- 智慧化企業整合期末報告 -

YouTube 熱門影片趨勢分析

109034540 胡心玫

CONTENTS

研究動機與目的

近年來有越來越多人投入Youtuber這個行業，為了成為成功的Youtuber，勢必得先了解Youtube發燒影片趨勢。

5W1H

What?

Youtube熱門影片趨勢分析

Where?

美國

When?

2017-2018

Why?

想要成為成功Youtuber勢必得了解Youtube發燒影片趨勢。

Who?

Youtuber

How?

資料預處理、LSTM、相關性分析與資料可視化、

研究方法

Trending YouTube Video Statistics
使用Kaggle的公開訓練資料集

LSTM
適合解時間序列問題

Python-keras
使用Python keras套件建模

相關性分析
透過相關性分析挑選適合的特徵

資料集描述

- Trending YouTube Video Statistics
  - 選擇**美國**的資料進行分析

1. **Video_id:** identification code for the YouTube Video
2. **Trending_date:** Date on which the Video was Trending
3. **Title:** Title of the YouTube Video
4. **Channel_title:** Title of the YouTube Channel uploading the Video
5. **Category_id:** Unique ID of the Video's Category (e.g. Entertainment, Music, Sports)
6. **Publish_time:** Date and Time in which the video was published
7. **Tags:** Hashtags added to the Video to make it easier to find by the public
8. **Views:** Number of Views the Video Obtained by the time the dataset was downloaded
9. **Likes:** Number of Likes the Video Obtained by the time the dataset was downloaded
10. **Dislikes:** Number of DisLikes the Video Obtained by the time the dataset was downloaded
11. **Comment_count:** Number of Comments the Video Obtained by the time the dataset was downloaded
12. **Thumbnail_link:** Thumbnail link of the Video (image representing a link)
13. **Comments_disabled:** Dummy Variable indicating whether the comments were disabled on the video
14. **Ratings_disabled:** Dummy Variable indicating whether the ratings were disabled on the video
15. **Video_error_or_removed:** Dummy Variable indicating the presence of a video error or removal of the video
16. **Description:** a piece of metadata that helps YouTube understand the content of a video. Well optimized descriptions can lead to higher rankings in YouTube search.

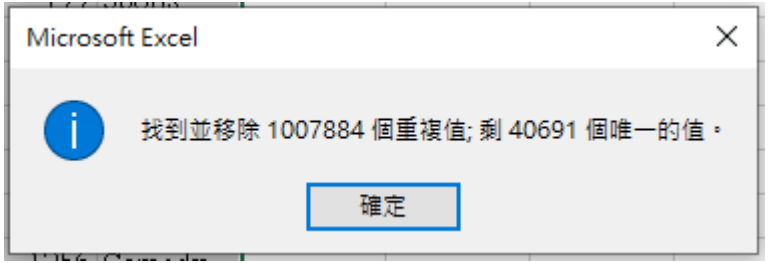
csv檔包含16種欄位，共40949筆資料

category_id	category_title	country_code
1	Film & Animation	US
2	Autos & Vehicles	US
10	Music	US
15	Pets & Animals	US
17	Sports	US
18	Short Movies	US
19	Travel & Events	US
20	Gaming	US
21	Videoblogging	US
22	People & Blogs	US
23	Comedy	US
24	Entertainment	US
25	News & Politics	US
31	Anime/Animation	US
32	Action/Adventure	US
33	Classics	US
34	Comedy	US
35	Documentary	US
36	Drama	US
37	Family	US
38	Foreign	US
39	Horror	US
40	Sci-Fi/Fantasy	US
41	Thriller	US
42	Shorts	US
43	Shows	US
44	Trailers	US

json檔經過處理後可得category_id對應的category_title及country_code

刪除、補值及標準化

- 利用Excel確認有無缺值，將重複的資料列刪除，計算tag的數量，將日期呈現方式標準化，最後輸出新的csv檔。



相關性分析與資料可視化

- 刪除不相干的欄位：thumbnail_link（縮圖連結）、comments_disabled（是否允許評論）、ratings_disabled（是否允許評分）、video_error_or_removed（影片是否損壞或移除）。
- 計算連續型資料和views的相關係數

	views
views	1
likes	0.850315
dislikes	0.472267
comment_count	0.619633
tags num	-0.02918

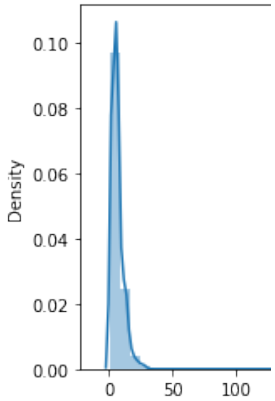
相關係數

✓ likes和views的相關係數大於0.8，納入訓練模型之特徵

相關性分析與資料可視化

- ✓ views為時間序列資料，因此將views的過去資料也納入訓練模型之特徵

	count
count	6282.0000
mean	6.518465
std	6.771645
min	1.000000
25%	3.000000
50%	6.000000
75%	8.000000
max	397.00000



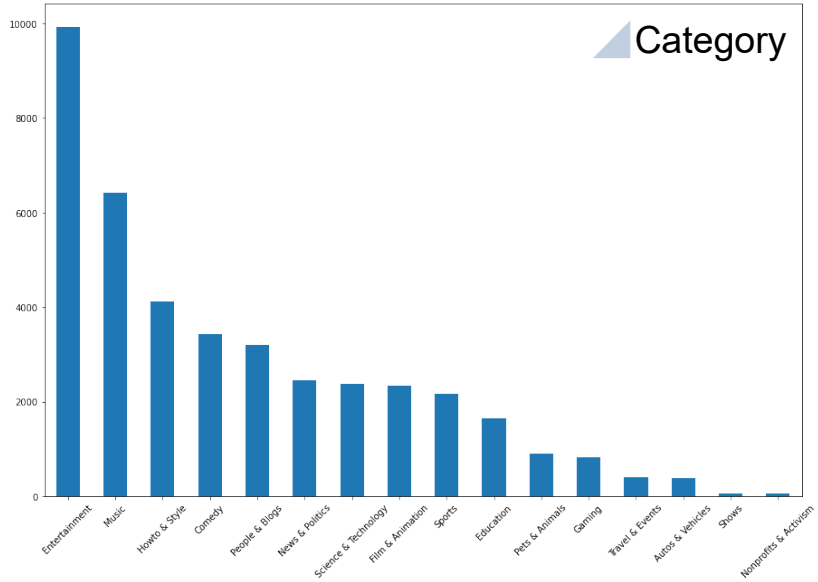
連續上榜天數

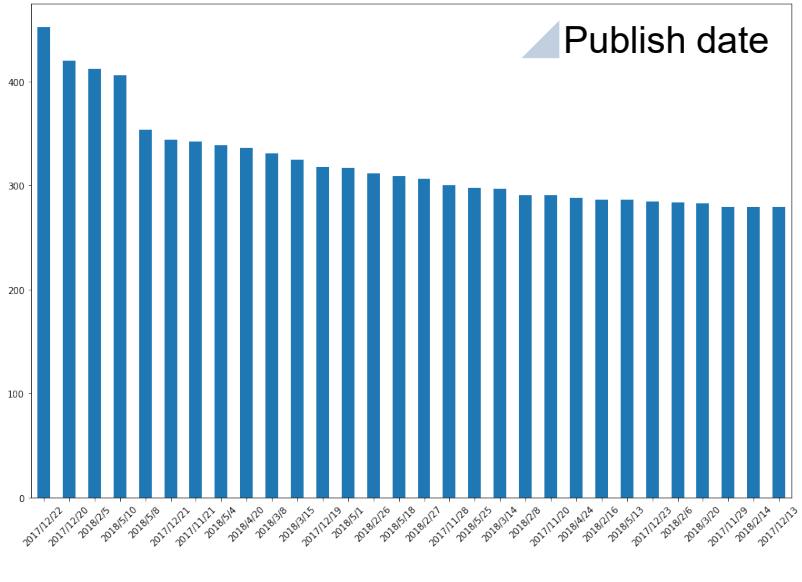
超過75%的熱門影片會連續上榜3天以上，因此我們設定time step=4天

相關性分析與資料可視化

- 對非連續型的資料做直方圖

最後僅挑選出兩項變數：likes和views(time step=4)
最後輸入模型的資料大小為(3942,4,2)





模型架構

```
model = Sequential()
model.add(LSTM(64, return_sequences=True,input_shape=(9, 2),activation='tanh'))
model.add(Dropout(0.2))
model.add(LSTM(128,return_sequences=False,activation='tanh'))
model.add(Dense(1, activation='linear'))#1 is output

optimizer = RMSprop(lr=0.005)
model.compile(loss='mean_squared_error', optimizer=optimizer)
model.summary()
```

- 透過加入Dropout層來減少模型過擬合發生的機率。

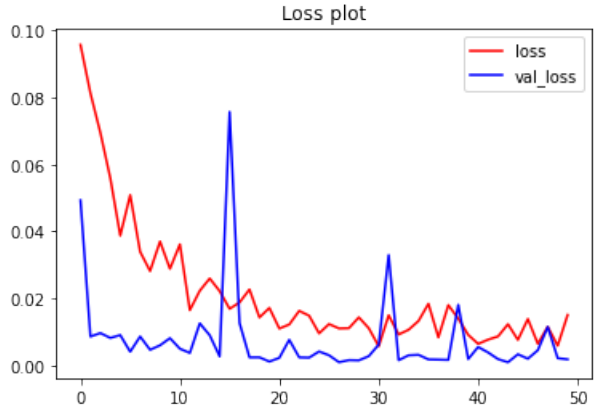
Model: "sequential_4"

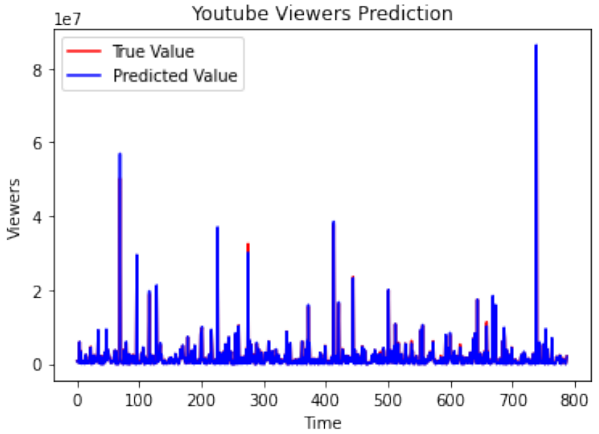
Layer (type)	Output Shape	Param #
lstm_8 (LSTM)	(None, 9, 64)	17152
dropout_4 (Dropout)	(None, 9, 64)	0
lstm_9 (LSTM)	(None, 128)	98816
dense_4 (Dense)	(None, 1)	129

Total params: 116,097
Trainable params: 116,097
Non-trainable params: 0

模型訓練成果

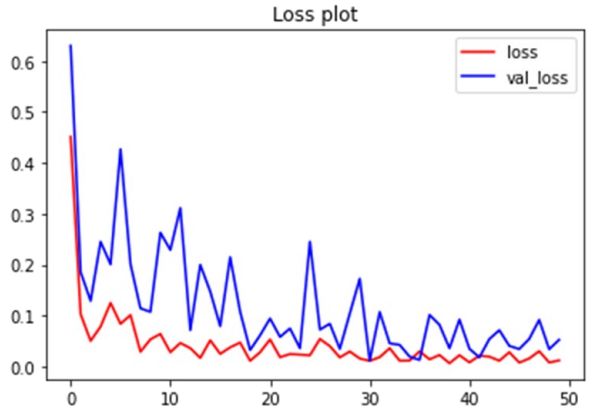
- 訓練集的 loss和驗證集的val_loss如所示，雖然偶有發生val_loss>loss的情形，但是當val_loss下降時，loss適時上升，最後收斂到0.02以下，代表此模型確實可以持續調整有效避免過擬合發生，並且可以精準預測目標值。

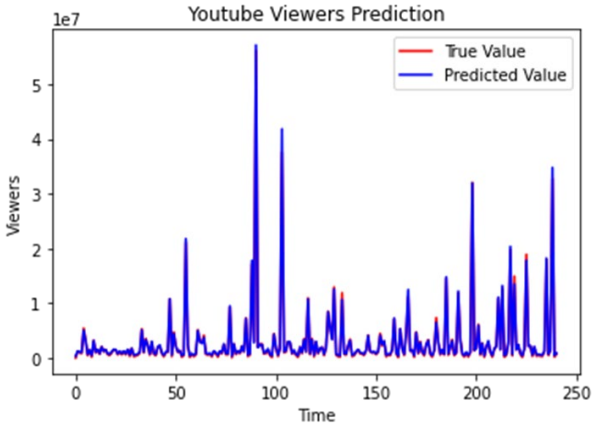




參數調整

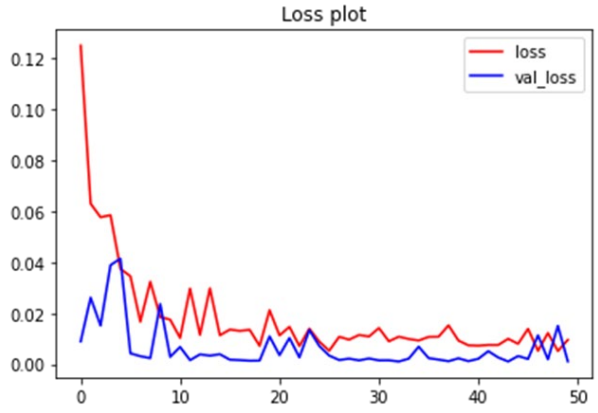
- 輸入likes及views(time step=9)

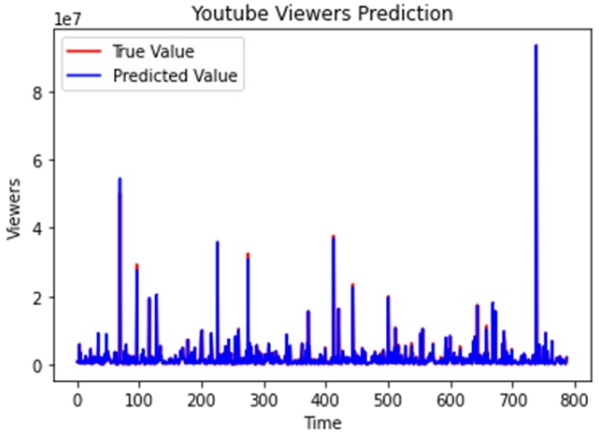




參數調整

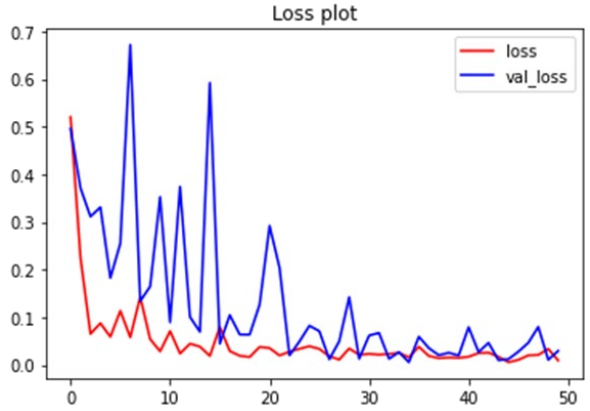
- 輸入views(time step=4)

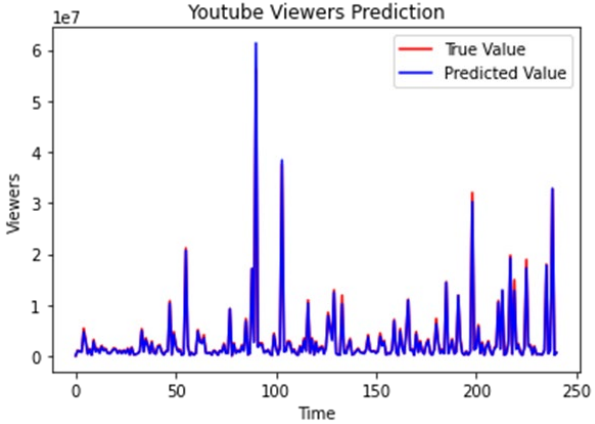




參數調整

- 輸入views(time step=9)





結論

- 以上的參數組合中，表現最好的是「輸入likes及views(time step=4)」，推測是因為連續上榜超過四天的熱門影片數較多，所以可以訓練的資料較多，因此預測較為精準。

	Train Score(RMSE)	Val Score(RMSE)
輸入likes及views(time step=4)	367757.64	338026.47
輸入likes及views(time step=9)	956439.82	408942.70
輸入views(time step=4)	484946.44	393396.12
輸入views(time step=9)	815038.07	480950.99

參考資料

- LSTM Prediction on Trending YouTube Videos Views
- YouTube热门视频榜单Excel数据分析
- LSTM_深度學習_股價預測
- ML Lecture 21-1: Recurrent Neural Network (Part I)

THANKS!

Do you have any questions?