# 深度學習 Final Compotetion

0852617 統計所碩一 曾鈺評

## 一、資料前處理

1. 初始資料如下，發現有部分資料的 keyword 顯示 NaN，我們在後續步驟會考慮將 NaN 予以去除

| | ID | label | label_name | title | keyword |
|---|---|---|---|---|---|
| 0 | 0 | 0 | news_entertainment | 古力娜扎再次成为焦点，这一身招摇大方，搞获了网友们的心 | 古力娜扎,粉丝 |
| 1 | 1 | 0 | news_entertainment | 如果张国荣，张学友，陈百强，王杰要排位的话怎么排？ | NaN |
| 2 | 2 | 0 | news_entertainment | 包贝尔带娇妻外出就餐被拍，大家把注意力放在了第3张！ | 娇妻,娇妻外出就餐,包贝尔 |
| 3 | 3 | 0 | news_entertainment | 娱乐圈娶了豪门的5位男星，事业开挂，最后一位想离婚门都没有 | 豪门,迟重瑞,周立波,吕良伟,石贞善 |
| 4 | 4 | 0 | news_entertainment | 陈学冬参加节目被爆童年照，果然胖子都是潜力股啊！ | 薛之谦,陈学冬,谭维维,跨界歌王 |

2. 觀察資料類別，發現大致上每個類別的資料量還算平均

| label | ID | label_name | title | keyword |
|---|---|---|---|---|
| 0 | 31506 | 31506 | 31506 | 31506 |
| 1 | 29982 | 29982 | 29982 | 29982 |
| 2 | 14141 | 14141 | 14141 | 14141 |
| 3 | 28676 | 28676 | 28676 | 28676 |
| 4 | 21572 | 21572 | 21572 | 21572 |
| 5 | 33310 | 33310 | 33310 | 33310 |
| 6 | 19987 | 19987 | 19987 | 19987 |
| 7 | 21530 | 21530 | 21530 | 21530 |
| 8 | 15500 | 15500 | 15500 | 15500 |
| 9 | 23425 | 23425 | 23425 | 23425 |

3. 利用 jieba 為 title 與 keyword 切成一個個詞，並蒐集起來當作字典

| | id | title | keyword | title_tokenized | keyword_tokenized |
|---|---|---|---|---|---|
| 0 | 0 | 新能源汽车充电桩的投资潜力和价值大吗？ | NaN | 新能源 汽车 充电 桩 的 投资 潜力 和 价值 大 吗 | |
| 1 | 1 | 黄磊问惠若琪：你们奥运会的金牌是纯金的吗？惠若琪耿直回应！ | 奥运会,运动员,黄磊,里约奥运,惠若琪 | 黄磊 问 惠若琪 你们 奥运会 的 金牌 是 纯金 的 吗 惠若琪 耿直 回应 | 奥运会 运动员 黄磊 里 约 奥运 惠若琪 |
| 2 | 2 | 宜家也有微信小程序了，O2O转化率提升142% | 宜家,宜家家居,程序,测量费,微信支付 | 宜家 也 有 微信 小 程序 了 O2O 转化率 提升 142 | 宜家 宜家 家居 程序 测量 费 微信 支付 |
| 3 | 3 | 南宁市户籍小学毕业生注意事项！ | 户口簿,西乡塘区,青秀区,监护人,人口居住证,不动产权证书,三代同堂,房屋所有权证 | 南宁市 户籍 小学 毕业生 注意 事项 | 户口簿 西乡 塘区 青秀区 监护人 人口居住证 不动产 权 证书 三代同堂 房屋所有权证 |
| 4 | 4 | 印度厕所“所长”，每月500块钱工资养5个小孩，吃住都在厕所！ | 印度,印度人,厕所所长,厕所,排泄物 | 印度 厕所 所长 每月 500 块钱 工资 养 5 个 小孩 吃住 都 在 厕所 | 印度 印度人 厕所 所长 厕所 排泄物 |

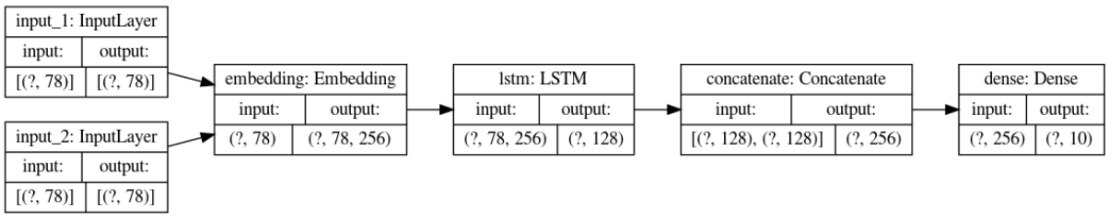
4. 基本參數設置

| 變數名稱 | 意義 | 設定 |
|---|---|---|
| NUM_CLASSES | 有幾個分類 | 10 |
| MAX_NUM_WORDS | 在語料庫裡有多少詞彙 | 500000 |
| MAX_SEQUENCE_LENGTH | 一個標題最長有幾個詞彙 | 78 |
| NUM_EMBEDDING_DIM | 一個詞向量的維度 | 256 |
| NUM_LSTM_UNITS | LSTM 輸出的向量維度 | 128 |

二、模型架構

因為詞為序列資料，故選用 LSTM 當作我們的訓練模型，架構如下：

1. Input layer: 有兩層，分別輸入 title 與 keyword 切成詞後所對應字典的數字序列(長度為 78)
2. Embedding: 對應字典的數字序列轉換成詞向量(78*256，因為設定一個向量的長度為 256)
3. LSTM: 訓練模型後輸入為 128 的長度
4. Concatenate: 以上我們將 title 與 keyword 的資料分開訓練，模型用的是同一個，在這裡我們再將訓練結果串接起來
5. Dense: 全連接層，最後用 softmax 預測每一類發生的機率，再選最大的機率預測是哪一類




【模型 summary】

| Layer (type) | Output Shape | Param # | Connected to |
|---|---|---|---|
| input_1 (InputLayer) | [(None, 78)] | 0 | |
| input_2 (InputLayer) | [(None, 78)] | 0 | |
| embedding (Embedding) | (None, 78, 256) | 128000000 | input_1[0][0]<br>input_2[0][0] |
| lstm (LSTM) | (None, 128) | 197120 | embedding[0][0]<br>embedding[1][0] |
| concatenate (Concatenate) | (None, 256) | 0 | lstm[0][0]<br>lstm[1][0] |
| dense (Dense) | (None, 10) | 2570 | concatenate[0][0] |

Total params: 128,199,690
Trainable params: 128,199,690
Non-trainable params: 0

三、結果與討論

1. Validation 的比例有嘗試用 0.1 與 0.2，但似乎沒有太大差別
2. batch size 選用 256, 512, 1024，發現 1024 表現得不好，256 與 512 則是沒差太多
3. Optimizer 選用 Adam 或 RMSprop 結果差不多，但 RMSprop 似乎稍為有相對好一點點
4. 由於此模型參數量多，大概訓練 3-4 個 epoch 便已達到訓練的極限，再訓練下去可能會有 overfitting 的問題
5. Epoch 從 1 到 10，Training set(藍線)與 validation set(橘線)的 accuracy 如下所示