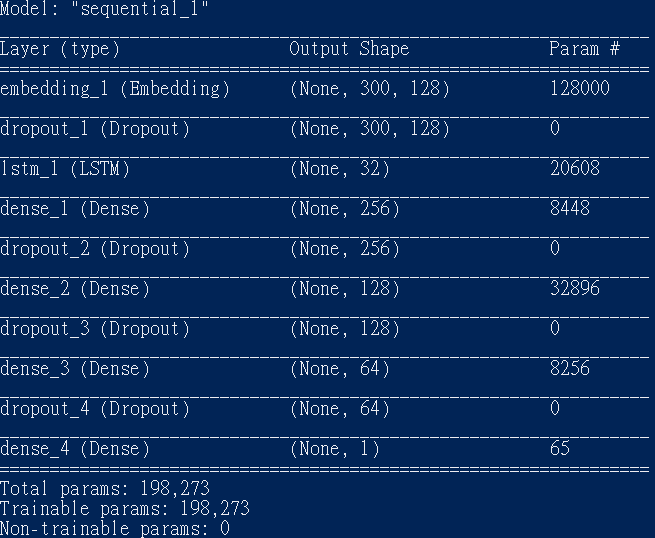


之後就可以開始訓練模型了。

## Training

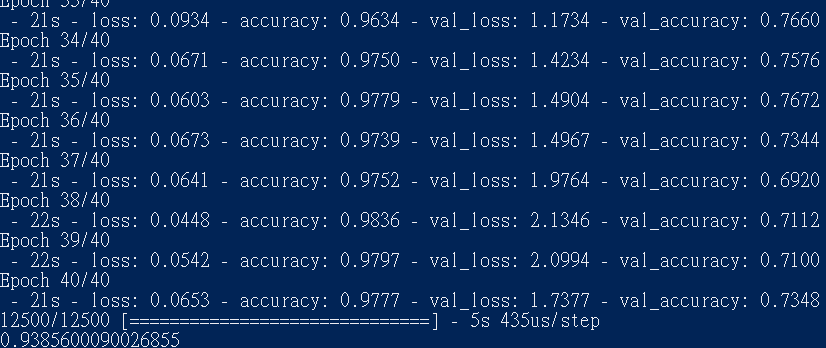
模型的架構如下，先建立一個Embedding層，輸入的維度(input\_dim)為1000，也就是剛剛Tokenizer字典的保留字數量，輸出的維度(output\_dim)為128，主要是經過多次訓練找出最適合的輸出，輸入長度(input\_length)為300，為剛才資料轉為數據之後限制的長度。中途皆採用Dropout將2成的特徵捨去，中途的 Dense層，使用relu來克服梯度消失的問題，最後由於是分類法，因此採用sigmoid來進行分類。模型中使用adam來進行梯度下降的演算。





## 訓練結果

訓練過程中，依多次不斷訓練調教，最終將Batch\_Size設定為512與Epochs設定為40取得目前經驗中最高的準確度(93%)。



# 結果

## 模型預測評論

隨機取一篇小丑電影的負面評論來驗證模型，輸出結果如下 :

