
統計學習與深度學習 期末專案

Jack Team

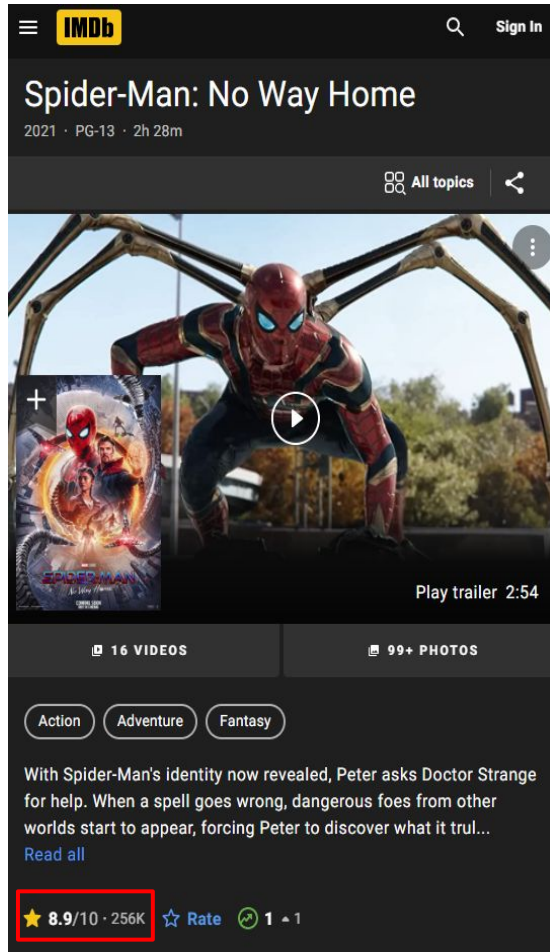
陳韋傑、石子仙、莊啟宏
周鈺淇、王博奕、江雨柔

大綱

- 任務簡介
- 資料集來源
- 資料探索與資料前處理
- 模型架設
- 結果探討

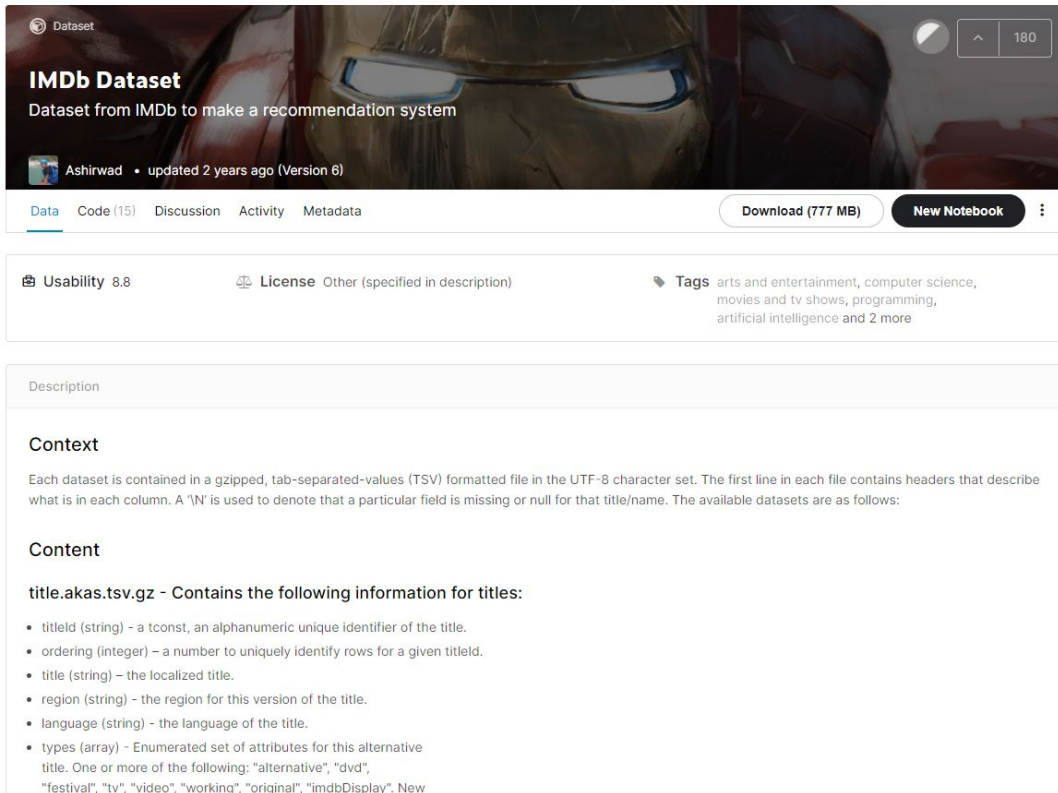
任務簡介

- 預測 IMDb 電影評分
 - 使用年份、片長、風格、演員等等變數
- Actor Embedding Learning
 - 使用 Actor Embedding 來取代 One-hot Encoding
- 比較不同演員的性質異同
 - 使用 Actor Embedding 來計算相似度



資料集來源

- Kaggle: IMDb Dataset
 - title.basics.tsv
 - title.ratings.tsv
 - title.principals.tsv



The screenshot shows the Kaggle page for the IMDb Dataset. The header features a background image of Iron Man's helmet with the text "IMDb Dataset" and "Dataset from IMDb to make a recommendation system". Below this, it says "Ashirwad • updated 2 years ago (Version 6)". A navigation bar includes links for "Data", "Code (15)", "Discussion", "Activity", and "Metadata", along with buttons for "Download (777 MB)" and "New Notebook".

Below the navigation bar, there are three sections: "Usability 8.8", "License Other (specified in description)", and "Tags" which include "arts and entertainment, computer science, movies and tv shows, programming, artificial intelligence and 2 more".

The "Description" section is expanded, showing the following content:

Context

Each dataset is contained in a gzipped, tab-separated-values (TSV) formatted file in the UTF-8 character set. The first line in each file contains headers that describe what is in each column. A 'N' is used to denote that a particular field is missing or null for that title/name. The available datasets are as follows:

Content

title.akas.tsv.gz - Contains the following information for titles:

- **titleid** (string) - a tconst, an alphanumeric unique identifier of the title.
- **ordering** (integer) - a number to uniquely identify rows for a given titleid.
- **title** (string) - the localized title.
- **region** (string) - the region for this version of the title.
- **language** (string) - the language of the title.
- **types** (array) - Enumerated set of attributes for this alternative title. One or more of the following: "alternative", "dvd", "festival", "tv", "video", "working", "original", "imdbDisplay". New

資料探索 – title.basics.tsv

資料筆數:6326545

欄位數:9

有空值的資料筆數:6299893

欄位名稱	空值個數	唯一值個數
tconst	0	6326545
titleType	0	10
primaryType	10	3175907
originalType	10	3191724
isAdult	0	2
startYear	371185	147
endYear	6272061	94
runTimeMinutes	4459212	X
genres	501660	X

資料探索 – title.ratings.tsv

資料筆數: 993821

欄位數: 3

資料缺失值: 0

欄位名稱	空值個數	唯一值個數
tconst	0	993821
averageRating	0	91
numVotes	0	17360

資料探索 – title.principals.tsv

資料筆數:36499704

欄位數:6

有空值的資料筆數:36499704

欄位名稱	空值個數	唯一值個數
tconst	0	5710740
ordering	0	10
nconst	0	3873199
category	0	12
job	30505144	32211
characters	17965033	2207711

資料前處理

- 利用 tconst(電影 id)欄位合併三個檔案
- averageRating為應變數(Y)，其他欄位皆為自變數(X)
- 資料篩選：
 - 1960 年後上映
 - 總時長 60 分鐘以上
 - 總評分票數 15 票以上
- 新增欄位 cast(演員)、crew(幕後人員)
 - 在 title.principals.tsv 中的 category 欄位，依其分類歸為演員或幕後人員
 - 為每一個電影新增兩個欄位，記錄該電影所有的演員以及幕後人員
- 新增欄位 change_name
 - 若電影的 primaryTitle(主要標題)與 originalTitle(原有標題)不同，該欄位為 1，否則為 0

資料探索 – 合併後資料

資料筆數: 203521

欄位數: 14

資料缺失值: 29146 (保留缺失值)

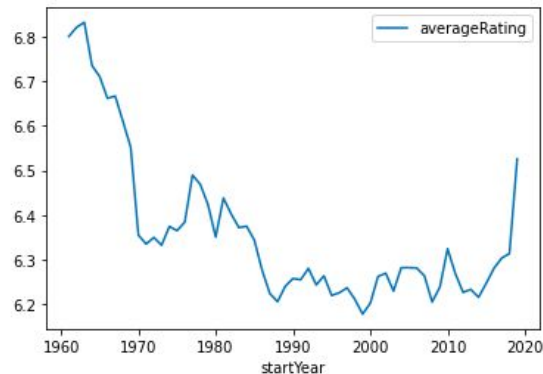
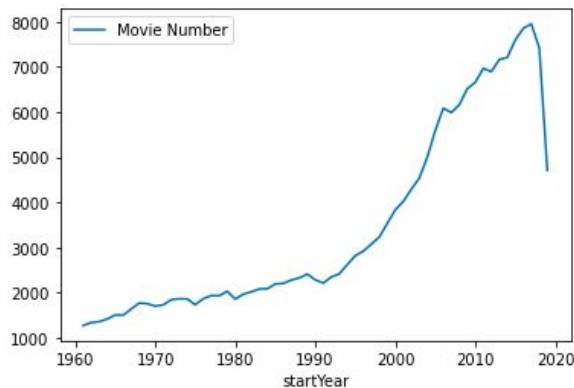
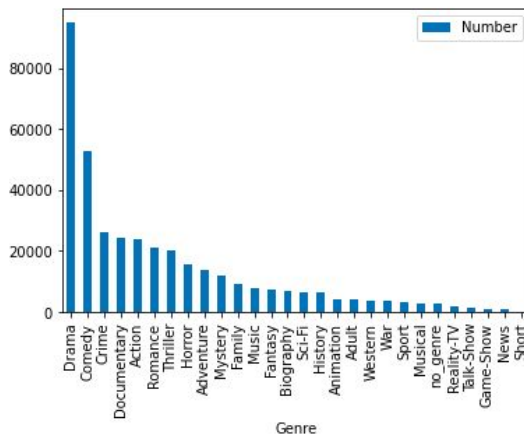
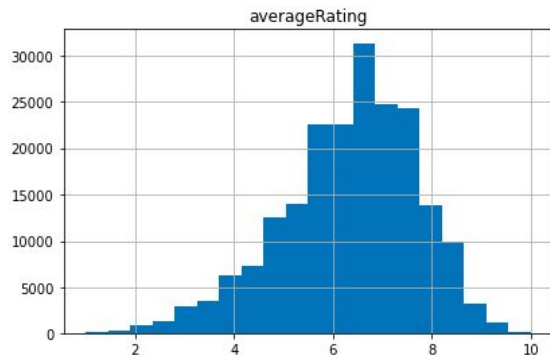
所有電影的分數總平均: 6.31 分

欄位名稱	空值個數	唯一值個數
cast	23068	X
crew	6078	X
change_name	0	2

欄位名稱	空值個數	唯一值個數
tconst	0	203521
titleType	0	10
primaryType	0	179791
originalType	0	184669
isAdult	0	2
startYear	0	59
endYear	0	62
runTimeMinutes	0	X
genres	0	X
averageRating	0	91
numVotes	0	15433

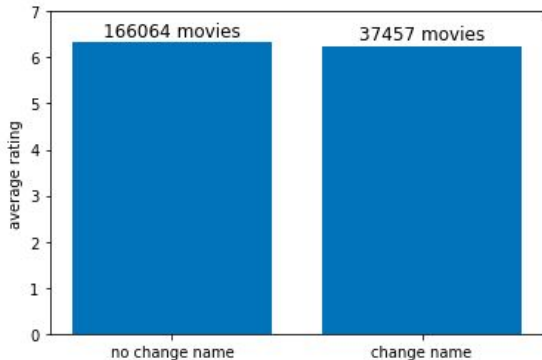
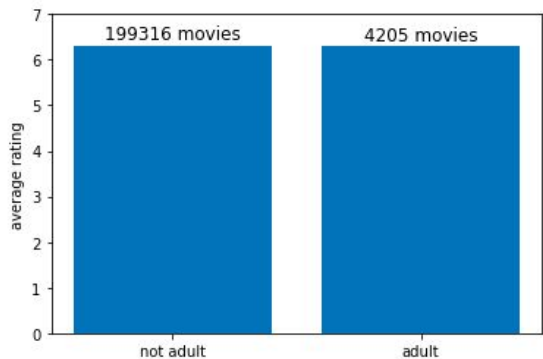
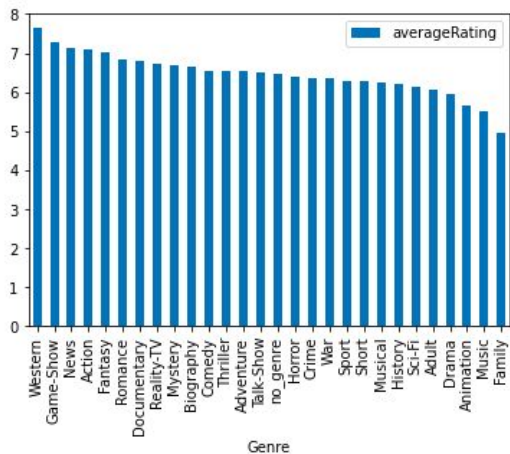
資料分布情形

- 大部分電影的平均分數落在 5 - 8 分之間
- 類型最多的電影為 Drama, 第二為 Comedy
- 近年越來越多電影上映, 然而電影分數卻越來越低



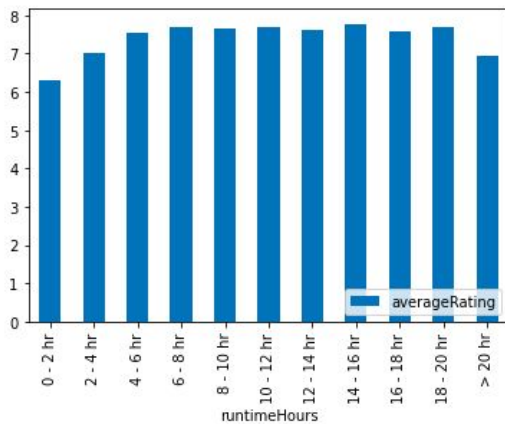
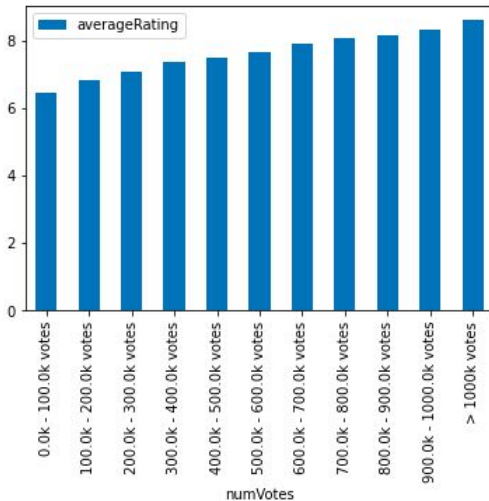
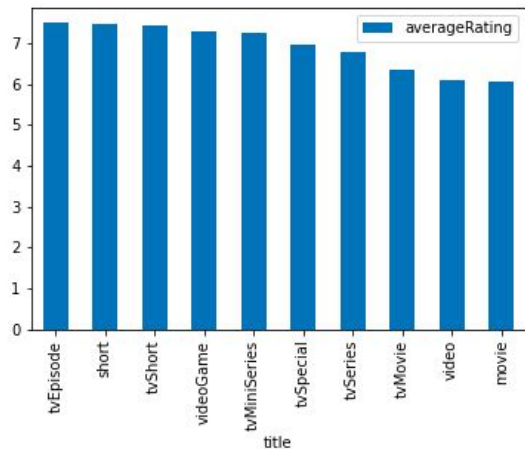
各欄位與平均分數關係

- 平均分數最高的電影種類為 Western, 最低為 Family
- 非成人電影遠多於成人電影, 兩者平均分數相近
- 無論是否改過電影標題, 兩者平均分數相近



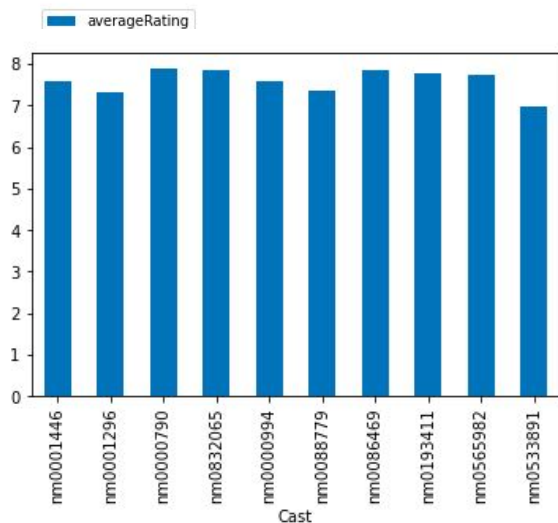
各欄位與平均分數關係

- 平均分數最高的類型為電視劇(tvEpisode), 最低為電影(movie)
- 票數越高的電影, 平均分數也越高
- 總長度在 4 小時以下、20 小時以上的電影, 平均分數略低



演出前 10 高的演員 (cast)

- 演出前 10 高的演員平均分數都在 7 分以上且差異不大



Michael Landon
nm0001446



Lorne Greene
nm0001296



James Arness
nm0000790



Milburn Stone
nm0832065



Raymond Burr
nm0000994



Dan Blocker
nm0088779



Amanda Blake
nm0086469



Ken Curtis
nm0193411



Doug McClure
nm0565982

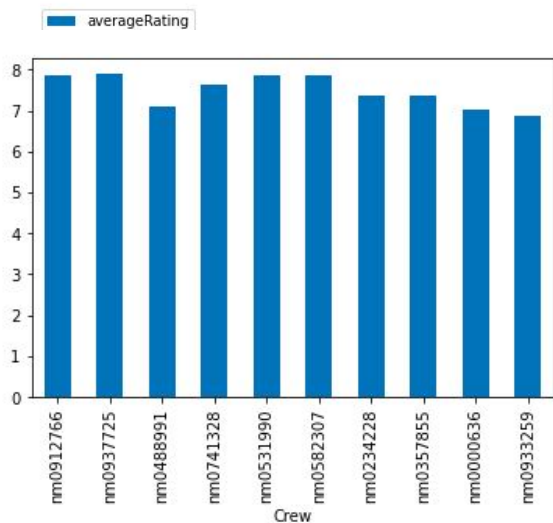


Gavin MacLeod
nm0533891



參與前 10 高的幕後人員 (crew)

- 參與前 10 高的幕後人員平均分數有些接近 8 分，只有一位低於 7 分



Charles Marquis Warren
nm0912766



Dick Wolf
nm0937725



Glen A. Larson
nm0488991



David Rose
nm0741328



Norman MacDonnell
nm0531990



John Meston
nm0582307



David Dortort
nm0234228



Fred Hamilton
nm0357855



William Shakespeare
nm0000636



Dave Wilson
nm0933259



預測模型

- 訓練資料: 基本資料集
 - 不包含cast和crew
- 使用基本資料集作為Baseline
- 用以比較加入cast/crew之成效



模型 / 評分	基本資料集	
	adjusted R ²	RMSE
Linear regression	0.298	1.159
RigdeCV	0.299	1.159
LassoCV	0.299	1.160
ElasticNetCV	0.296	1.161
Random Forest	0.391	1.091
XGBoost	0.487	1.030
NN	-	1.088

預測模型

- 訓練資料: 加入cast/crew
 - 使用 one-hot encoding 表示
- 耗用過多資源
 - 記憶體容量不足, 有數十萬個欄位
 - 無法運算
- 以出演頻率篩選, 降低維度
 - 設定最小出演次數為 50
 - 剩下1000個欄位左右



模型 / 評分	加入cast/crew(one-hot)	
	adjusted R ²	RMSE
Linear regression	0.351	1.115
RigdeCV	0.351	1.114
LassoCV	0.328	1.128
ElasticNetCV	0.322	1.133
Random Forest	0.381	1.091
XGBoost	0.466	1.026
NN	-	1.075

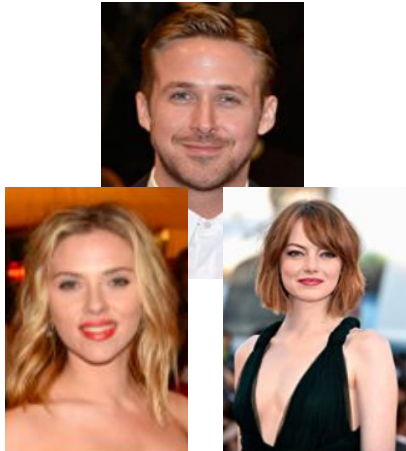
模型架設 – Word2Vec

- One-hot Encoding 維度過多、資訊量較低，難以訓練
- 使用 gensim.models.word2vec 來訓練 Actor Embedding
- 使用演員們共同出演電影的紀錄作為訓練資料
- 將一部電影視作一個句子，一個演員視作一個詞語
- 其他類似模型：
 - Fasttext - 考慮 sub-word information, 不適合此任務
 - BERT - 考慮 context information, 不適合此任務

$$pr(w_o|w_i) = \frac{\exp(v_{w_o}^T v_{w_i})}{\sum_{j=1}^W \exp(v_{w_o}^T v_{w_j})}$$

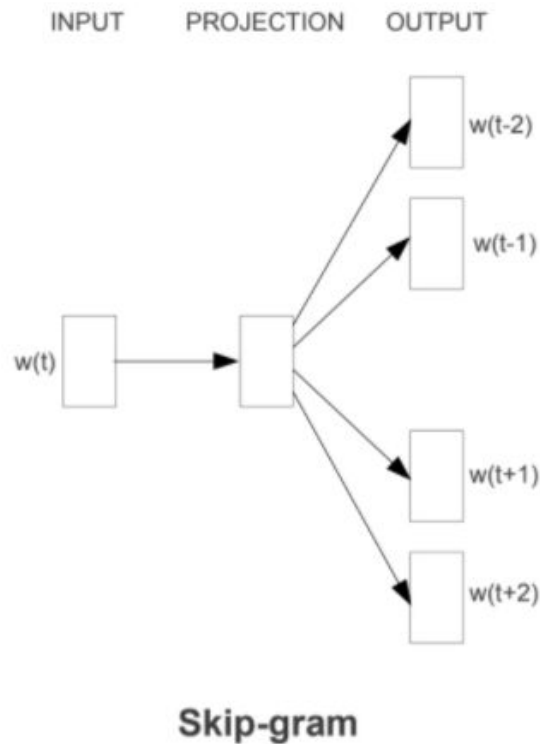
模型架設 – Word2Vec

- After Parameters Tuning:
 - 使用Skip-gram Algorithm
 - Min Count = 50
 - Dimensions = 300



Word2Vec

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_{n-1} \\ a_n \end{bmatrix}$$



預測模型

- 訓練資料: 加入cast/crew
 - 使用word2vec表示
- 有效降低使用的資源
 - 訓練時間、所需記憶體下降
- 仍保有良好的預測能力



模型 / 評分	加入cast/crew (one-hot encoding)		加入cast/crew (word2vec)	
	adjusted R ²	RMSE	adjusted R ²	RMSE
Linear regression	0.351	1.115	0.364	1.102
RigdeCV	0.351	1.114	0.364	1.102
LassoCV	0.328	1.128	0.358	1.106
ElasticNetCV	0.322	1.133	0.355	1.108
Random Forest	0.381	1.091	0.409	1.081
XGBoost	0.466	1.026	0.565	0.977
NN	-	1.075	-	1.008

預測模型

- 訓練資料: 加入cast/crew與title
 - 使用word2vec表示
 - 將電影名稱納入考量
- 使用Pre-trained BERT
 - 將title轉為Embedding
- 考量所有現有資訊



模型 / 評分	加入cast/crew與title (w2v+bert)	
	adjusted R^2	RMSE
Linear regression	0.374	1.099
RigdeCV	0.373	1.096
LassoCV	0.363	1.100
ElasticNetCV	0.360	1.102
Random Forest	0.467	1.082
XGBoost	0.606	0.979
NN	-	1.016

預測模型

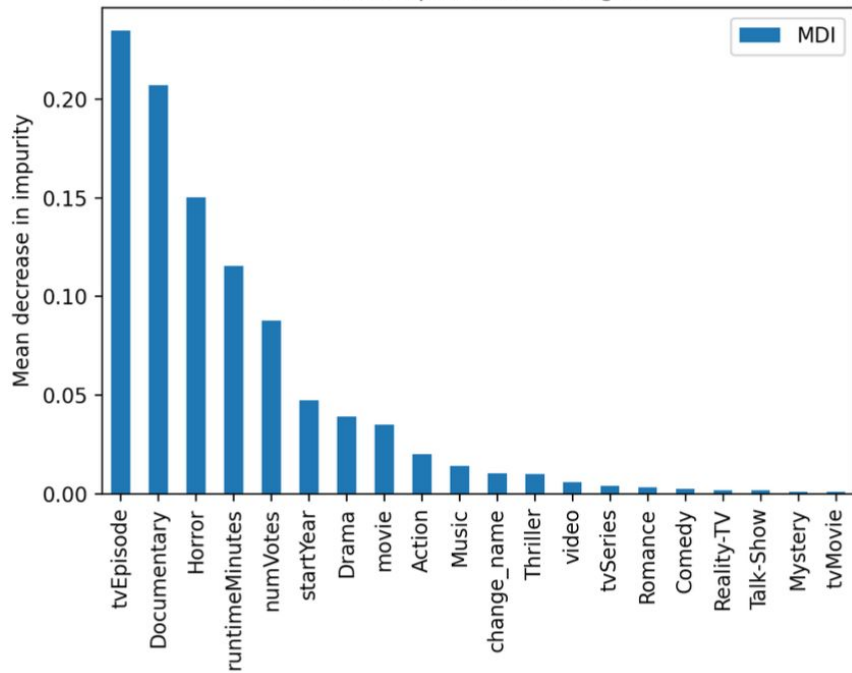
模型 / 評分	基本資料集		加入cast/crew (one-hot encoding)		加入cast/crew (word2vec)		加入cast/crew與title (w2v+bert)	
	adjusted R ²	RMSE	adjusted R ²	RMSE	adjusted R ²	RMSE	adjusted R ²	RMSE
Linear regression	0.298	1.159	0.351	1.115	0.364	1.102	0.374	1.099
RigdeCV	0.299	1.159	0.351	1.114	0.364	1.102	0.373	1.096
LassoCV	0.299	1.160	0.328	1.128	0.358	1.106	0.363	1.100
ElasticNetCV	0.296	1.161	0.322	1.133	0.355	1.108	0.360	1.102
Random Forest	0.391	1.091	0.381	1.091	0.409	1.081	0.467	1.082
XGBoost	0.487	1.030	0.466	1.026	0.565	0.977	0.606	0.979
NN	-	1.088	-	1.075	-	1.008	-	1.016



Feature Selection

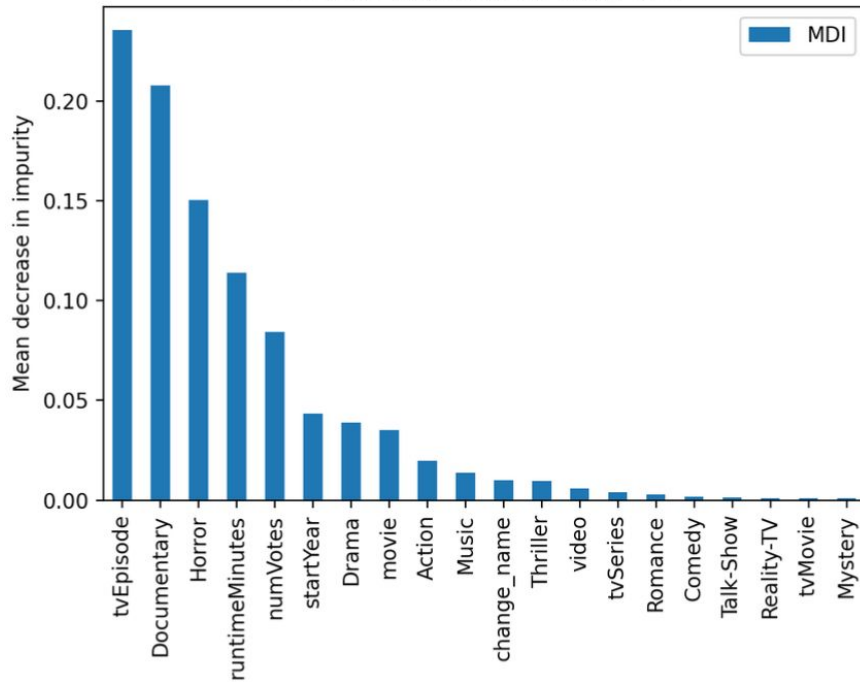
無cast/crew

Feature importances using MDI



有cast/crew

Feature importances using MDI



Word2Vector – 演員間之關聯性

- 透過計算 cosine similarity 可以找出有共同出演的演員

與李奧納多相似的演員

Leonardo DiCaprio
nm0000138



Russell Crowe
nm0000128



Vincent Cassel
nm0001993



Kate Winslet
nm0000701



John C. Reilly
nm0000604



Tom Cruise
nm0000129



Carey Mulligan
nm1659547



Aaron Eckhart
nm0001173



Emily Blunt
nm1289434



Rebecca Ferguso
nm0272581

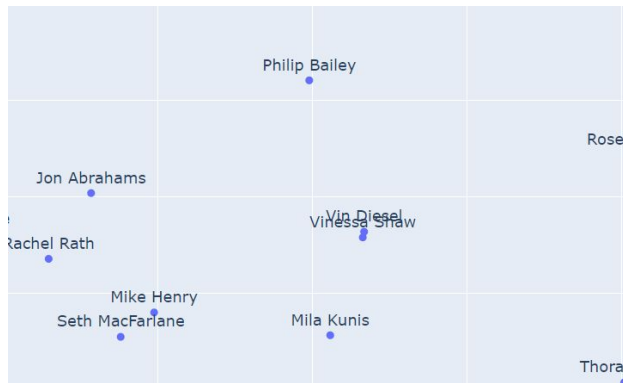


Word2Vector – 演員間之關聯性

Vin Diesel
nm0004874



John David Washington
nm0913475



Donnie Yen
nm0947447



Barry Pepper
nm0001608



Vin Diesel

- 與 Vin Diesel 相關的主要都是動作片的演員，其中包含了出演『天能』的 John David Washington；我們熟悉的甄子丹，他與馮迪索共同主演『限制級戰警』；還有出演『搶救雷恩大兵』的 Barry Pepper

Word2Vector – 演員間之關聯性

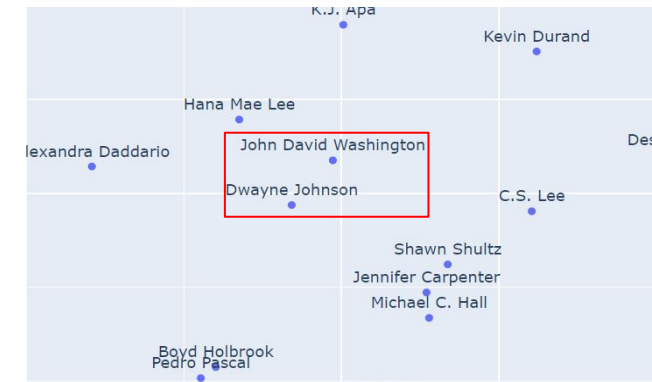
Dwayne Johnson

nm0425005



John David Washington

nm0913475



Paul Levesque

nm0505391



Mick Foley

nm0284201



Glenn Jacobs

nm0414417



Dwayne Johnson

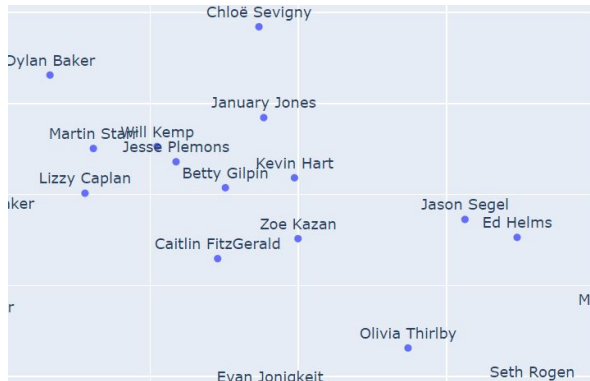
- 與 Dwayne Johnson 相關的主要都是摔角選手(肌肉猛男), 包含共同參與世界摔角冠軍的 Paul Levesque, 曾在 WWE 擊敗 Dwayne Johnson 拿下冠軍的 Mick Foley, 還有 WWE 選手 Glenn Jacobs
- John David Washington 則是和 Dwayne Johnson 曾共同出演『好球天團』

Word2Vector – 演員間之關聯性

Kevin Hart
nm0366389



Ed Helms
nm1159180



Mike Epps
nm0258402



Jason Segel
nm0781981



J. Cole
nm3359577



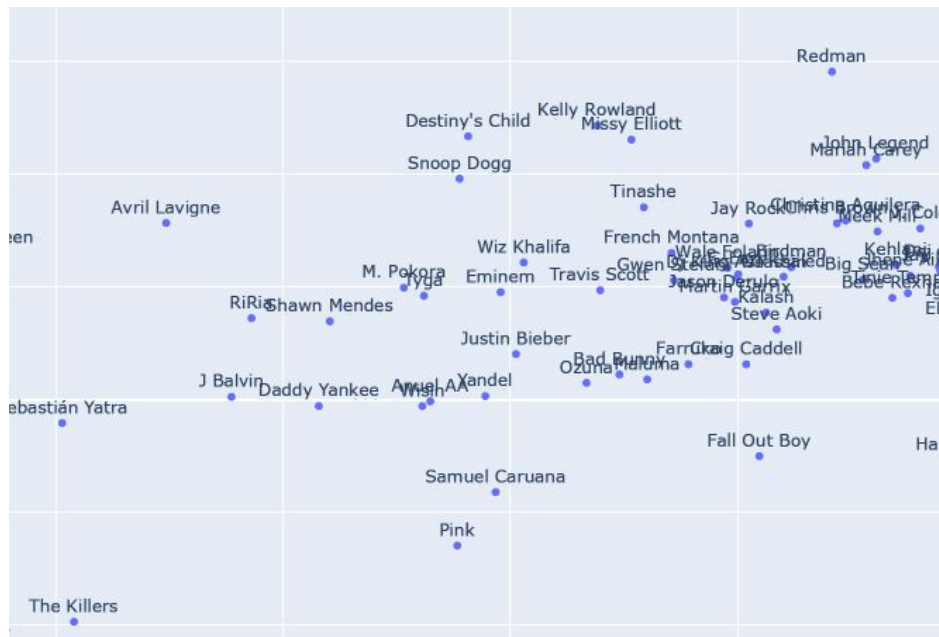
Kevin Hart

- 與 Kevin Hart 相關的多為喜劇演員，包含 Ed Helms、Mike Epps 和 Jason Segel，也和他們分別演出了『凱文哈特：現在要怎樣？』、『麥克伊皮斯：一支獨秀』與『五年之約』
- Kevin Hart 也曾經出演 J. Cole 的 MV 演出

Word2Vector – 其他群體

- 歌手群

- Avril Lavigne
- Wiz Khalifa
- Justin Bieber
- ...



K-means 演員分群

- 利用 K-means 為演員分群，並拿結果與 Word2Vector 的 similarity 比較，發現十個人中會有四個相同，表示分群結果與 W2V 效果差異不大

From W2V

Andrés García
nm0305955



Ofelia Medina
nm0575754



Griselda Nogueras
nm0634092



Laura Fabian
nm0264573



Anaís de Melo
nm0210265



From K-means

Andrés García
nm0305955



Ofelia Medina
nm0575754



Pablo Azar
nm1021126



Vanessa Villela
nm0889234



Jimena Araya
nm10358571



Laura Fabian
nm0264573



Griselda Nogueras
nm0634092



K-means 演員分群

- 因為演員數量多，利用 K-means 分成 500 群之後，隨機挑選兩個群體觀察
- 分群的結果在國籍上具有一致性

多為墨西哥演員

Silvia Navarro
nm0623021



Sergio Basañez
nm0059639



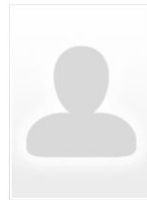
Víctor González
nm0328514



Anette Michel
nm0584795



Margarita Galia
nm0334482

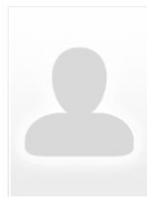


Sergio Kleiner
nm0459073

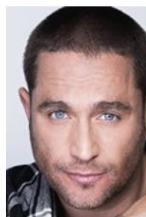


多為阿根廷演員

Luciano Castro
nm0145624



Michel Brown
nm0112868



Carola Reyna
nm0721487



Trinidad Alcorta
nm0017294



Pablo Novak
nm0636904



Cecilia Dopazo
nm0233292



結論

- 預測 IMDb 電影評分
 - RMSE預測評分大約誤差正負 1
- Actor Embedding Learning
 - 節省大量運算資源和時間, 且在分數上有更佳的表现
- 比較不同演員的性質異同
 - 成功使用 Actor Embedding 來計算相似度, 能夠將演員特質量化並且放在一群

Thanks For Listening

附錄：預測結果（微調後）

- 用GridSerach 之後得出的結果

Fitting 4 folds for each of 60 candidates, totalling 240 fits

Best parameters found: {'colsample_bytree': 0.7, 'eta': 0.1, 'max_depth': 10}

Lowest RMSE found: 0.96295253171448

- 最後得到：
 - adjusted r_square : 0.737086
 - RMSE : 0.949862

XGBoost	XGBoost_fine
0.567154	0.738123
0.565440	0.737086
0.976618	0.949862
163.998187	254.020915

附錄：預測模型

- 任務：預測電影評分
- 預期：電影的卡司與幕後員工會影響電影之評分
- 方法：使用 5 種 sklearn 中的回歸預測模型
 - 資料集
 - 一：不包含卡司與幕後員工
 - 二：卡司與幕後員工之 one hot encoding
 - 三：卡司與幕後員工之 Word2vec embedding
 - 四：卡司與幕後員工之 Word2vec embedding + 電影名稱Bert
 - 回歸模型
 - Linear Regression, RidgeCV, LassoCV, ElasticNet, XGBoost

附錄：預測結果

評分 / 模型	資料集一	資料集二	資料集三	資料集四
最佳模型	XGBoost	XGBoost	XGBoost	XGBoost
adjusted R ²	0.486669	0.465692	0.565440	0.605935
RMSE	1.030348	1.026167	0.976618	0.978936

附錄：預測結果(調整)

- 用GridSerach 之後得出的結果

Fitting 4 folds for each of 60 candidates, totalling 240 fits

Best parameters found: {'colsample_bytree': 0.7, 'eta': 0.1, 'max_depth': 10}

Lowest RMSE found: 0.96295253171448

- 最後得：
 - adjusted r_square : 0.737086
 - RMSE : 0.949862

XGBoost	XGBoost_fine
0.567154	0.738123
0.565440	0.737086
0.976618	0.949862
163.998187	254.020915

附錄: Feature Selection

- 使用 sklearn 之中 feature_selection 套件裡的 SelectKBest
 - 選取前 20 個重要的欄位來做訓練
 - 資料集一
 - ['short', 'History', 'Crime', 'change_name', 'Talk-Show', 'Horror', 'Documentary', 'Romance', 'Biography', 'Game-Show', 'Animation', 'videoGame', 'tvEpisode', 'tvMovie', 'Comedy', 'averageRating', 'tvShort', 'startYear', 'War']
 - 資料集二
 - ['tvEpisode', 'Horror', 'Documentary', 'movie', 'Thriller', 'Music', 'Drama', 'Sci-Fi', 'Comedy', 'History', 'Biography', 'Action', 'tvMiniSeries', 'nm0490375', 'tvSeries', 'nm0912766', 'Crime', 'numVotes', 'tvSpecial', 'nm0937725']

附錄: Word2Vector – 演員間之關聯性

Emma Stone

nm1297015



Ryan Gosling

nm0331516



Jonah Hill

nm1706767



Stanley Tucci

nm0001804



Emma Stone

- Ryan Gosling 與 Emma Stone 共同出演三部電影, 包含了廣受好評的 LA LA LAND。根據 Emma Stone 採訪中她也表明 she couldn't "even imagine what my life would be without Ryan." 可見兩人之間的密切關聯
- Jonah Hill 則與 Emma Stone 共同出演改編自挪威的同名電視劇 *Maniac*, 許多影評讚賞該劇的視覺效果以及 Stone 與 Hill 的演技。另外, Stone 的首秀也是與 Hill 合作出演的
- Stone 在演藝上的突破受惠於 2010 年的喜劇片『破處女王』, 該劇的合作演員包含了 Stanley Tucci

附錄: Word2Vector – 演員間之關聯性

Javier Bardem

nm0000849



Jasmine Trinca

nm0872910



Penélope Cruz

nm0004851



Lola Dueñas

nm0240318

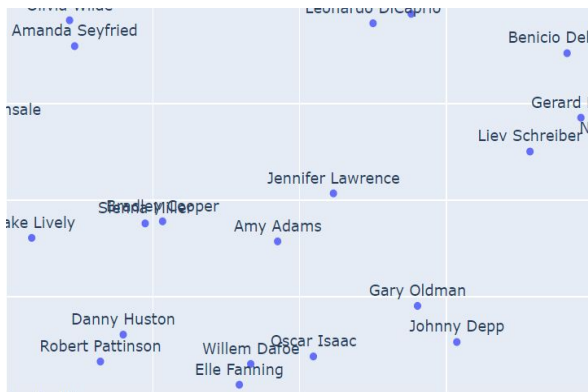


Javier Bardem

- Javier Bardem 與 Jasmine Trinca 及 Lola Dueñas 分別演出『The Gunman』和『The Sea Inside』
- Penélope Cruz 則是 Javier Bardem 的妻子。綜合 Javier Bardem 以及 Nicole Kidman 的分類可以發現, 具有伴侶關係是除了共同演出之外的歸類因子

附錄: Word2Vector – 演員間之關聯性

Jennifer Lawrence
nm2225369



Bradley Cooper
nm0177896



Meryl Streep
nm0000658



Constance Wu
nm2090422

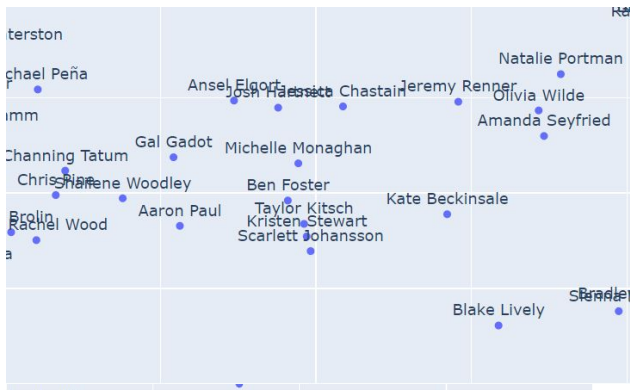


Jennifer Lawrence

- Jennifer Lawrence 與 Bradley Cooper 為螢幕情侶，曾共同出演四部戲
- 另外，其他相關的演員跟 Jennifer Lawrence 一樣多為大咖女星，包含正在一起拍戲的 Meryl Streep，還有曾共同出演『舞孃騙很大』的 Constance Wu

附錄: Word2Vector – 演員間之關聯性

Scarlett Johansson
nm0424060



Justin Timberlake
nm0005493



Hugh Jackman
nm0413168



Nathan Fillion
nm0277213



Matt Damon
nm0000354



Scarlett Johansson

- 黑寡婦 Scarlett Johansson 很漂亮且緋聞不斷，可以發現與他相關的演員多為男性，包含非常好的朋友休傑克曼，兩人曾共同演出『頂尖對決』。史嘉蕾過去和萊恩雷諾斯結婚時，休傑克曼則開玩笑的向萊恩雷諾斯開戰
- 另外與 Justin Timberlake 曾共同出演 MV，與 Nathan Fillion 曾共同出演『美人心機』，與 Matt Damon 曾共同演出『我們買了動物園』