影評自然語言分析

專題編號：108-CSIE-S014
執行期限：107 年第 1 學期至 108 年第 1 學期
指導教授：劉傳銘
專題參與人員： 105590022 劉聖鴻
105590032 張哲源
105590037 杜晉宇

一、摘要

本專題主旨在於利用 IMDB 網路資料庫中已被整理且標記完成的資料集經過資料前處理後，用以訓練已建構完成的深度學習模型，使此模型能對使用者所輸入的影評文字進行評論正負評的判斷，且將此功能嵌入以 Django 為框架的網頁來對電影分析其評論的正評率、討論熱度、綜合排名等等。

關鍵詞：deep learning、Django、RNN、Python、Tensorflow、keras。

二、緣由與目的

食衣住行育樂為人民主要的六大需求，而在近年影視產業發展迅速，時常會有許多不同類型的電影公開上映，許多強檔大片更是同學、朋友間熱烈討論的話題，在現今收看影視作品是極為方便的，走進電影院、亦或是收看線上的頻道，皆是常見的方法，但並不是每部電影都是極其完美的，有些電影討論熱度較低，但其極具教義意義、有些電影討論熱度很高，但實際上卻是我們俗稱的大爛片，因此我們小組想著手架出一個電影資訊平台，其中又可讓使用者於平台上留下觀後心得，並且提供給使用者電影的正評、熱度排行做為參考，使用者亦可參與討論。
且此類型的專案應用的領域十分多元及富含商業價值，如：可用來提早得知顧客對公司或產品觀感，以即時調整銷售策略方向。

三、使用技術方法

（一）Model 部份：

1.進行資料前處理：先載入資料處理套件如下：pandas、numpy、matplotlib；檔案路徑處理套件如下：urllib.request、os、tarfile；儲存 json 檔案處理套件：json。

首先到 IMDB 上的資料集網址將我們所要用來訓練模型的資料載下來，並且進行資料處理，處理完後將資料正規化，接下來我們建立一個字典來存放我們訓練集影評資料中，最常出現的前三千個單字，且經過停用詞的處理，並且將這個字典儲存成一個 json 檔案以供使用。

2.建立深度學習模型：

載入 Tensorflow、keras 後，分別使用 RNN 和 LSTM 對上方所創建出的兩千字的字典進行訓練，建造出模型且儲存下來以供嵌入網站做評論正負判斷之用，最後使用 model.evaluate() 來對模型評估準確度。

（二）網站部份：

1.網站部分，採用的是 Django 是一個開放原始碼的 web 應用框架，由 Python 寫成。採用了 MVT 的軟體設計模式，Django 內建也提供了很多內建套件，像可延伸的認證系統、動態站點管理頁面、分頁器，以及資料庫等。

2. BeautifulSoup4，用來將 IMDB 網站上的電影評論爬蟲下來，以進行正負平分析，並加進網站中給予參與總結評分。

四、架構流程

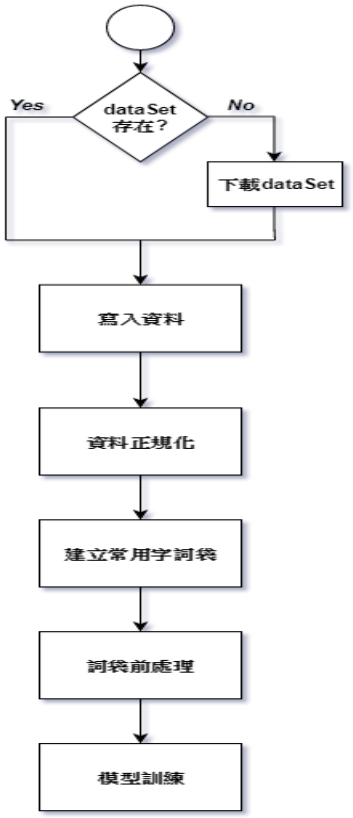

圖 1.資料處理流程圖

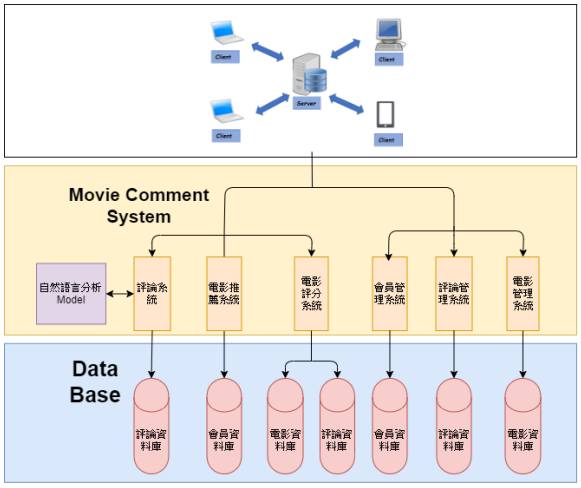

圖 2.系統架構流程圖

使用者於一開始註冊成為會員後可選擇個人的電影喜好，網站會依個人偏好與當前電影評分，推薦適合觀看的電影給使用者，瀏覽該電影後亦可對該電影進行評論，網站系統會自動判別其為正負評，也會對該電影的總評分有影響，判別正負評的系統，是來用已訓練好的模型，此模型的訓練集用於IMDB網站上公開的資料集去做自然語言的分析，再將訓練完的模型和詞袋載入到網站的後端去進行使用。該網站後端亦具備資料庫管理，隨時對資料庫的資料進行管理，包括新增電影、刪除評論等，而使用者則只能進行註冊、新增評論、瀏覽網站。

五、實驗結果

模型的部份經過實驗則發現，對詞袋進行停用詞處理、冗贅詞處理、時態處理後，預測準確度明顯較高，而同樣的 batchSize 下，進行多次訓練後準確率反而下降，則應是因為外部訓練的資料多樣性不足，無法有效的使準確率提升。

RNN									
Batch_Size	100	100	100	64	64	64	32	32	32
epochs	5	10	15	5	10	15	5	10	15
外部訓練準確度	78.68%	77.68%	76.18%	75.12%	76.80%	77.56%	81.22%	76.56%	81.62%
未處理準確度	77.12%	73.98%	69.84%	76.46%	72.80%	74.58%	71.56%	76.46%	77.46%
LSTM									
Batch_Size	100	100	100	64	64	64	32	32	32
epochs	5	10	20	5	10	15	5	10	15
外部訓練準確度	85.76%	83.74%	76.76%	80.66%	75.94%	81.18%	79.90%	80.94%	81.04%
未處理準確度	76.82%	81.22%	74.42%	75.72%	74.06%	77.86%	73.36%	75.32%	76.96%

圖 3.模型訓練紀錄

目前網站使用的部分仍在試驗階段，還未上線，目前僅有參與專題的幾位同學使用過該網站。

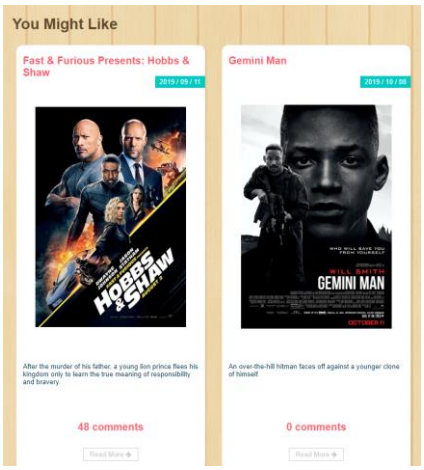

圖 4 .網站主頁

參考文獻

[1] Keras - IMDb 網路電影資料集：http://puremonkey2010.blogspot.com/2017/09/toolkit-keras-imdb.html

[2] Django 使用指南教學: https://djangogirlstaipei.gitbooks.io/django-girls-taipei-tutorial/