# 自然語言處理專案實作：電影評論分析

製作者：龔慕祥

# 專案大綱

- 前言
- 資料前處裡
- 類神經網路模型設計
- 真實電影評論實際應用

# 前言

- 隨著人工智慧技術的迅猛發展，自然語言處理在電腦科學領域取得了令人矚目的成就。通過訓練和優化模型，我們能夠賦予電腦一定程度的語言理解能力，從而實現對語句意義的精確判斷。

- 在本專題中，我們選擇了Kaggle上的電影評論數據集作為訓練和測試數據來源。該數據集涵蓋了各種類型的電影評論，包括正面評價、負面評價和中立評價等。我們打算運用循環神經網絡（RNN）結合LSTM模型，通過對數據集的訓練和優化，使模型能夠準確識別和分類不同類型的電影評價。

- 透過該專題的實施，我們期望能夠探索並提升自然語言處理在電影評論分析方面的應用能力。這將有助於影院、網絡平台等電影相關業務更好地理解觀眾對電影的喜好和評價，從而提供更優質的用戶體驗和相關推薦服務。

# 資料前處裡

取得資料集

文字前處理

文字轉數字

# 取得資料集

從 KAGGLE 取得

SENTIMENT ANALYSIS ON MOVIE REVIEWS資料集

| | Phrase | Sentiment |
|---|---|---|
| | A series of escapades demonstrating the adage ... | 1 |
| | A series of escapades demonstrating the adage ... | 2 |
| | A series | 2 |
| | A | 2 |
| | series | 2 |

```
[156060 rows x 2 columns]>
```

```
# 0 - negative
# 1 - somewhat negative
# 2 - neutral
# 3 - somewhat positive
# 4 - positive
```

# 文字前處理

將常見的 stop words 與非文字去除

```
[原本] Hi! My name is Eric.
[原本] Hello Eric, it is nice to meet you!
[後來] hi name eric
[後來] hello eric nice meet
```

# 文字轉數字

使用 Tokenize 手法把人類讀的文字改成機器讀的數字，並且利用 Pad 手法讓每則評論長度相同(補零)

**原文**

- series escapades demonstrating adage good goose also good gander occasionally amuses none amounts much story

**Tokenize**

- [217, 13863, 5839, 6387, 8, 2819, 91, 8, 10657, 550, 9874, 460, 1980, 12, 6]

**Pad**

- [0, 0, 0, 0, 0, 0, 0, 0, 0 ,0, 0, 0, 0, 0, 0 217, 13863, 5839, 6387, 8, 2819, 91, 8, 10657, 550, 9874, 460, 1980, 12, 6]

# 循環神經網路-LSTM模型的設計

## 循環神經網路 RNN(LSTM)

- Embedding : 128 (讓模型去了解不同文字之間的關係)
- Bidirectional(LSTM) : 256 (此單元輸出的矩陣長度，由於是雙向因此為 256*2 = 512)
- 30 -> 512 (節點由30個增加成512個，再丟進DNN訓練)
- activation="relu"

## 深度神經網路 DNN

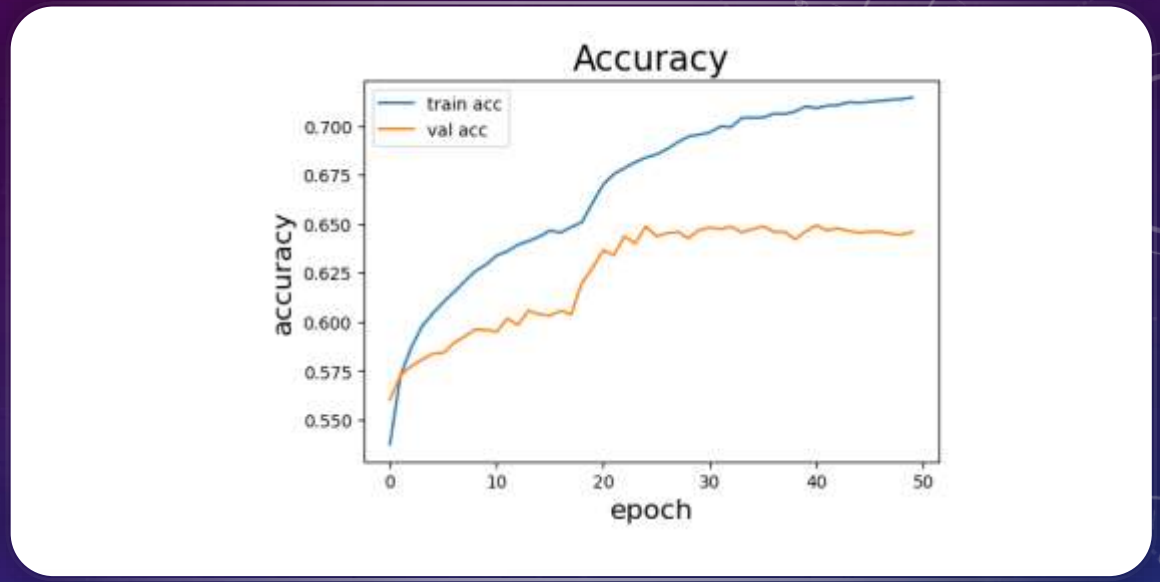
- Layer1 128units
- Layer2 5units
- activation="relu"

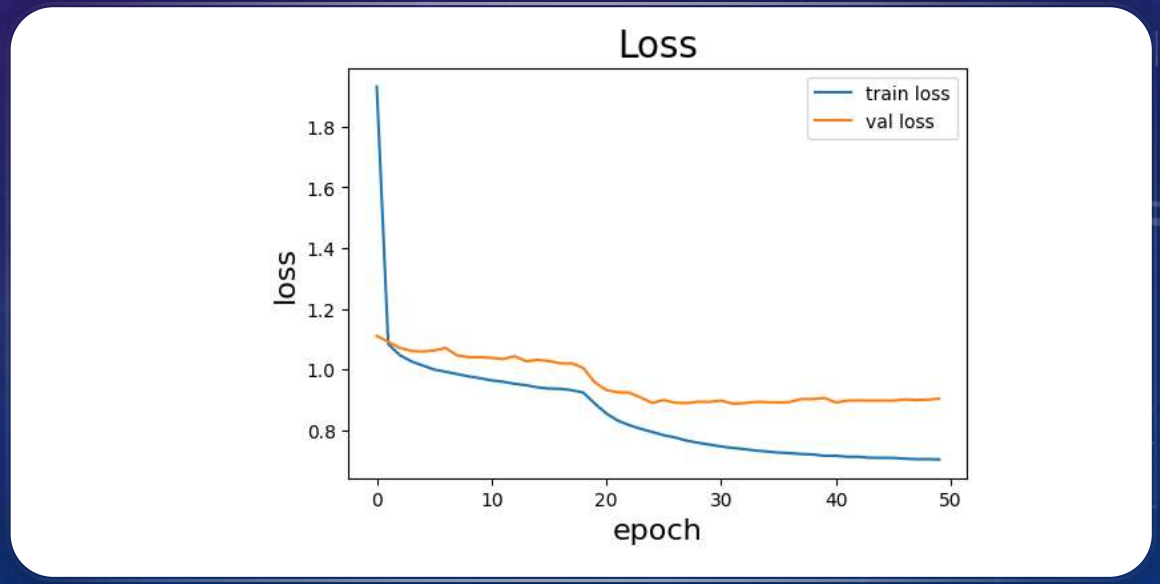
## 輸出層

- units=5
- activation="softmax"
- 1 x 5 (由於5個類別的分類問題，最終輸出為五個節點)

# 模型績效

- 由圖可知在模型訓練過程，準確率持續上升；LOSS持續下降
- 可以說明模型算是成功的



# 真實評論實際應用

The film was okay, but nothing special. It didn't really leave a lasting impression.("neutral")

- somewhat positive

It has all the classic elements that make a really good romantic comedy with two very charming leads.("positive")

- positive

I found the movie to be disappointing. The acting was subpar and the plot was predictable.("negative")

- negative

# 結論&未來展望

- 綜合以上結果，本專案的自然語言處理模型在文本分類方面雖未無法100%正確判斷評論評價，但還是有一定的判斷水準。未來，我們可以進一步優化模型，針對參數進行調整，以期望取得更好的結果。
- 以下是一些未來改進和展望的方向：

1. 資料增強：通過增加更多多樣性的訓練數據，包括不同領域、不同風格的文本，可以增加模型的泛化能力。此外，也可以利用生成模型生成合成數據，以擴充訓練集的規模和多樣性。
2. 模型集成和融合：採用模型集成和融合技術，如模型投票、模型融合等方法，可以結合多個不同的模型，從而提高整體預測的穩定性和準確性。
3. 參數調整和優化：進行系統性的參數調整和優化，例如優化學習率、正則化項、批量大小等，可以使模型更快收斂並提高泛化能力。