**NLP學習筆記\_以函數式API建立非序列式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。

使用Sequential()設計的序列式模型，資料流皆得依序經過每一道神經層，解決方案受限。

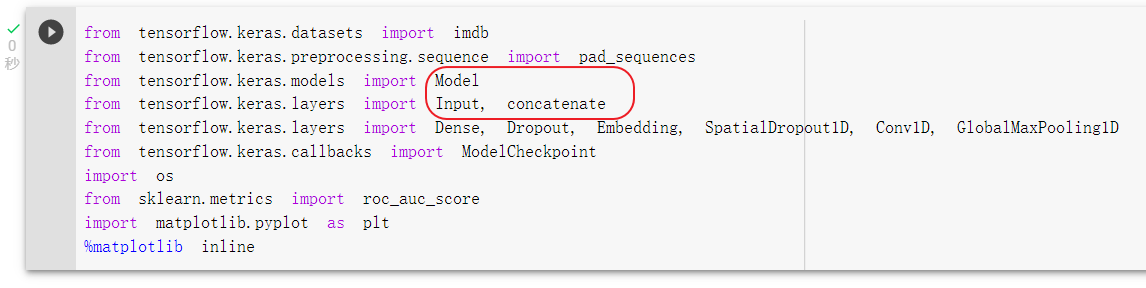
如果使用非序列化模型設計，例如將某神經層輸出的激活值與其他層共享等等設計，是否可以有更好的識別率呢?

**以函數式API建立非序列式自然語言NLP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非序列化CNN分類模型，架構是在嵌入層(Embedding())後加上三道並行的卷積層，讓資料分流計算，而並行的卷積層濾鏡長度不同。



以tf.Keras的函數API取代Sequential()。



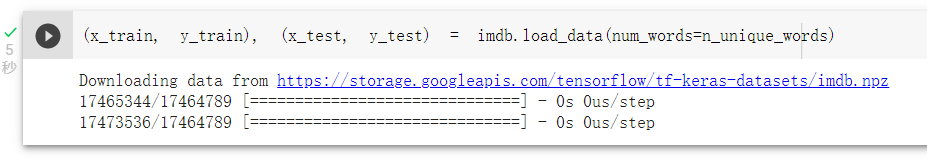
1.設定訓練週期為4。

2.CNN的n\_unique\_words設定為5000個字。

3. max\_review\_length為400。

4.共三並列的卷積層，其濾鏡長度分別為3，2，4。

#### 步驟二，載入影評資料集



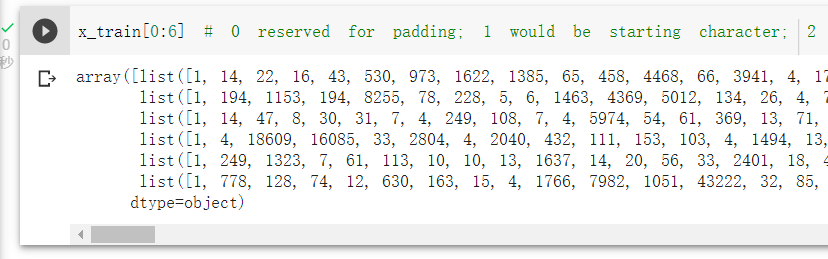
該資料集是Andrew Mass等人於2011年收集lMDb(imdb.com)的影評，作為訓練資料用，共有50000則影評，一半為訓練資料集，另一半是測試資料集。評分等級最高10顆星，二元標籤依據為:

1.4星以下為負評(=0)

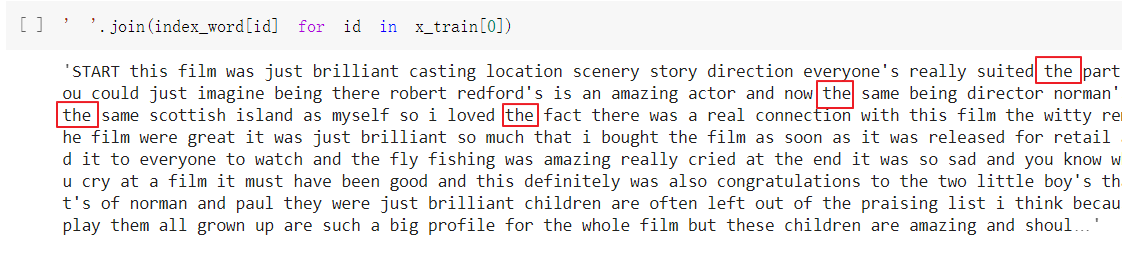
2. 7星以上為正評(=1)

3.5~6星不列入資料集

我們來看每筆影評資料長相如下。

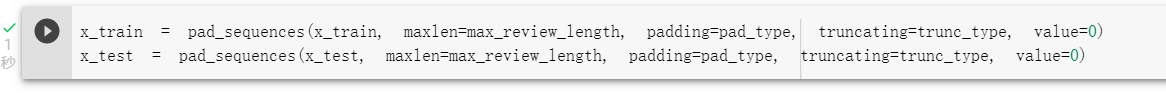
每個單字已被處理成整數格式，值越小代表出現次數越多。

那原文長怎樣呢?



原來the的值為4。

#### 步驟三，資料預處理

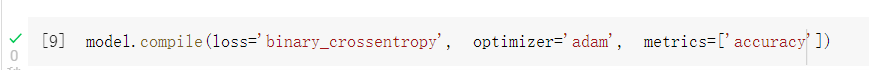


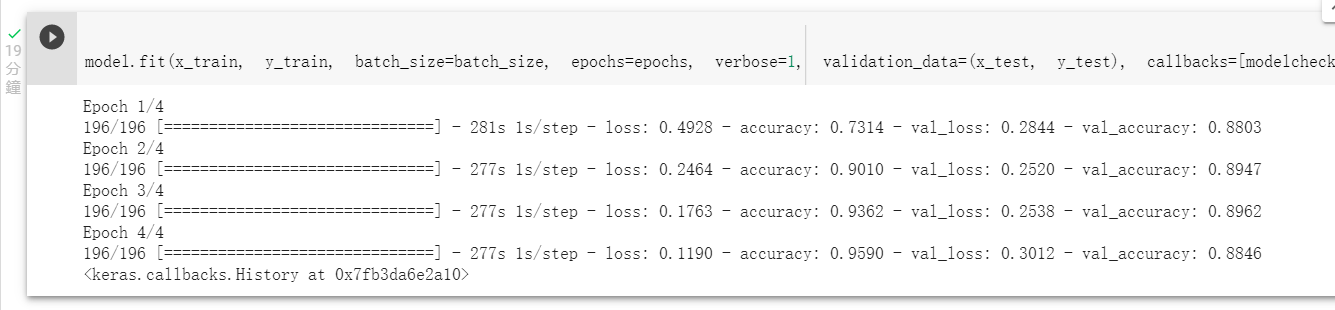
#### 步驟四，非序列化神經網路架構



並行的三個卷積層各由一Conv\_ID層+GlobalMaxPooIingID層組合而成。

#### 步驟五，編譯模型及訓練模型



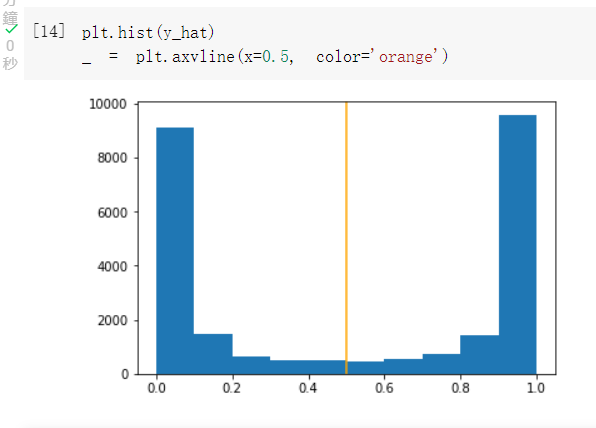
****

訓練結果以第三周期準確率0.8962 & Val\_loss:0.2538最佳，故我們要將模型參數回朔至第三周期結束時的值。



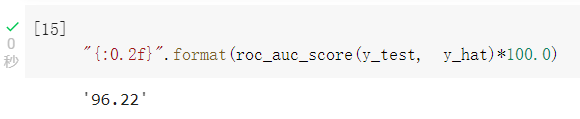
#### 步驟六，模型評估

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96.22分。

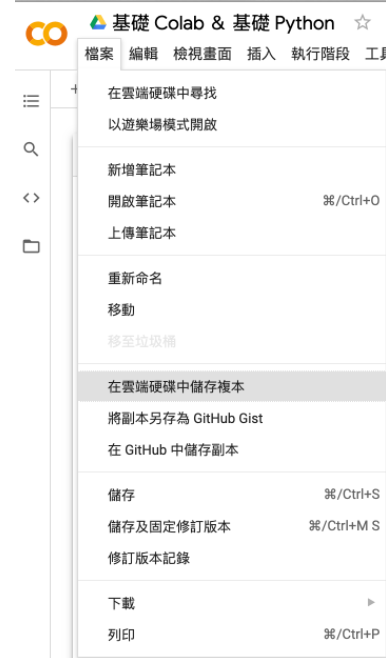
結果，在資料集的規模不大的情況下，非序列化函數式API的CNN模型會是一個不錯的選擇。

#### **附錄:前置作業Colab環境準備**

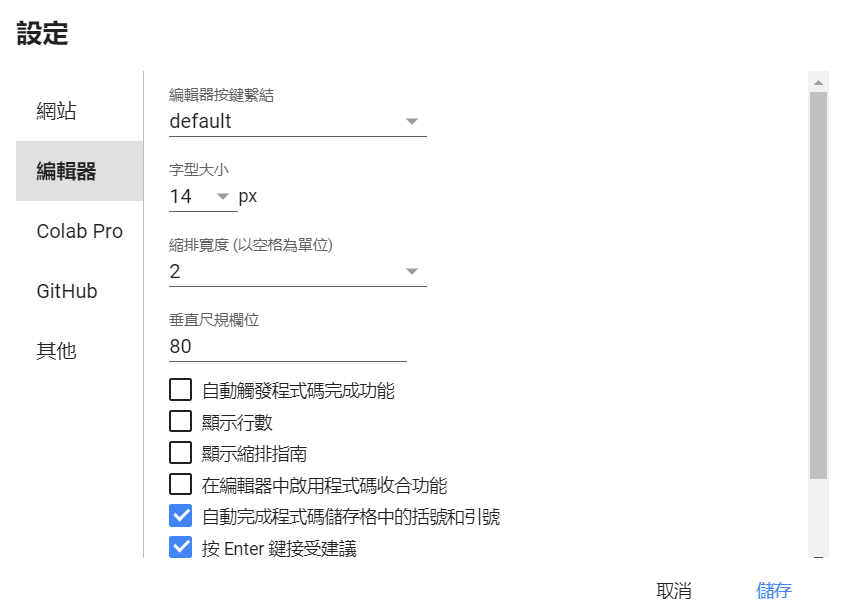
1.必須有自己的google帳號:使用自己的個人的 Google 帳號才可以執行自己 Colab上的程式。

2.先下載 Chrome瀏覽器，並登入 Google 帳號。

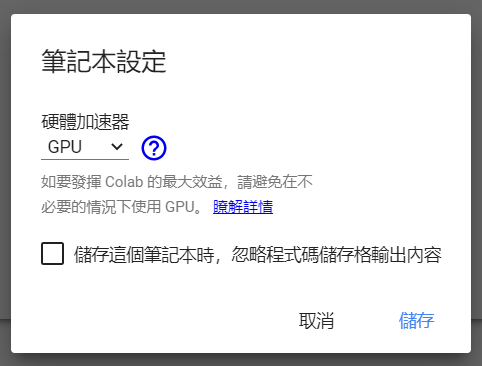
3.建立副本至個人目錄，否則不能存檔。



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。



