Bert Chinese Model

1. 使用套件

整個訓練過程是使用Huggingface這個公司開發的套件，huggingface是個紐約的聊天機器人初創服務商，在nlp領域有大型的開源社群，這次的專案我從其中的預訓練模型庫Transformer拿了一個中文的模型來進行訓練得到結果。

1. 處理資料

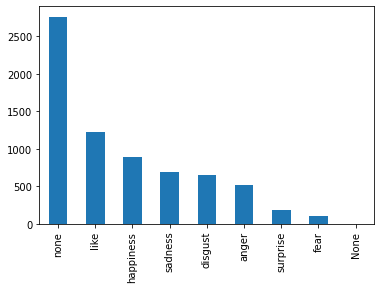
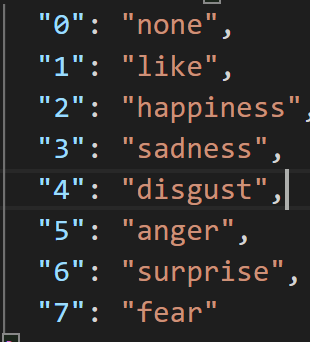
原始資料集是簡體微博貼文資料集，原始資料已經有情感的標註。其格式為一個貼文下有好幾條句子，且每句句子皆有情感標註，而總共有2000筆貼文。經過討論以及處理後，整理成包含了訓練資料集以及測試資料集，格式皆為一行情緒label一行句子，訓練集有5616筆資料，測試集有1404筆資料，並且已用openCC轉換成了繁體資料，以契合之後的習慣輸入用語。

資料型態如下，皆打包成tsv檔以方便之後拿去訓練

一張含有 文字 的圖片

自動產生的描述

資料集中的情緒標記分布 這是label的分類

1. 訓練過程

這次拿出來再訓練的模型是bert-base-chinese，bert模型的訓練過程需要先將資料轉換為bert可以被餵入的資料型態，在處理bert的資料時需要回傳四種list，所以為了方便使用，同樣也是用huggingface自帶的分詞工具

一張含有 文字 的圖片

自動產生的描述

並且因為data是用dataloader的形式load進來的，label和sentence分開處理，如下圖

一張含有 文字 的圖片

自動產生的描述

最終整理出適合被拿去訓練的資料集，之後開始訓練

如同上面提到的，我使用了bert-base-chinese的模型作為我的預處理模型，並且下去使用我的資料集來訓練。

一張含有 文字 的圖片

自動產生的描述

訓練結果如下 訓練了5個epoch

一張含有 文字 的圖片

自動產生的描述

訓練玩了之後將模型存到資料夾裡，以便日後可以拿去做我們的機器人對話，可以看到這裡資料夾分別存了三個檔案，即為我們會用到的資料。

一張含有 文字 的圖片

自動產生的描述

最後作一些model性能好不好的評估，雖然在這裡準確率只有60%，但實際上測的時候我們覺得這樣的準確率還可以在聊天的範圍內接受，所以沒有改變。

一張含有 文字 的圖片

自動產生的描述

1. 提取模型

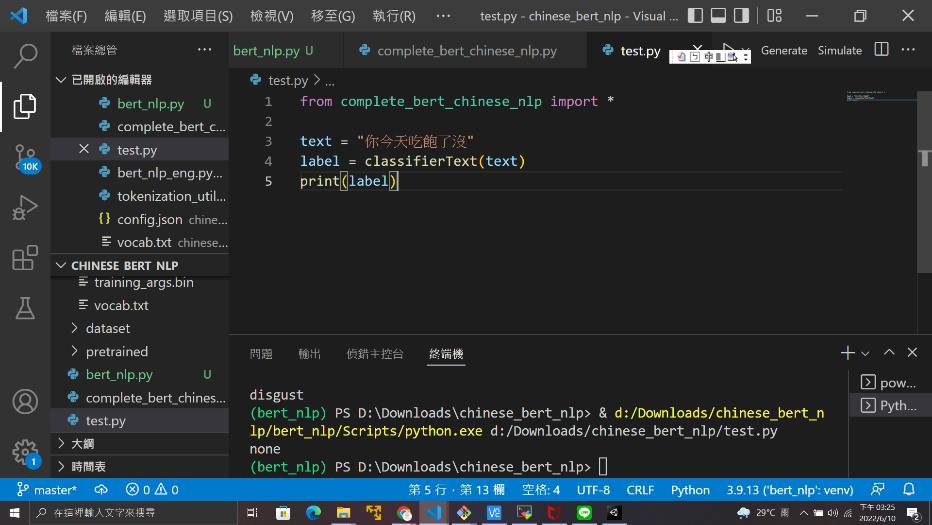
在完成了訓練之後，我們擁有了以下格式的資料夾

一張含有 文字 的圖片

自動產生的描述

就可以用來測試輸入的text是否為合理的label了

這是最後的label輸出結果，藉由這個function來跟機器人的接口來做結合



這是實際上如何call自己建好的model的函式部份 一張含有 文字 的圖片

自動產生的描述

1. 參考條目
2. 簡體微博貼文資料集 : <http://tcci.ccf.org.cn/conference/2013/pages/page04_tdata.html>
3. Huggingface 官網 : <https://huggingface.co/>

負責組員 : 吳應玉