**NLP學習筆記\_用LSTM建立NLP模型**

自然語言處理Natural Language Processing(NLP)為語言學，電腦科學及人工智慧等三大科技文明領域集結之應用。

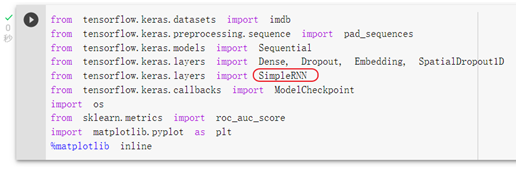
了解了自然語言處理one-hot編碼(one-hot encoding)與詞向量(Word vector)的概念後，我們要試著訓練循環神經網路Recurrent Neural Networks,RNN模型來處理自然語言，RNN不僅可以處理自然語言等依照特定順序出現的資料，連金融資訊，溫度變化，疾病發生率等任何時間序列資料皆可以處理。

**自然語言NLP循環神經網路模型訓練範例**

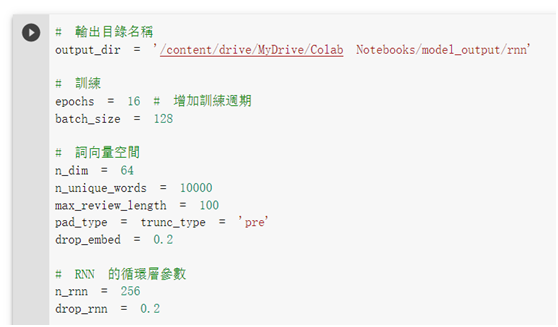
同樣的，我們使用Cola開放平台(可參考文末說明)練習此次的課題。

實作流程如下:

#### 步驟一， 載入模組，使用tf.Keras模組快速建立能夠分析影評內容為正評或負評的RNN模型。



之前以ANN，CNN示範的隱藏層中的神經元網路，此次以SimpleRNN()層取代。

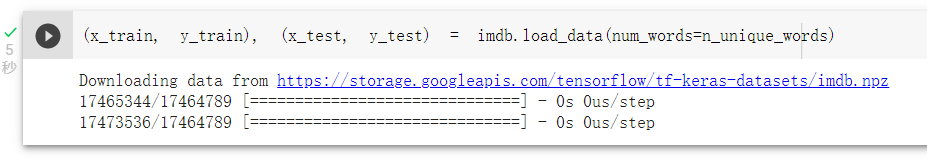


1.設定訓練週期為16。

2.相較於ANN，CNN，使用RNN時詞向量的詞彙需要加倍，所以n\_unique\_words設定為10000個字。

3.RNN在做反向傳播時，梯度可能約10個時步就完全消失，我們先嘗試將max\_review\_length降為100。

#### 步驟二，載入影評資料集



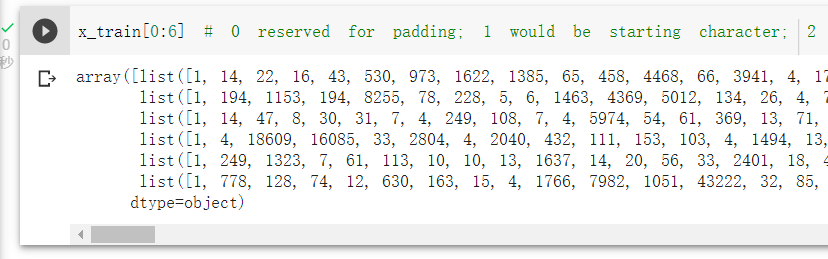
該資料集是Andrew Mass等人於2011年收集lMDb(imdb.com)的影評，作為訓練資料用，共有50000則影評，一半為訓練資料集，另一半是測試資料集。評分等級最高10顆星，二元標籤依據為:

1.4星以下為負評(=0)

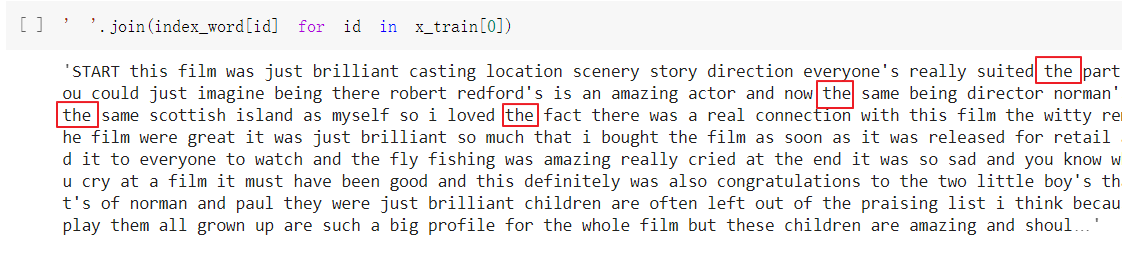
2. 7星以上為正評(=1)

3.5~6星不列入資料集

我們來看每筆影評資料長相如下。

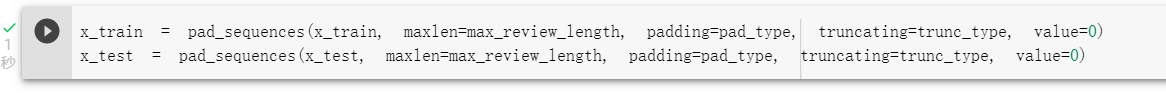
每個單字已被處理成整數格式，值越小代表出現次數越多。

那原文長怎樣呢?

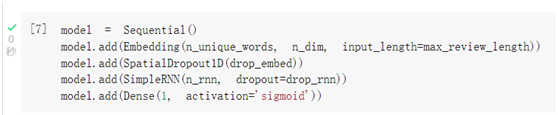


原來the的值為4。

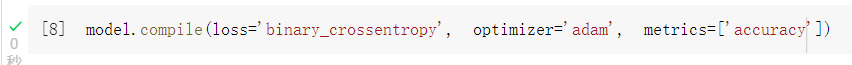
#### 步驟三，資料預處理

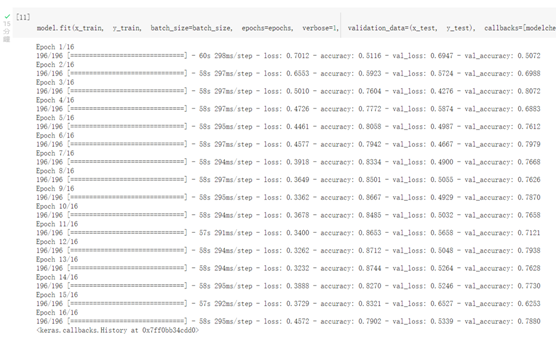


#### 步驟四，RNN模型架構



#### 步驟五，編譯模型及訓練模型



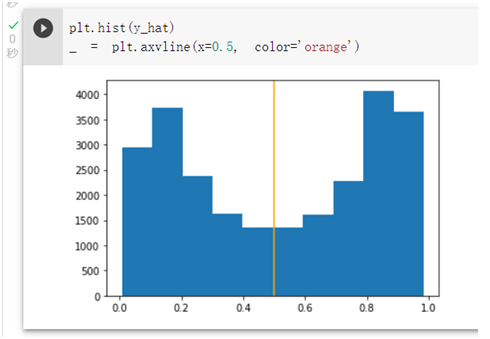


訓練結果以第三周期最佳，故我們要將模型參數回朔至第三周期結束時的值。



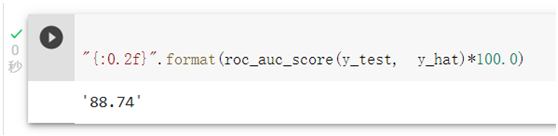
#### 步驟六，模型評估

方法一直方圖



PS:機率超過0.5代表正評。

方法二ROC AUC分數



ROC AUC分數常用於二元分類的評估指標，只要算出來的分數越接近1(我們乘100比較好比較)，代表模型的預測能力越好，本案為88.74分。

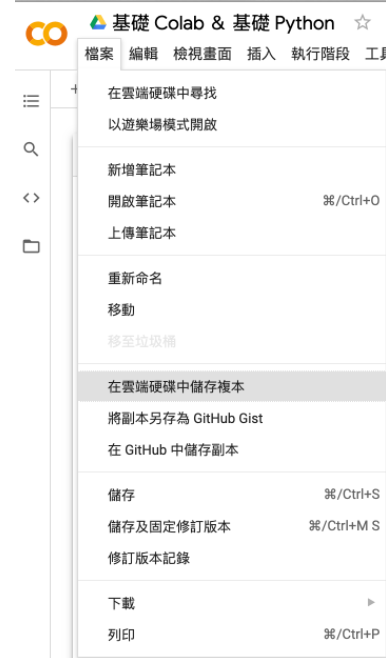
驗證資料(val­\_loss)有增加趨勢，在本案例中單純的RNN在反向傳播時，梯度大約只能撐10個時步，再多會有記憶消失問題，也就是能保持短期記憶，但很難保持長期記憶。

#### **附錄:前置作業Colab環境準備**

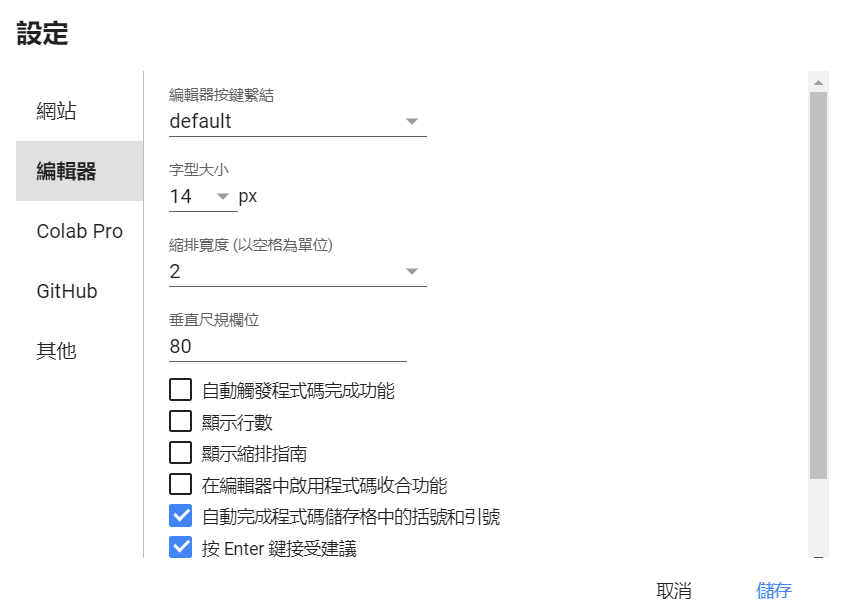
1.必須有自己的google帳號:使用自己的個人的 Google 帳號才可以執行自己 Colab上的程式。

2.先下載 Chrome瀏覽器，並登入 Google 帳號。

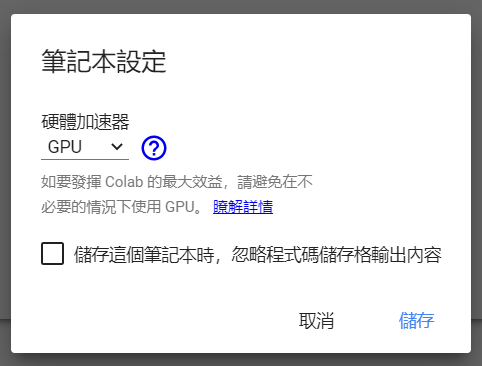
3.建立副本至個人目錄，否則不能存檔。



4.在⼯具 > 設定>編輯器中,反勾選「⾃動觸發程式碼完成功能」。



5.(必須做)設定執行階段類型。



6.療癒模式(非必須)，在⼯具 > 設定>其他, 勾選 「柯基⽝模式」和「貓咪 模式」，就有可愛的貓 貓狗狗來陪你寫程式了。



