

Blockbuster

-Application of Random Forests and Regression
Model for Opening Weekend Gross Forecasting

爬蟲、 ETL
、資料分析

爬蟲、資料庫、 ETL
、 資料視覺化

莊季陶

爬蟲、 ETL、
Hadoop叢集、
資料視覺化

周昱宏

指導老師
郭惠民

黃鉉鈞

爬蟲、 ETL
、資料分析
、資料視覺化

江浩平

ETL 、
資料視覺化

邱柏龍

TITLE

DIRECTOR

小組成員

分析目的



利用過去電影之票房與其相關資訊，預測未來上映電影之首周票房，並期望將此預測資訊提供給電影院業者，使業者在安排播映場次上可達到最佳化。

資料來源



- <http://www.imdb.com/>
- 蒐集2000年至2014年間，其成本超過30萬的電影之全美票房與其相關資訊。



Overview

- 系統建置
- 資料蒐集
- 資料庫建構
- 資料ETL
- 資料分析
- 視覺化呈現
- 結論與展望



系統建置

系統流程

Data Collection



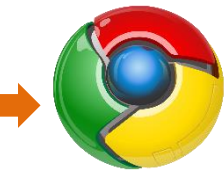
python™



Microsoft®
SQL Server™

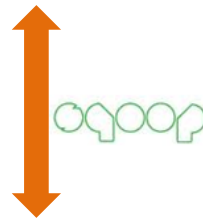


Shiny



HIGHCHARTS

Data ETL and Analysis



MapReduce

HDFS

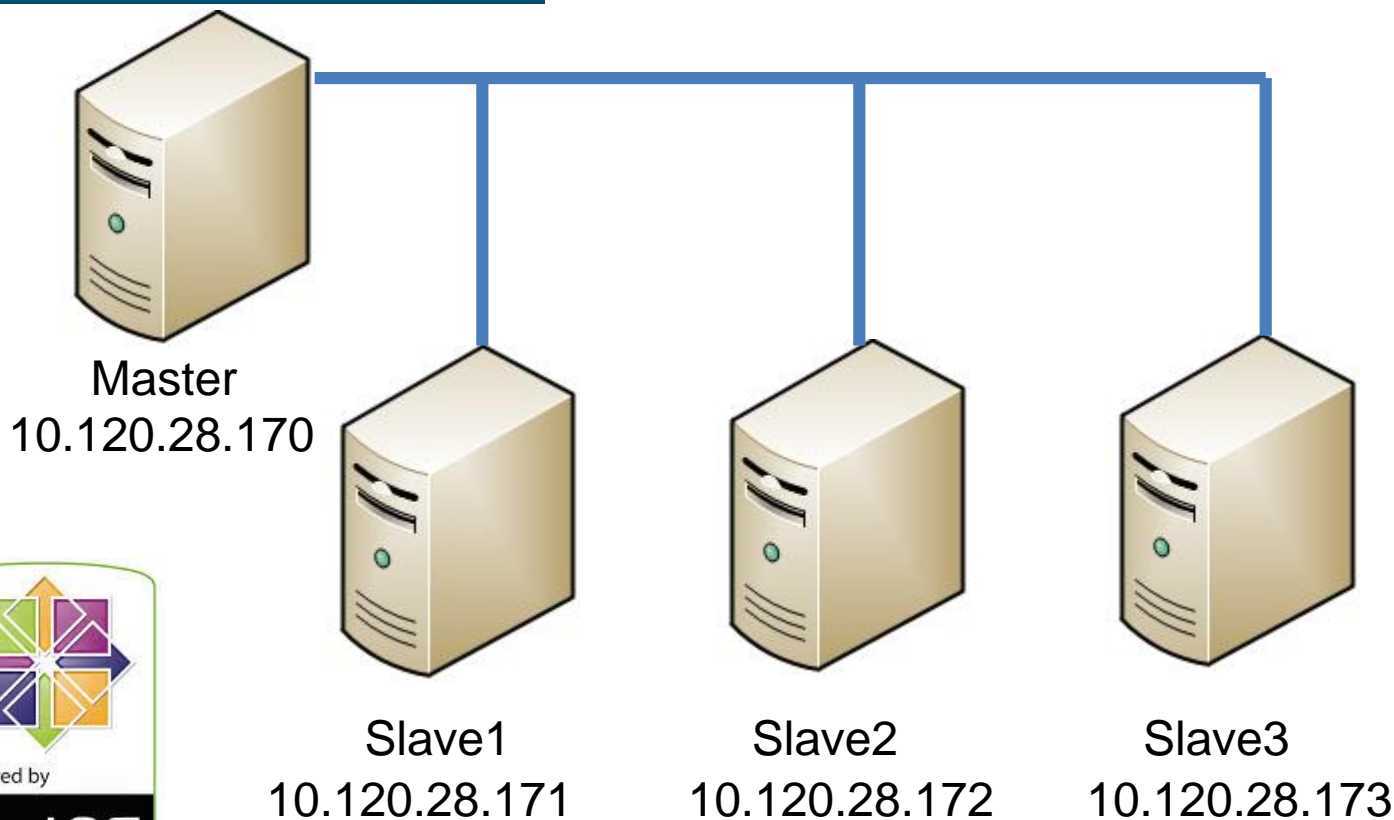
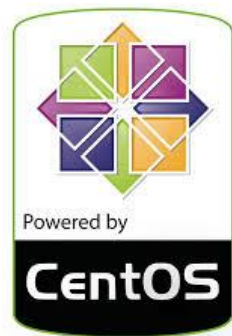


REVOLUTION
ANALYTICS

Hadoop叢集架構

cloudera manager

- 1.主機名稱
- 2.IP位置
- 3.hosts



Cluster install

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This will also enable health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using [patterns](#) .

master,slave1,slave2

SSH Port:

22

Search

Cluster Install

Cluster Installation

Provide SSH login credentials.

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As: ☒ root
☐ Another user

You may connect via password or public-key authentication for the user selected above.

Authentication Method: ☒ All hosts accept same password
☐ All hosts accept same private key

Enter Password:

Confirm Password:

SSH Port:

Number of Simultaneous Installations: (Running a large number of installations at once can consume large amounts of network bandwidth and other system resources)

◀ Back

1 2 3 4 5 6 7 8

▶ Continue

Cloudera Manager

cloudera manager Home Clusters Hosts Diagnostics Audits Charts Administration

Home Status All Health Issues All Configuration Issues 4 All Recent Commands

Cluster 1 (CDH 5.3.0, Parcels)

Hosts	4
HBase	
HDFS	
Hive	
Hue	
Impala	
Key-Value St...	
Oozie	
Solr	
Spark	
Sqoop 2	
YARN (MR2 I...	
ZooKeeper	

Cloudera Management Service

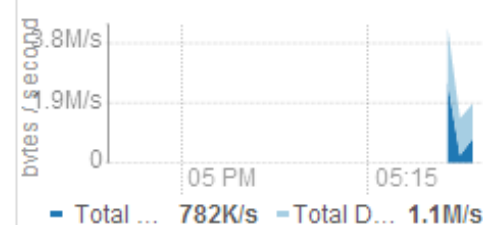
Cloudera Ma...	
----------------	--

Charts

Cluster CPU



Cluster Disk IO



HDFS IO



Spark Standalone

啟動 master

`#./sbin/start-master.sh`

master Web UI: <http://master.iii.com:8080/>



Spark Master at `spark://master.iii.com:7077`

URL: `spark://master.iii.com:7077`

Workers: 0

Cores: 0 Total, 0 Used

Memory: 0.0 B Total, 0.0 B Used

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Spark Standalone

worker加入master

```
./bin/spark-class
```

```
org.apache.spark.deploy.worker.Worker \
```

```
spark://master.iii.com:7077
```

Spark Standalone



Spark Master at spark://master.iii.com:7077

URL: spark://master.iii.com:7077

Workers: 4

Cores: 8 Total, 0 Used

Memory: 30.7 GB Total, 0.0 B Used

Applications: 0 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers

Id	Address	State	Cores	Memory
worker-20150116102357-master.iii.com-55173	master.iii.com:55173	ALIVE	2 (0 Used)	10.6 GB (0.0 B Used)
worker-20150116102441-slave1.iii.com-51049	slave1.iii.com:51049	ALIVE	2 (0 Used)	6.7 GB (0.0 B Used)
worker-20150116102510-slave2.iii.com-36160	slave2.iii.com:36160	ALIVE	2 (0 Used)	6.7 GB (0.0 B Used)
worker-20150116102530-slave3.iii.com-48234	slave3.iii.com:48234	ALIVE	2 (0 Used)	6.7 GB (0.0 B Used)

Running Applications

ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
----	------	-------	-----------------	----------------	------	-------	----------

Completed Applications



資 料
蒐 集



抓取頁面

- ✓ 電影列表
- ✓ 電影主頁
- ✓ 電影演員與工作人員頁面
- ✓ 電影票房頁面
- ✓ 電影拍攝地點頁面
- ✓ 電影圖片頁面
- ✓ 匯率網站



電影列表

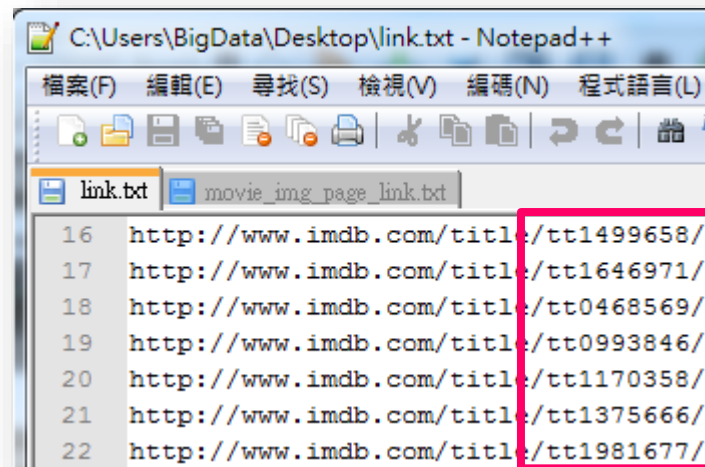
抓取全部電影連結

1-50 of 5,014 titles.

Next »

Sort by: **MOVIEmeter** | A-Z | User Rating | Num Votes | US Box Office | Runtime | Year | US Release Date

-  **Foxcatcher** (2014) Add to Watchlist
★★★★★ 7.4/10
The greatest Olympic Wrestling Champion brother team joins Team Foxcatcher led by multimillionaire sponsor John E. du Pont as they train for the 1988 games in Seoul - a union that leads to unlikely circumstances.
Dir: Bennett Miller With: Steve Carell, Channing Tatum, Mark Ruffalo
Biography | Drama | Sport | Thriller 129 mins. R
-  **Whiplash** (2014) Add to Watchlist
★★★★★ 8.7/10
A promising young drummer enrolls at a cutthroat music conservatory where his dreams of greatness are mentored by an instructor who will stop at nothing to realize a student's potential.
Dir: Damien Chazelle With: Miles Teller, J.K. Simmons, Melissa Benoist
Drama | Music 107 mins. R
-  **Boyhood** (2014) Add to Watchlist
★★★★★ 8.3/10
The life of a young man, Mason, from age 5 to age 18.
Dir: Richard Linklater With: Ellar Coltrane, Patricia Arquette, Ethan Hawke
Drama 165 mins. R
-  **Divergent** (2014) Add to Watchlist
★★★★★ 6.8/10
In a world divided by factions based on virtues, Tris learns she's Divergent and won't fit in. When she discovers a plot to destroy Divergents, Tris and the mysterious Four must find out what makes Divergents dangerous before it's too late.
Dir: Neil Burger With: Shailene Woodley, Theo James, Kate Winslet
Action | Adventure | Sci-Fi | Thriller 139 mins. PG-13





電影各類屬性頁面

電影主頁

movie_list table(電影清單)

genres table(電影類型表)

電影演員與工作人員頁面

combine table(導演與作者表)

cast table(演員表)

電影票房頁面

budget table(預算表)

boxoffice table(票房表)


電影拍攝地點頁面

locations table(拍攝地點表)



匯率資料

- 根據年度及歷史匯率修正個電影預算成美元USD
- 歷史匯率資料來源: <http://www.oanda.com/>



Currency I Have:		Currencies I Want:				
US Dollar USD		AUD	CAD	CNY	CZK	DKK
RANGE:	Custom	Jan 1, 2004	Dec 31, 2014	INTERBANK: +/- 0%		
PRICE:	Bid	VALUES:	Rate	FREQUENCY:	Annual	
	USD / AUD	USD / CAD	USD / CNY	USD / CZK	USD / DKK	
2014	1.1094	1.1041	6.1432	20.7388	5.6173	
2013	1.0362	1.0298	6.1905	19.5326	5.6168	
2012	0.9658	0.9996	6.3034	19.5461	5.7913	
2011	0.9687	0.9888	6.4544	17.6568	5.3552	
2010	1.0900	1.0302	6.7605	19.0670	5.6206	
2009	1.2805	1.1411	6.8212	18.9961	5.3541	
2008	1.1961	1.0660	6.9404	17.0334	5.0934	
2007	1.1947	1.0738	7.5972	20.2603	5.4429	
2006	1.3277	1.1340	7.9646	22.5497	5.9426	
2005	1.3115	1.2111	8.1838	23.9170	5.9932	
2004	1.3590	1.3008	8.2664	25.6567	5.9877	

historical_rate table(匯率表)

日期	ARS	AUD	BEF	BRL
2015/1/11	8.5854	1.2185	34.567	2.6264
2015/1/10	8.5851	1.2271	34.567	2.6511
2015/1/9	8.5745	1.2338	34.567	2.6771
2015/1/8	8.5534	1.2394	34.567	2.6912
2015/1/7	8.5464	1.2314	34.567	2.7008
2015/1/6	8.5666	1.2379	34.567	2.6976
2015/1/5	8.5527	1.2359	34.567	2.702
2015/1/4	8.5518	1.2359	34.567	2.6912
2015/1/3	8.5243	1.2294	34.567	2.6724
2015/1/2	8.4474	1.225	34.567	2.6527





資料表

movie_list table(電影清單)

電影代碼	電影名稱
tt0479997	Season of the Witch
tt2404461	Le passé
tt1091191	Lone Survivor
tt0359950	The Secret Life of Walter Mitty
tt0993846	The Wolf of Wall Street
tt1335975	47 Ronin

combine table(導演與作者表)

電影代碼	職稱	導演與作者順序	人物代碼	導演與作者
tt0993846	director	1	nm0000217	Martin Scorsese
tt0993846	writer	1	nm1010540	Terence Winter
tt0993846	writer	2	nm0067789	Jordan Belfort

genres table(電影類型表)

電影代碼	電影類型
tt0993846	Biography
tt0993846	Comedy
tt0993846	Crime
tt0993846	Drama

cast table(演員表)

電影代碼	演員順序	人物代碼	演員
tt0993846	1	nm0000138	Leonardo DiCaprio
tt0993846	2	nm1706767	Jonah Hill
tt0993846	3	nm3053338	Margot Robbie
tt0993846	4	nm0000190	Matthew McConaughey
tt0993846	5	nm0151419	Kyle Chandler



資料表

budget table(預算表)

電影代碼	預算貨幣型態	預算
tt0993846	USD	100000000

locations table(拍攝地點表)

電影代碼	國家
tt0993846	USA

boxoffice table(票房表)

電影代碼	週數	票房貨幣型態	票房	日期	上映廳數
tt0993846	1	USD	9000000	2014/1/12	2521
tt0993846	2	USD	7500000	2014/1/19	1930
tt0993846	3	USD	5478368	2014/1/26	1804
tt0993846	4	USD	3400780	2014/2/2	1607
tt0993846	5	USD	2570000	2014/2/9	1167
tt0993846	6	USD	2288672	2014/2/16	751
tt0993846	7	USD	1302871	2014/2/23	627
tt0993846	8	USD	688931	2014/3/9	359

historical_rate table(匯率表)

日期	ARS	AUD	BEF	BRL
2015/1/11	8.5854	1.2185	34.567	2.6264
2015/1/10	8.5851	1.2271	34.567	2.6511
2015/1/9	8.5745	1.2338	34.567	2.6771
2015/1/8	8.5534	1.2394	34.567	2.6912
2015/1/7	8.5464	1.2314	34.567	2.7008



預算幣值修正

budget table(預算表)

電影代碼	預算貨幣型態	預算
tt1740707	NOK	19900000

boxoffice table(票房表)

電影代碼	週數	票房貨幣型態	票房	日期	上映廳數
tt1740707	1	USD	5585	2011/6/12	1
tt1740707	2	USD	20125	2011/6/26	9
tt1740707	3	USD	19222	2011/7/10	19
tt1740707	4	USD	10369	2011/7/17	20
tt1740707	5	USD	10305	2011/7/24	21
tt1740707	6	USD	7588	2011/7/31	12
tt1740707	7	USD	7409	2011/8/7	16
tt1740707	8	USD	8559	2011/8/14	14
tt1740707	9	USD	2406	2011/8/21	5



電影代碼	日期	預算貨幣型態	預算
tt1588337	2011/2/27	EUR	4000000
tt1600524	2011/2/27	CAD	600000
tt1020773	2011/3/13	EUR	7000000
tt1339161	2011/2/13	EUR	1000000
tt1740707	2011/6/12	NOK	19900000
tt1668200	2011/7/24	EUR	10000000
tt1270262	2011/7/31	EUR	15000000
tt1236371	2011/8/7	EUR	2500000



預算幣值修正

電影代碼	日期	預算貨幣型態	預算
tt1588337	2011/2/27	EUR	4000000
tt1600524	2011/2/27	CAD	600000
tt1020773	2011/3/13	EUR	7000000
tt1339161	2011/2/13	EUR	1000000
tt1740707	2011/6/12	NOK	19900000



電影代碼	日期	預算貨幣型態	預算
tt1588337	2011/2/27	USD	5503577
tt1600524	2011/2/27	USD	614187
tt1020773	2011/3/13	USD	9735744
tt1339161	2011/2/13	USD	1355564
tt1740707	2011/6/12	USD	3634636



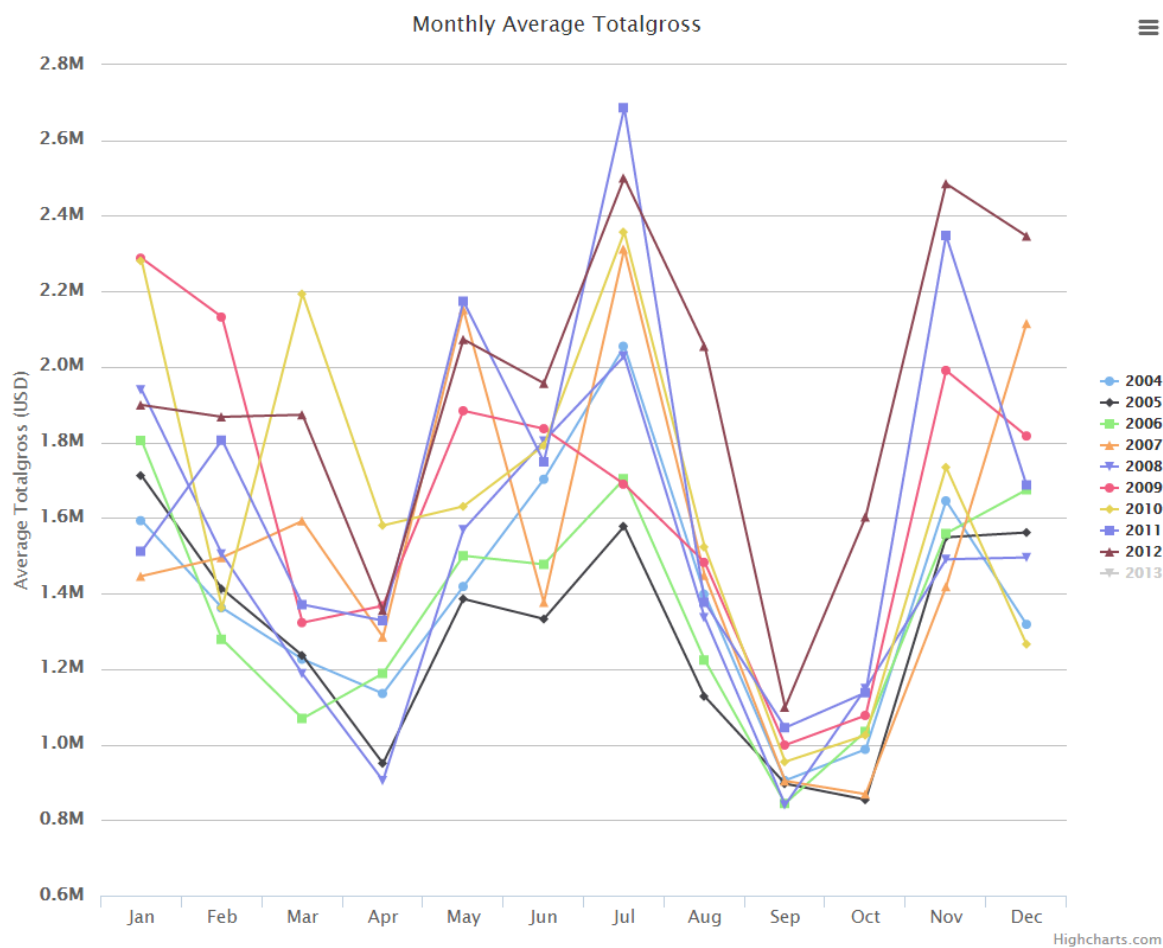
資料表

movie_list table(電影清單)

電影代碼	電影名稱	首位導演	電影類型	預算(萬)		上映日期	票房(萬)	上映廳數
tt0993846	The Wolf of Wall Street	Martin Scorsese	Comedy	1000		2014/1/12	90	2521

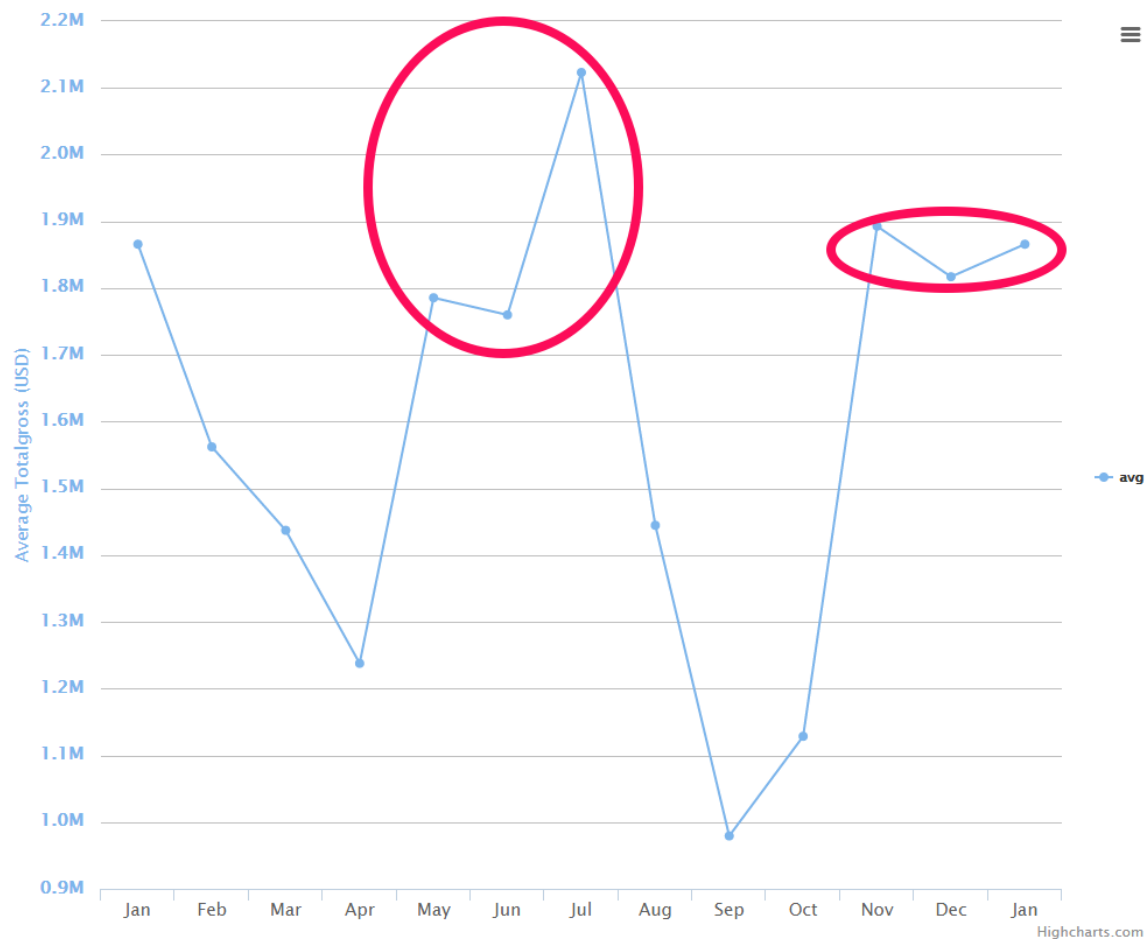


上映月份與票房的相關性



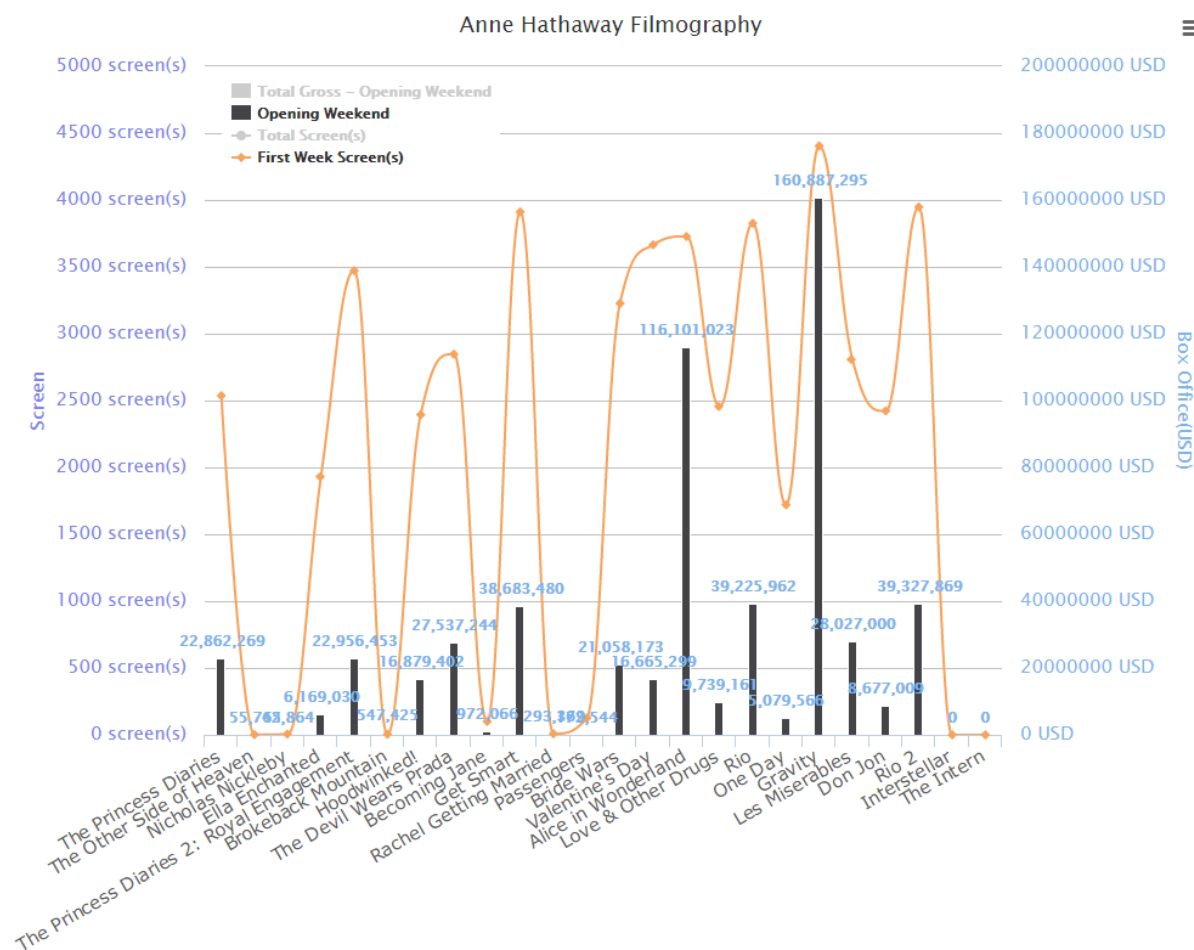


上映月份與票房的相關性





上映廳數與票房的相關性





考慮屬性



- 演員
- 導演
- 電影成本
- 上映月份
- 電影類型
- 上映院廳數



判斷演員重要程度

只要是**正妹**
就很重要!?



依照google上的搜尋數來
判斷此演員是否重要



以2013年全美片酬第500高
的明星搜尋數為基準，即
搜尋數大於13萬5千為有名
的演員

抓Google Search Item



emma watson



網頁

圖片

影片

新聞

地圖

更多 ▾

搜尋工具

約有 41,200,000 項結果 (搜尋時間：0.24 秒)

相關搜尋： [emma watson instagram](#) [emma watson高登](#) [emma watson 洗澡](#)

提示： 只顯示繁體中文搜尋結果。您可以在使用偏好中指定搜尋語言

[Emma Watson - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Emma_Watson ▾ 翻譯這個網頁

Emma Charlotte Duerre **Watson** (born 15 April 1990) is an English actress and model.

She rose to prominence portraying Hermione Granger in the Harry Potter ...

[Regression - Brown University](#) - [HeForShe](#) - [My Week with Marilyn](#)



艾瑪·華森

503錯誤

如要繼續，請輸入以下字元：

2371521

https://www.google.com.tw/

503

Yuki Amami

+star%22

提交

為何顯示此頁

我們的系統偵測到您的電腦網路送出的流量有異常情況。這頁是為了確認要求確實出自您本人，不是由自動程式發出。[為什麼會發生這種情況？](#)

IP 位址：59.115.85.231

時間：2014-12-08T00:17:52Z

網址：https://www.google.com.tw/search?

解決方法



Home

About Tor

Do

Anonymity Online

Protect your privacy. Defend yourself against network surveillance and traffic analysis.



Download Tor 

- ➔ Tor prevents people from learning your location or browsing habits.
- ➔ Tor is for web browsers, instant messaging clients, and more.
- ➔ Tor is free and open source for Windows, Mac, Linux/Unix, and Android

建立proxy連線

建立proxy連線

```
session = requests.session()|
session.proxies = {
    'http': 'socks5://127.0.0.1:9150',
    'https': 'socks5://127.0.0.1:9150'
}
```

重新建立連線

```
def newI():
    controller = Controller.from_port(port = 9151)
    try:
        controller.authenticate()
        controller.signal(Signal.NEWNYM)
    finally:
        controller.close()
```

困境解決

```
https://www.google.com.tw/search?q=%22Theodule+Carre-Cassaig
```

```
200
```

```
25
```

```
Theodule Carre-Cassaigne==>ok
```

```
https://www.google.com.tw/search?q=%22Jerome+Kircher%22+%22m
```

```
200
```

```
1540
```

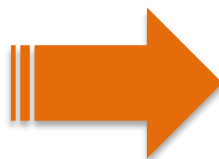
```
Jerome Kircher==>ok
```

Actor_item.TXT

19274	Emma Watson, <u>425000</u> >135,000=有名
19275	Alyson Hannigan, 374000
19276	John Cullum, 14500
19277	Carly Pope, 28600
19278	Annika Pergament, 392

屬性數值化

電影	Iron Man 3
演員1	Robert Downey Jr.
演員2	Gwyneth Paltrow
演員3	Don Cheadle
導演	Shane Black
上映月份	5
類型	Action
成本	\$200, 000, 000
上映廳院數	4253
明星個數	6



電影	Iron Man 3
演員1	1
演員2	0.567
演員3	0.413
導演	0.912
上映月份	0.691
類型	0.702
成本	0.940
上映廳院數	0.891
明星個數	0.862



計算各個屬性權重

月份	歷史平均票房
1	1985513
2	1658829
3	1485571
4	1263961
5	1870371
6	2005406
7	2247056
8	1678666
9	1027425
10	1162892
11	1890273
12	1883724

$$W(\omega) = \frac{\omega - \min(\omega)}{\max(\omega) - \min(\omega)}$$



月份	歷史平均票房
1	0.7856
2	0.5177
3	0.3756
4	0.1939
5	0.6911
6	0.8019
7	1
8	0.534
9	0
10	0.1111
11	0.7075
12	0.7021



計算各個屬性權重

類型	歷史平均票房
Adventure	55452400
Sci-Fi	44419167
Fantasy	49290056
Animation	56585510
Action	39719300
Family	47945038
Thriller	22025164
Mystery	21581004
Horror	17000625
Western	19105297
Comedy	19583716
Crime	16930205
Sport	17840539

$$W(\omega) = \frac{\omega - \min(\omega)}{\max(\omega) - \min(\omega)}$$



類型	歷史平均票房
Adventure	0.98
Sci-Fi	0.7849
Fantasy	0.871
Animation	1
Action	0.7018
Family	0.8473
Thriller	0.3891
Mystery	0.3812
Horror	0.3002
Western	0.3374
Comedy	0.3459
Crime	0.299
Sport	0.3151



補遺失值

針對上映廳院數屬性：

- null 表示此部電影的上映廳數為遺失值
- 在此用全部電影上映廳院數的平均數來補遺失值

Movie_id	screens
tt0120667	3602
tt0121164	5
tt0121766	3661
tt0167190	3028
tt0204313	2803
tt0206634	null
tt0211933	2002



數值化方法

Hive

- join過去票房資訊表格

Hadoop Streaming

- 使用python撰寫map與reduce腳本



First Step



Map

cast	year	gross
Jason Statham	2013	15.74
Jason Statham	2012	17.68
Izabela Vidovic	2013	15.74
James Franco	2013	15.74
Ansel Elgort	2012	17.68
Nat Wolff	2012	17.68



Reduce

cast	2010	2011	2012	2013
Jason Statham	0.7049	0.7121	0.6499	0.6186
James Franco	0.7743	0.7743	0.6742	0.6742

Input data:

Movie_id	gross	cast	year
tt2312718	15.74	Jason Statham	2013
tt2312718	15.74	James Franco	2013
tt2312718	15.74	Izabela Vidovic	2013
tt2582846	17.68	Jason Statham	2012
tt2582846	17.68	Ansel Elgort	2012
tt2582846	17.68	Nat Wolff	2012



Second Step

Input data:

cast	2010	2011	2012	2013
Jason Statham	0.7049	0.7121	0.6499	0.6186

Movie_id	cast	year
tt2312718	Jason Statham	2013
tt2312718	James Franco	2013
tt2312718	Izabela Vidovic	2013

Map

Movie_id	cast	year
tt2312718	Jason, James, Izabela	2013

Reduce

Movie_id	cast1_weight	cast2_weight	cast3_weight
tt2312718	0.6499	0.265	0.1606



資 料
分 析



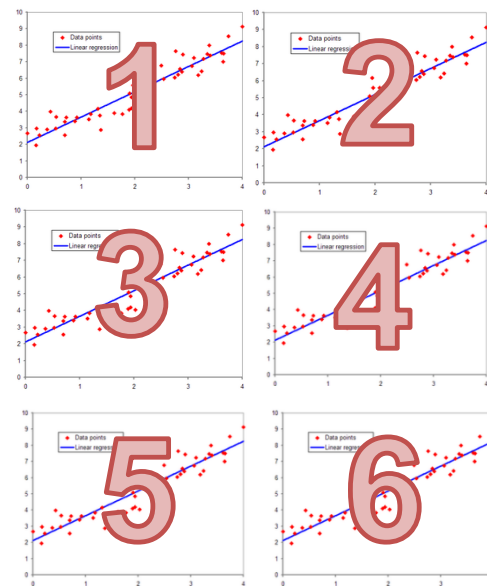
分析流程



分成6類



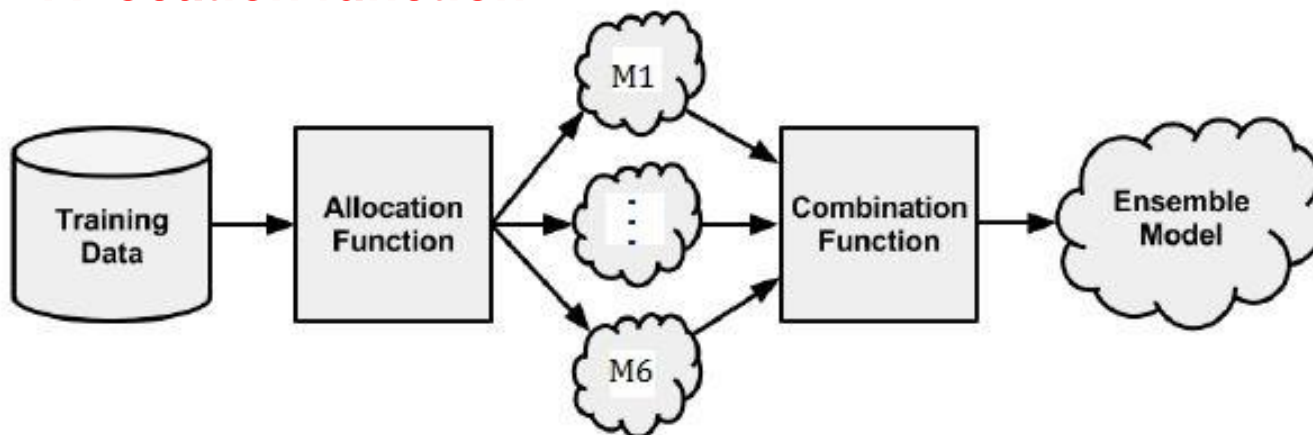
建立迴歸模型





Ensembles

Allocation function



原始訓練資料

切割訓練資料

建立各別模型

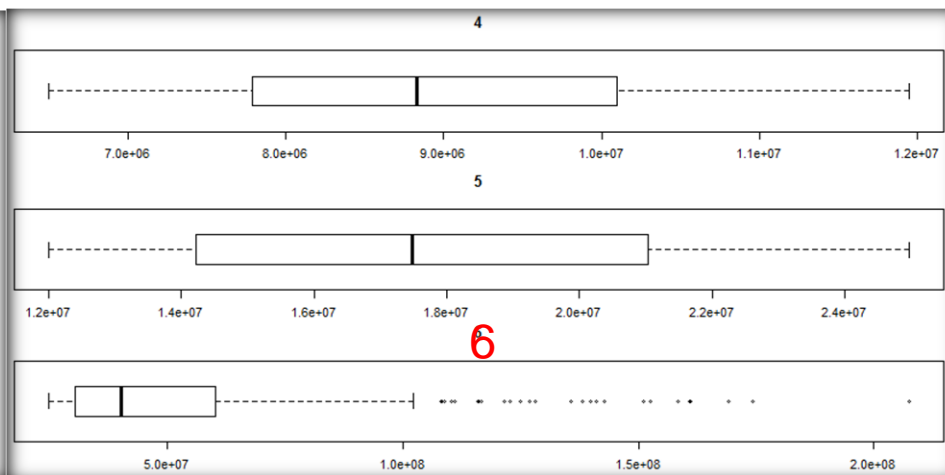
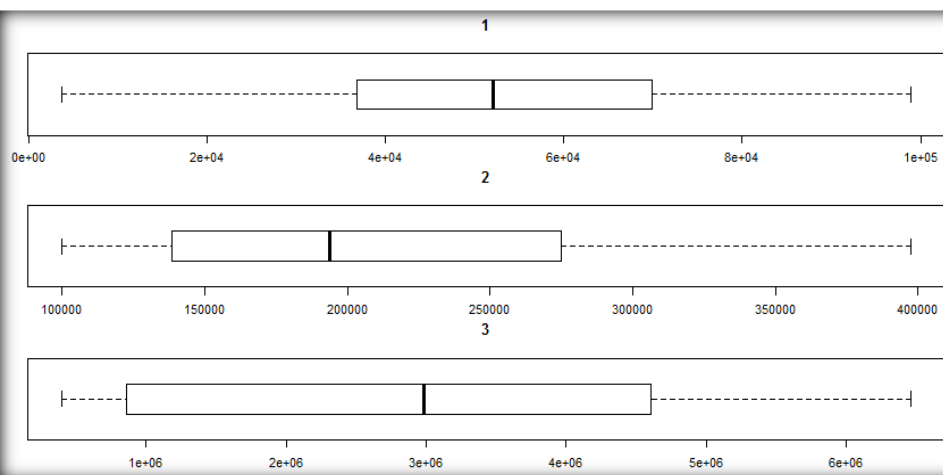
結合訓練資料

結合各別模型



票房等級區間

Level	1	2	3	4	5	6
Interval 單位:萬元	[0,1)	[1, 4)	[4,65)	[65,120)	[120,250)	[250,Inf)
#	181	209	395	274	400	349





Principal Component Analysis

```
> summary(queryResults.pr)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9
Standard deviation	0.4447372	0.2733328	0.2242776	0.20429856	0.15949076	0.13376954	0.11665807	0.09028861	0.07600633
Proportion of Variance	0.4542642	0.1715872	0.1155243	0.09585884	0.05842149	0.04109756	0.03125584	0.01872266	0.01326787
Cumulative Proportion	0.4542642	0.6258514	0.7413758	0.83723459	0.89565608	0.93675364	0.96800948	0.98673213	1.00000000

96.8%

- 將資料維度縮減
- 取得相互垂直的新軸
- 使用少數主成分取代原變數來解釋原始資料



分類演算法

--類神經網路演算法--

- 可以充分逼近任意複雜的非線性關係
- 採用並行分佈處理方法，使得快速進行大量運算成為可能

--隨機森林演算法--

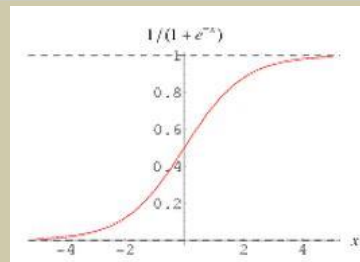
- 可以處理大量的輸入變數，並處理其重要性
- 遇到大量遺失的資料，仍可以維持準確度
- 學習過程是快速的



類神經網路演算法

模型設計

- 輸入層節點個數為7
 - 兩層隱藏節點個數分別為15與10
 - 輸出層節點個數為6
-
- 遞迴式網路結構
 - 隨機指定初始權重
-
- 使用sigmoid激發函數
 - 收斂參數為0.1





隨機森林演算法

模型設計

- 拔靴集合法 (bootstrap aggregation, bagging)

- 以簡單多數表決 (simple majority vote) 決定最後之結果

- 樹的數量: 1000

- 子集合包含變數數量為原始集合的80%



模型比較

Confusion Matrix

隨機森林

	1	2	3	4	5	6
1	28	5	0	0	0	0
2	17	40	11	0	0	0
3	4	9	77	8	0	1
4	0	0	29	52	35	9
5	0	0	4	20	86	38
6	0	0	0	0	11	62

Accuracy

55%

1-Away

97%

類神經網路

	1	2	3	4	5	6
1	22	10	5	0	0	0
2	8	22	13	0	0	1
3	0	7	50	7	5	0
4	0	0	11	13	17	2
5	0	1	4	24	46	12
6	0	1	3	3	19	51

Accuracy

54%

1-Away

92%



Linear Regression

- Formula: $Y = X\beta + \varepsilon$

- Response Variables: $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

- Explanatory Variables: $X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}$

- Coefficients: $\hat{\beta} = (X^T X)^{-1} X^T Y$



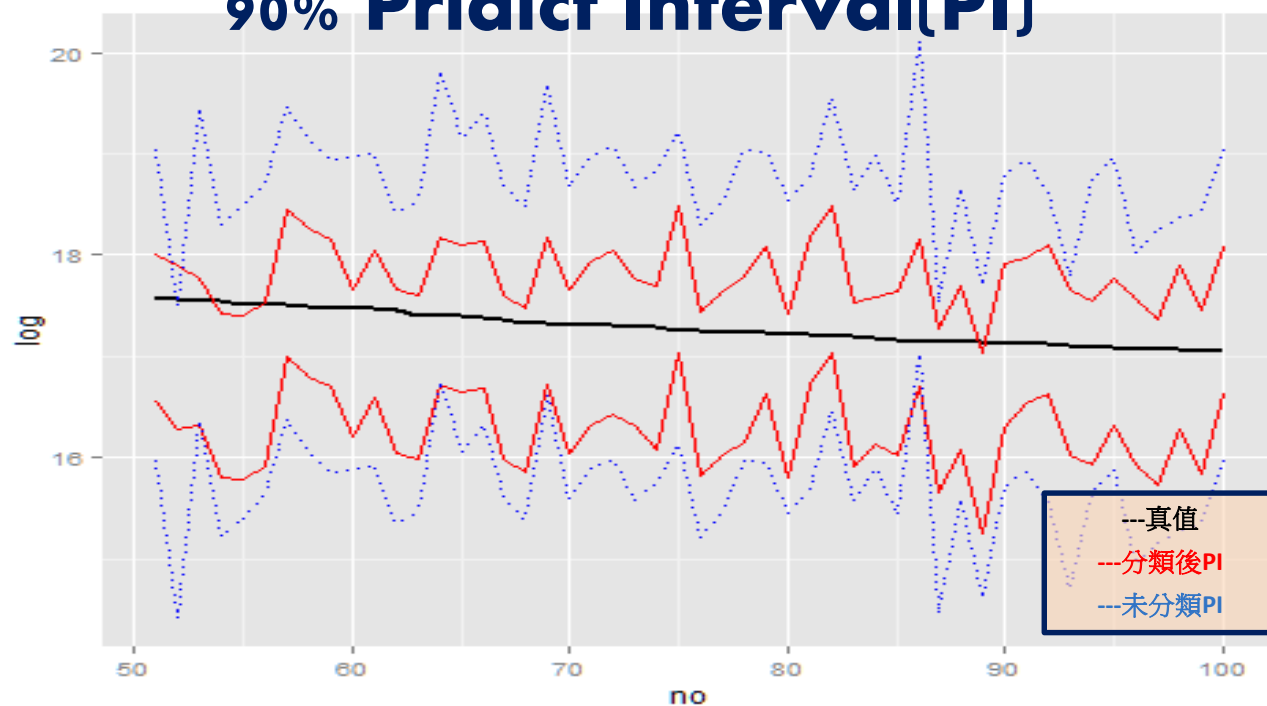
Linear Regression with MR

- MapReduce job1:
 - Calculating the $\mathbf{x}^T \mathbf{x}$ value
 - **Map:** `keyval(1, list(t(Xi) %*% Xi))`
 - **Reduce:** `sum(value)`
- MapReduce job2:
 - Calculating the $\mathbf{x}^T \mathbf{y}$ value
 - **Map:** `keyval(1, list(t(Xi) %*% Yi))`
 - **Reduce:** `sum(value)`
- Deriving the **coefficient** values $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$



預測結果

90% Pridict Interval(PI)

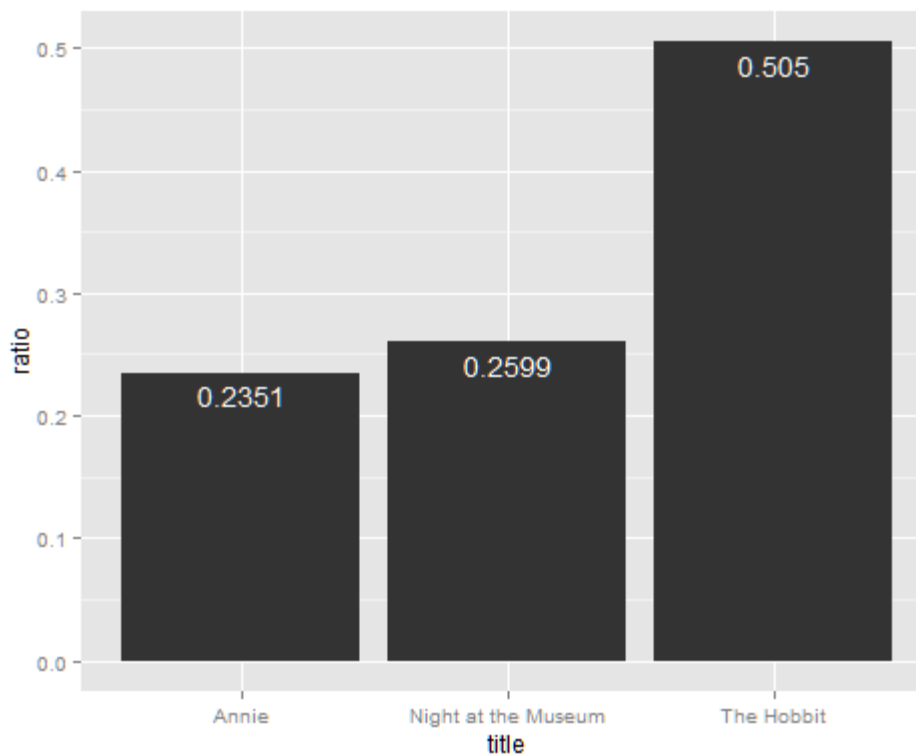


	未分類	分類後
correlation	0.75	0.79



預測結果

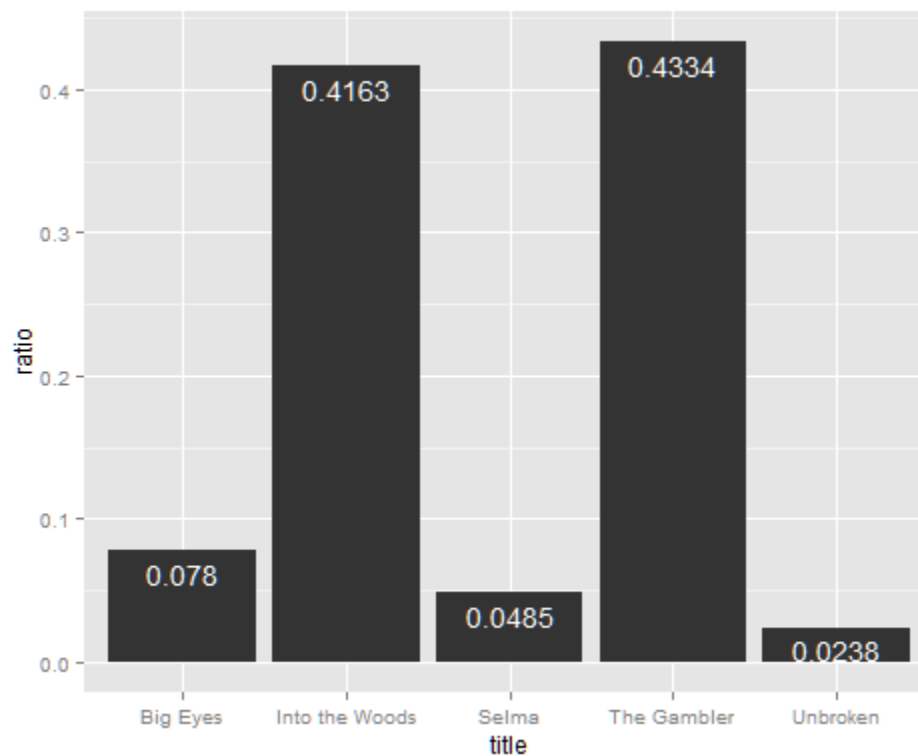
2014-12-19





預測結果

2014-12-26





未來展望

- 使用台灣電影院的資料，重新建立預測模型
- 蒐集電影上映前預告片評論，增加精確度
- 以Spark-MLlib完成資料分析部分



參考文獻

- [1] 台灣電影票房績效模型影響因素之研究
- [2] Predicting box-office success of motion pictures with neural networks
- [3] Forecasting box office revenue of movies with BP neural network



***Thanks for your attention
Q&A***