



元智大學

Yuan Ze University

基於辭典的意圖探勘方法

及其在電影票房預測的應用

A lexicon-based approach for intention mining and its
application in movie box office prediction

2023/07/05

研究生：鄭鈺均

指導教授：楊錦生

目錄

01 緒論

02 文獻回顧

03 研究方法

04 實驗結果與評估

05 結論與未來研究方向

緒論

研究背景、動機、目的

研究背景與動機

- 網路的普及促使社群媒體蓬勃發展，使用者可以更自由的在社群平台上表達自己的觀點，過去十年間用戶生成內容(User-generated content, UGC)與口碑(Word-of-Mouth, WOM) 的參考價值也隨著使用者數量的增加而持續提升。
- 對於電影等體驗類型的產品來說，由於其具有較強的感性屬性和情感性，因此UGC和WOM更是扮演了重要的角色。消費者會透過觀看其他人的評價和分享，來獲取更多關於電影的資訊和體驗，從而影響其最終的消費決策。
- 意圖探勘透過分析使用者行為和語言等方面，推斷他們的意圖。可用於預測線上購物者的購買意願，商業上，意圖探勘提供有價值的洞察和建議。

研究動機、目的

意圖探勘方法有分為機器學習法以及辭典法。參考Ahmad (2020)，研究提出以辭典法為基礎，透過網路字典進行電影意圖字的擴充，但這個方法存在以下缺點：

- 只侷限於單一單詞的替換，例如: recommend (O), can't wait(X)
- 字典內容缺乏即時性

針對以上缺點，本研究提出以Word2vec詞嵌入方法擴充意圖字，此方法有以下優點：

- 可以產生N-gram的同義詞
- 生成的同義字會包含熱門趨勢詞
- 可以依照專業領域，訓練不同的Domain Model

研究目的、問題



本研究挑選IMDb網路電影資料庫網站，作為本次數據來源。

延續意圖探勘字典法的應用，透過自行訓練一個與電影相關的Word2vec詞嵌入模型，接著進行意圖字擴充，建立了意圖辭典，並進一步探討意圖特徵與電影票房之間的關係。

研究問題

1. 使用Word2vec詞嵌入方式擴充同義詞是否能更全面地捕捉使用者可能使用的詞彙？
2. 擴充意圖字的次數對意圖辭典的建立是否產生影響？
3. 意圖特徵和機器學習方法對電影票房的影響程度如何相互比較？

文獻回顧

意圖探勘方法

使用者意圖探勘於商業應用中的研究案例

文獻回顧

- Dhaoui (2017) ; Khattak (2021)意圖探勘的文本分析大致分為兩種技術層：辭典法、機器學習法。辭典法使用意圖辭典進行意圖檢測；機器學習法則透過大量文本訓練模型自動判別意圖。
- Liu et al. (2016)使用微博和中國電影網站Wangpiao的資料集，研究了電影票房的預測。並提出了六個特徵類別來探測購買意圖，包括：詞袋模型、提及其他使用者、含有網址連結、含有表情符號、文字長度超過30字元以及含有意圖字。
- Zhang (2021)以Airbnb為例，研究中使用了LDA主題模型和情感分析方法來衡量基於線上評論的感知價值，並應用辭典法提取重複購買意圖變數。結果顯示，社會關係價值（服務態度、主客關係）對重複購買意願影響最為重要。
- Ahmad (2020)研究從Youtube電影預告片下方評論中提取電影購買意圖並進行情感分析。並辭典法，對評論文本進行TF-IDF提取前200個重要的Bigram詞組，然後利用人工判斷是否具有購買意圖，並以此過濾意圖種子字。接著使用線上字典therasus.com進行擴充。

研究方法

研究流程圖

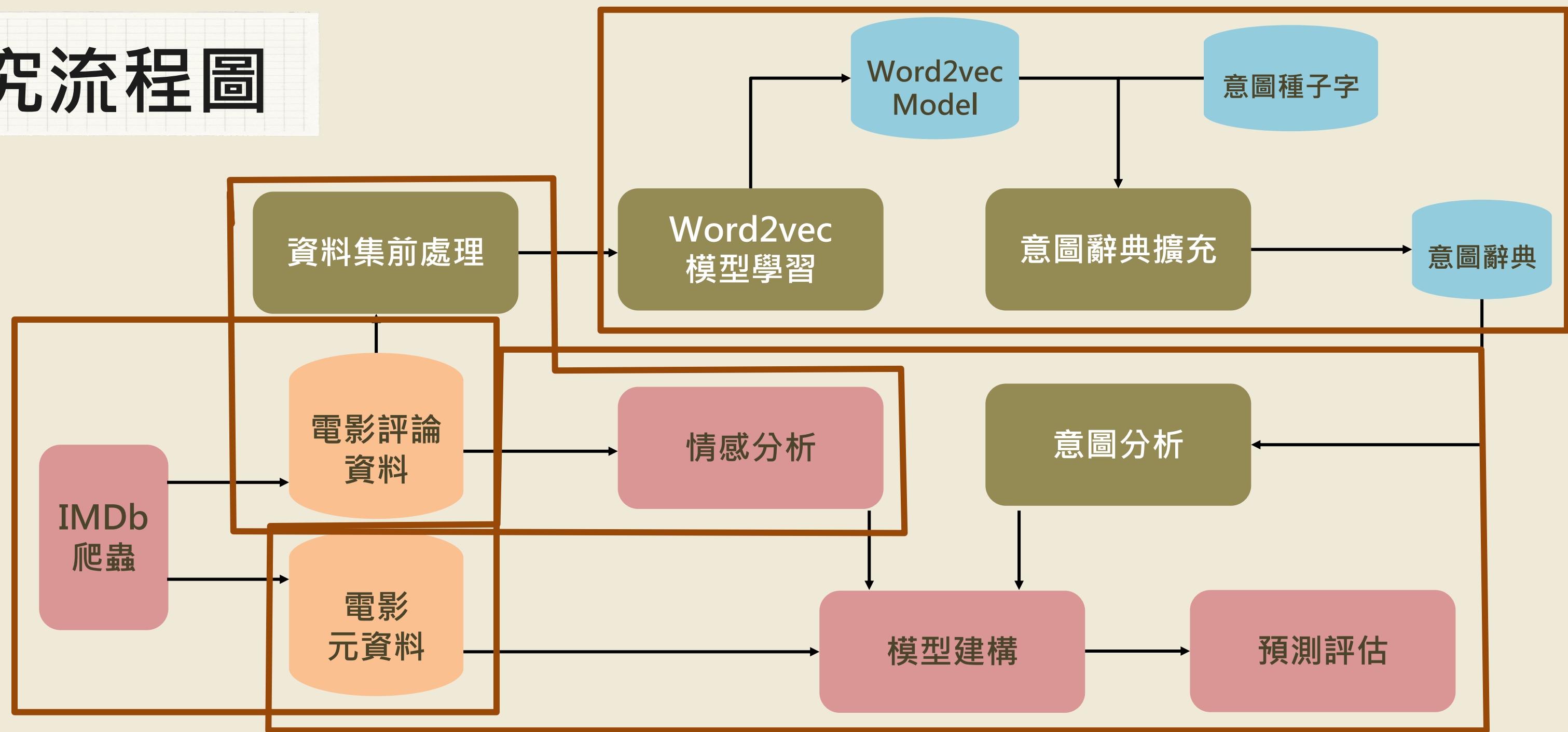
網路爬蟲與資料前處理

Word2vec模型學習與意圖辭典擴充

意圖分析與情感分析

模型建構與預測評估

研究流程圖



- 從IMDb網站抓取電影相關資料，並分成電影評論資料、電影元資料兩個資料庫
- 將要訓練Word2vec的文本進行前處理，同時也將實驗用的評論資料集進行情感分析
- 將前處理後的文本進行Word2vec模型學習，並產生Word2vec Model，將入意圖種子字匯入模型進行意圖辭典擴充，生成意圖辭典
- 考慮電影元資料、情感分析、意圖分析並建構模型，最後根據預測結果進行評估

網路爬蟲與資料前處理

- 本研究使用的數據集來自亞馬遜旗下的IMDb網站，全球最大的網路電影資料庫。平台提供了豐富的電影相關資訊，包括演員、導演、劇情、投票評分、票房等。
- 資料範圍是2020年至2021年的電影訓練資料集，評論資料集則涵蓋了2019年至2022年的電影，用於Word2vec詞嵌入模型的學習。
- 資料前處理只針對Word2vec模型學習的文本進行。Symeonidis (2022)曾經提到，動詞片語的組成(助動詞+動詞)更容易表現出使用者購買意圖。

網路爬蟲與資料前處理

紅色框分別代表:

- 1: 電影名稱、上映年分、電影分級、電影時長；
- 2: 電影類型；
- 3: 投票次數、IMDb Rating；
- 4: Metascore、使用者評論數；
- 5: 製作國家；
- 6: 預算、北美票房、北美首周周末票房、全球票房；
- 7: 使用者評論內容

沙丘

Original title: Dune: Part One

2021 · 6+ · 2h 35m

IMDb RATING

★ 8.0/10

669K

YOUR RATING

☆ Rate

POPULARITY

74

▲ 25

+

38 VIDEOS

99+ PHOTOS

Action

Adventure

Drama

A noble family becomes embroiled in a war for control over the galaxy's most valuable asset while its heir becomes troubled by visions of a dark future.

Director

Denis Villeneuve

Writers

Jon Spaihts · Denis Villeneuve · Eric Roth

Stars

Timothée Chalamet · Rebecca Ferguson · Zendaya

5.7K User reviews

518 Critic reviews

74 Metascore

Details

Release date

Countries of origin

Official sites

Languages

Also known as

Filming locations

Production companies

See more company credits at IMDbPro

Box office

Budget

Opening weekend US & Canada

Gross US & Canada

Gross worldwide

★ 6/10

Started off sensational, but eventually overlong with too much going on for too little happening.

paulclaassen 26 October 2021

Although the film is called 'Dune', the opening title refers to it as 'Dune Part One'. I knew, when I saw this, it probably should have been better to wait for Part Two before watching it. As a result some characters felt underdeveloped, and some simply vanished halfway through the movie. They also kept talking about Paul Atreides (Timothée Chalamet) being 'The One', but the one for what? This somehow reminded me of Neo from 'The Matrix', also being 'The One'.

Regardless, 'Dune' is a spectacle of note. From the stunning visuals, state of the art CGI, production design, and cinematography, to good performances from a stellar cast and a great score, this is one amazing movie. Sure, the film won't satisfy everyone's palate,

282 out of 458 found this helpful. Was this review helpful? Sign in to vote.

Permalink

Word2vec模型學習與意圖辭典擴充

- Word2vec為一種詞嵌入方法，它將詞轉為向量，並計算兩詞之間的相關性，本研究透過電影文本的匯入，自行訓練Word2vec模型，希望能藉此模型獲得相關同義字。
- 使用Word2vec模型進行意圖詞典擴充，將種子字進行第一次迭代，其生成的前十個同義詞，加上原先的種子字再刪除重複詞，就形成第一次擴充結果。第二次迭代的種子字為第一次迭代的結果(只有Iteration 1，不包含Seed)，一樣將生成的前十個同義詞加上第一次擴充的結果(Seed + Iteration 1)，刪除掉重複出現的詞，作為第二次擴充的結果。

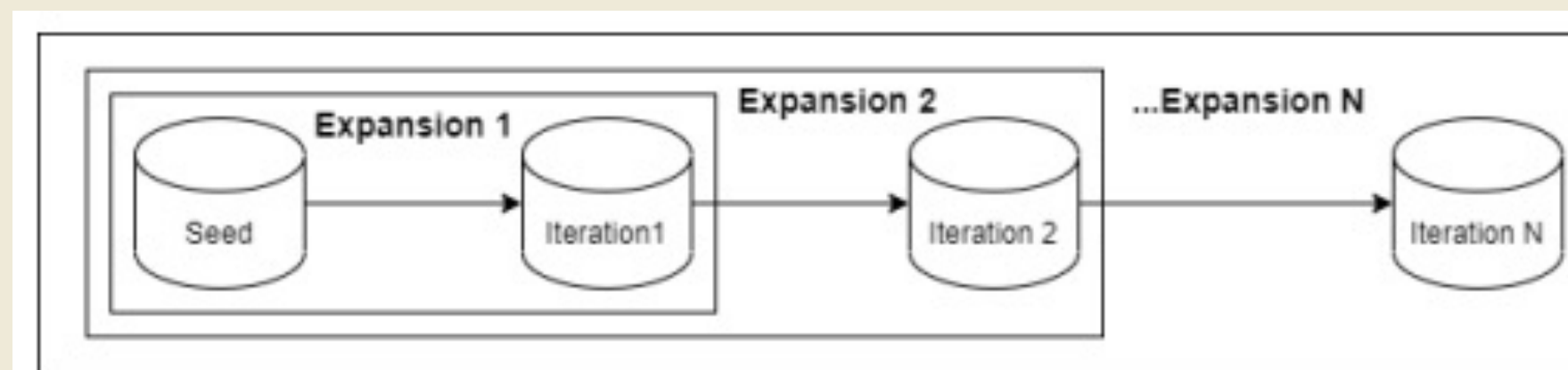


圖 5 意圖辭典擴充示意圖

意圖分析與情感分析

- 使用意圖探勘方法中的字典法，分析與電影元資料匹配的使用者評論資料。根據意圖辭典中的意圖字，在評論中出現符合辭典的字則標註為1，否則標註為0。計算每部電影評論中具有意圖的筆數，並生成相應的意圖特徵值。
- 採用VADER情感分析方法，情感值範圍在-1到1之間，分為中性、正面和負面三類。考慮到評論數量對分析的影響力，使用分位數將評論數量分為十個分位數，並根據評論數量給予不同的權重。權重計算公式參考Ahmad (2020)的方法，根據正面情感和負面情感的評論數量以及權重值來計算情感分析權重。

$$WPNratio = \frac{\text{positive sentiment}}{\text{negative sentiment}} * W$$

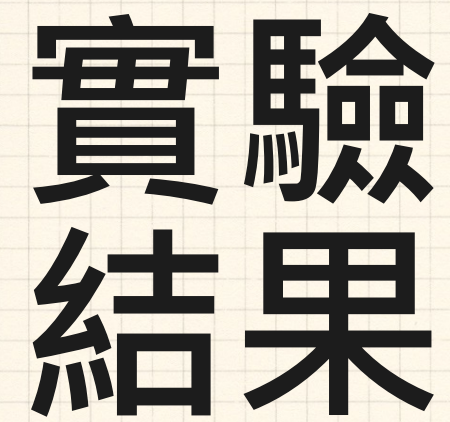

模型建構與預測評估

使用回歸型的機器學習法包含以下四種：

- 支援向量回歸(Support Vector Regression, SVR)
- 線性回歸(Linear Regression, LR)
- 決策樹(Decision Tree, DT)
- 隨機森林(Random Forest, RF)

接著資料進行十折交叉驗證，本研究使用以下五個回歸模型評估指標：

- 皮爾森積動差相關係數(Pearson Correlation Coefficient, CC)
- 平均絕對誤差(Mean Absolute Error, MAE)
- 均方根誤差(Root Mean Square Error, RMSE)
- 相對絕對誤差(Relative Absolute Error, RAE)
- 相對平方根誤差(Root Relative Squared Error, RRSE)



實驗結果

資料集與意圖種子字

敘述統計、相關分析

實驗結果、敏感度分析

資料集與意圖種子字

- 使用IMDb作為數據集來源。時間設定為2020年至2021年，網頁搜索中設定一些限制條件，如：長片、超過1000次投票、在美國上映，透過這些限制，共獲得835部電影。為了符合研究需求，只保留包含票房值的電影，同時，使用者評論數量需大於100筆。經篩選後，總共收集了91部電影作為實驗樣本。
- 用於Word2vec模型的評論集，時間則是設定在2019~2022之間，在沒有過濾票房缺失值的條件下，一共蒐集1,454部電影。

Symeonidis (2022) ; Ahmad (2020) ; Zhang (2021)以上三篇參考文獻，共整理出28個與意圖相關的字詞，過濾掉與電影無關的意圖字，並參照剩下的字再另外加入一些新的字詞，最後一共產生8個電影相關的意圖種子字。

表 2 研究資料集		
資料集	資料筆數	說明
電影元資料集	91 筆電影資料	所有與電影相關的屬性資料，為實驗模型主要的數據集
使用者評論集 2020~2021	130,985 筆使用者評論	從電影元資料 91 部電影中抓取的使用者評論，用於情感分析、意圖分析
使用者評論集 2019~2022	618,125 筆使用者評論	為 Word2vec 模型的訓練文本

表 3 意圖種子字	
recommend	next time
cannot wait	looking forward
must watch	would watch
should watch	watch again

敘述統計

表 4 敘述統計表-離散變數

變數名稱	各類別 占比
電影分級	G: 2(2%) ; PG: 13(14%) ; PG-13: 40(44%) ; R: 36(40%) ; NC-17: 0(0%)
電影續集	非續集: 49(54%) ; 續集: 42(46%)
電影類型	Action: 39(73%) ; Adventure: 35(38%) ; Drama: 32(35%) ; Comedy: 29(32%) ; Horror: 21(23%) ; Crime: 16(18%) ; Thriller: 14(15%) ; Mystery: 13(14%) ; Fantasy: 12(13%) ; Animation: 10(11%) ; Sci-Fi: 16(9%) ; Biography: 7(8%) ; Romance: 5(5%) ; Music: 5(5%) ; Family: 3(3%) ; History: 2(2%) ; Sport: 1(1%)
製作國家	United States: 79(87%) ; Canada: 13(14%) ; United Kingdom: 10(11%) ; China: 8(9%) ; Australia: 5(5%) ; Japan: 5(5%) ; Germany: 4(4%) ; South Korea: 2(2%) ; France: 2(2%) ; Mexico: 2(2%) ; South Africa: 1(1%) ; Hong Kong: 1(1%) ; Luxembourg: 1(1%) ; Sweden: 1(1%) ; Bulgaria: 1(1%) ; Spain: 1(1%) ; Chile: 1(1%)

- 電影分級: 最常見的電影分級是PG-13 (44%)，其次是R級(40%)，而且沒有NC-17級的電影。
- 電影續集: 54%的電影不是續集，而非續集電影數量稍多於續集電影。
- 電影類型: 很多時候一部電影會包含多種電影類型。73%的電影屬於動作類型，其次是冒險、劇情、喜劇、恐怖等類型。
- 製作國家: 很多時候一部電影會由多國家的工作團隊共同監製。美國佔了87%的比例，其他佔比超過5%的國家包括加拿大、英國、中國、澳洲和日本。

敘述統計與相關分析

- 全球票房的數值落差範圍非常大。
- 在意圖分析中，使用種子字的情況下，平均每部電影含有87則具有意圖的評論。
- 在進行兩次擴充後，意圖評論的平均數分別增長到604則和1,108則。

表 5 敘述統計表-連續變數

變數名稱	平均值	標準差	最小值	最大值
電影時長	115.81	24.89	83	242
投票次數	127,967.74	138,270.93	5,706	749,188
IMDb Rating	6.27	0.90	3.90	8.30
Metascore	55.75	15.73	22	91
使用者 評論數	1,439.40	1,585.04	121	8,070
預算	69,424,180	65,396,460	1,100,000	250,000,000
北美票房	59,060,400	96,515,830	93,147	814,115,100
北美首週 週末票房	19,523,060	32,623,820	42,165	260,138,600
全球票房	151,174,900	244,616,400	266,963	1,916,307,000
情感分析	7.22	6.89	0.78	35.88
意圖分析(Seed)	87.18	93.17	4	480
意圖分析 (Expansion 1)	604.88	707.74	41	3,495
意圖分析 (Expansion 2)	1,108.81	1,239.02	93	5,951

表 6 變數相關分析

變數名稱	相關係數
北美票房	0.9562
北美首週週末票房	0.9206
投票次數	0.7206
意圖分析(Seed)	0.6329
預算	0.6244
意圖分析(Expansion 2)	0.5252
意圖分析(Expansion 1)	0.5224
使用者評論次數	0.5085
情感分析	0.4194
IMDb Rating	0.3351
電影時長	0.2880
Metascore	0.1327

- 北美票房和北美首週周末票房與全球票房之間的相關係數均大於0.9，顯示它們之間存在著高度正相關。
- 投票次數、意圖分析、預算以及使用者評論次數與全球票房之間的相關係數均超過0.5，表示它們之間存在著中度正相關。

實驗結果-原始數據

以下包含四種不同的變數模型：
 No Intention -> 沒有加入意圖分析
 Seed -> 只使用意圖種子字的意圖分析
 Expansion 1 -> 擴充一次的意圖分析
 Expansion 2 -> 擴充兩次的意圖分析

- SVR以及LR方法在No Intention和Seed變數相對於其他Expansion結果表現最佳。
- DT方法，No Intention和Expansion 2變數在指標結果上達到相同的最佳結果。
- RF方法，Seed和Expansion 1的變數在評估結果上呈現最佳性能。

表 7 支援向量回歸(SVR)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.7217	0.7169	0.7148	0.7156
MAE	84,149,325	86,710,802	84,611,610	85,337,776
RMSE	174,446,451	173,571,081	175,902,469	175,499,209
RAE	59.0808	60.8792	59.4054	59.9152
RRSE	70.8651	70.5095	71.4566	71.2928

註:同一個表格中，若模型的指標結果最佳，則以粗體字表示

表 9 決策樹(DT)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.6381	0.582	0.5795	0.6381
MAE	105,077,752	105,895,437	107,354,936	105,077,752
RMSE	192,458,687	203,938,154	205,514,064	192,458,687
RAE	73.7746	74.3487	75.3734	73.7746
RRSE	78.1822	82.8455	83.4857	78.1822

表 8 線性回歸(LR)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.6227	0.6011	0.5841	0.6186
MAE	111,728,594	111,212,172	118,049,460	112,517,084
RMSE	198,442,001	200,904,771	204,910,461	199,151,292
RAE	78.4441	78.0815	82.8819	78.9977
RRSE	80.6128	81.6133	83.2405	80.9009

表 10 隨機森林(RF)模型評估結果

	No Intention	Seed	Expansion 1	Expansion 2
CC	0.4855	0.5097	0.5228	0.5119
MAE	96,852,737	94,225,344	94,320,697	96,965,321
RMSE	213,996,624	209,607,933	207,506,465	209,232,317
RAE	67.9998	66.1551	66.2221	68.0789
RRSE	86.9315	85.1487	84.2951	84.9961

實驗結果-原始數據

右側彙整表包含七種模型：

- Google以及Wiki為預訓練模型
- Phrases則是使用phrases模塊，進行短語句處理
- Unigram使用單詞的電影評論文本，不進行Bigram前處理

這些模型都做了兩次的擴充，但表中只會挑一個結果最佳的

- SVR在沒有意圖分析和意圖種子字模型結果較好。
- LR方法，Unigram Expansion 1模型的效果最佳。
- DT方法，No Intention和Expansion 2模型的結果相同。
- RF方法，預訓練模型的效果較佳。

整體而言，原始數值中擴充兩次的模型結果會比只擴充一次的模型效果更佳。

表 19 支援向量回歸(SVR)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.7217	0.7169	0.7156	0.7134	0.7213	0.7175	0.7193
MAE	84,149,325	86,710,802	85,337,776	86,365,802	85,268,926	84,655,792	84,467,152
RMSE	174,446,451	173,571,081	175,499,209	174,812,055	173,913,106	175,245,229	174,867,763
RAE	59.0808	60.8792	59.9152	60.637	59.8669	59.4364	59.4654
RRSE	70.8651	70.5095	71.2928	71.0137	70.6485	71.1896	71.0363

表 20 線性回歸(LR)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 1
CC	0.6227	0.6011	0.6186	0.5876	0.6272	0.6213	0.6339
MAE	111,728,594	111,212,172	112,517,084	114,082,910	109,857,002	112,176,879	103,736,268
RMSE	198,442,001	200,904,771	199,151,292	203,462,909	196,026,896	198,890,909	189,524,756
RAE	78.4441	78.0815	78.9977	80.097	77.13	78.7588	72.8327
RRSE	80.6128	81.6133	80.9009	82.6523	79.6317	80.7952	76.9904

表 21 決策樹(DT)模型評估結果彙整表

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 1	Unigram 2
CC	0.6381	0.582	0.6381	0.6349	0.5697	0.6213	0.6341
MAE	105,077,752	105,895,437	105,077,752	106,706,762	110,905,246	108,767,098	105,319,884
RMSE	192,458,687	203,938,154	192,458,687	194,960,560	208,591,333	196,811,629	194,166,205
RAE	73.7746	74.3487	73.7746	74.9183	77.866	76.3648	73.9446
RRSE	78.1822	82.8455	78.1822	79.1986	84.7358	79.9505	78.8759

表 22 隨機森林(RF)模型評估結果彙整表

	No Intention	Seed	Expansion 1	Google 2	Wiki 1	Phrases 2	Unigram 2
CC	0.4855	0.5097	0.5228	0.5254	0.5305	0.5147	0.5253
MAE	96,852,737	94,225,344	94,320,697	91,711,848	92,989,574	92,524,459	91,764,581
RMSE	213,996,624	209,607,933	207,506,465	207,092,221	206,324,883	209,101,486	207,135,197
RAE	67.9998	66.1551	66.2221	64.3904	65.2875	64.961	64.4274
RRSE	86.9315	85.1487	84.2951	84.1268	83.8151	84.943	84.1442

實驗結果-過濾數據

敘述統計表發現，全球票房數據中存在一些數值差異極大的離群值。為了提高模型的訓練效果，因此決定刪除全球票房值小於10,000,000或大於1,000,000,000的10部電影作品。最終，保留了81部電影作為本小節的訓練資料樣本。

- SVR方法，意圖種子和Google擴充一次的結果較好。
- LR方法，Seed和Unigram擴充一次的結果較佳。
- DT方法，Wiki和Unigram擴充兩次的效果最佳。
- RF方法，Phrases擴充一次和Unigram擴充兩次的結果較好。

整體而言，過濾後數值的模型擴充兩次的結果較只擴充一次的模型更佳。

表 35 支援向量回歸(SVR)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 1	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.7171	0.7293	0.7146	0.727	0.7191	0.6847	0.7142
MAE	73,199,210	72,172,107	74,440,388	72,135,746	73,191,600	88,897,937	74,918,448
RMSE	112,283,671	109,790,796	112,711,444	110,206,391	111,547,083	125,401,957	112,811,575
RAE	61.4745	60.6119	62.5169	60.5814	61.4681	74.6587	62.9184
RRSE	69.6634	68.1168	69.9288	68.3747	69.2065	77.8023	69.9910

表 36 線性回歸(LR)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 1
CC	0.6805	0.678	0.6735	0.6633	0.6785	0.698	0.7015
MAE	89,093,453	81,627,450	88,208,570	84,371,201	88,384,580	83,629,490	85,408,431
RMSE	125,940,820	120,577,744	125,860,702	123,669,321	125,742,748	120,367,669	118,956,115
RAE	74.8229	68.5528	74.0797	70.857	74.2276	70.2341	71.7281
RRSE	78.1367	74.8093	78.087	76.7274	78.0138	74.679	73.8032

表 37 決策樹(DT)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 1	Wiki 2	Phrases 2	Unigram 2
CC	0.6933	0.6758	0.6935	0.6794	0.6882	0.6956	0.7016
MAE	73,625,268	73,965,512	73,887,358	74,023,924	72,128,804	73,017,590	72,666,941
RMSE	116,950,429	119,142,205	116,992,785	118,299,670	117,227,957	116,379,611	115,170,732
RAE	61.8323	62.1181	62.0524	62.1671	60.5756	61.322	61.0275
RRSE	72.5588	73.9186	72.5851	73.3959	72.731	72.2047	71.4546

表 38 隨機森林(RF)模型評估結果彙整表(過濾後)

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 1	Unigram 2
CC	0.6857	0.6897	0.6851	0.6916	0.6747	0.6897	0.6958
MAE	74,971,525	76,351,235	75,866,525	75,009,110	74,841,969	72,165,674	72,860,330
RMSE	116,811,006	116,140,723	116,765,799	115,920,267	118,134,099	116,128,313	115,386,996
RAE	62.963	64.1217	63.7146	62.9945	62.8541	60.6065	61.1899
RRSE	72.4723	72.0565	72.4443	71.9197	73.2932	72.0488	71.5888

敏感度分析- 意圖種子字數量

這部分直接使用未過濾的二十八個意圖種子字進行意圖分析，並觀察其結果。

- SVR方法在四種方法中表現最佳，其中不使用意圖種子字的模型效果最好。
- LR和RF方法的最佳結果都是在Unigram擴充一次的情況下。
- DT方法的最佳結果是在進行Wiki擴充一次時獲得的。

整體而言，在二十八個意圖種子字的模型中，擴充一次的結果優於擴充兩次。

表 39 支援向量回歸(SVR)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	0.7217	0.7148	0.7144	0.7174	0.7166	0.7160	0.7179
MAE	84,149,325	85,947,716	85,484,580	85,067,596	84,920,588	85,168,153	85,109,875
RMSE	174,446,451	175,402,190	175,666,743	175,154,621	175,187,262	116,128,313	174,943,421
RAE	59.0808	60.3435	60.0183	59.7255	59.6223	59.7961	59.7552
RRSE	70.8651	71.2534	71.3609	71.1528	71.1661	71.2862	71.067

表 40 線性回歸(LR)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 1	Google 1	Wiki 2	Phrases 1	Unigram 1
CC	0.6227	0.6369	0.6319	0.6192	0.7166	0.6313	0.6312
MAE	111,728,594	108,560,134	107,280,387	110,757,793	113,523,942	107,626,139	106,748,384
RMSE	198,442,001	193,235,600	195,129,308	199,300,502	199,347,078	195,153,040	194,357,626
RAE	78.4441	76.2195	75.321	77.7625	79.7046	75.5638	74.9475
RRSE	80.6128	78.4978	79.2671	80.9616	80.9805	79.2767	78.9536

表 41 決策樹(DT)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 1	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.6381	0.6437	0.6408	0.6433	0.6450	0.6380	0.6438
MAE	105,077,752	104,661,245	104,737,651	103,938,448	103,475,098	106,045,280	104,040,431
RMSE	192,458,687	190,481,873	191,478,062	190,213,395	189,596,839	191,974,230	190,181,074
RAE	73.7746	73.4821	73.5358	72.9747	72.6493	74.4539	73.0463
RRSE	78.1822	77.3792	77.7839	77.2701	77.0197	77.9854	77.2570

表 42 隨機森林(RF)模型評估結果彙整表(二十八個意圖字)

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 1	Unigram 1
CC	0.4855	0.4944	0.5195	0.5034	0.5046	0.5220	0.5330
MAE	96,852,737	95,801,116	93,836,488	94,252,521	94,915,204	95,826,005	91,857,568
RMSE	213,996,624	212,793,482	208,097,286	211,174,357	210,968,053	207,849,833	205,894,457
RAE	67.9998	67.2615	65.8821	66.1742	66.6395	67.279	64.4927
RRSE	86.9315	86.4428	84.5351	85.7851	85.7013	84.4345	83.6402

敏感度分析- 北美票房

本小節以北美票房作為目標變數進行分析。

- SVR方法在本研究提出的模型擴充一次時效果最好。
- LR方法，Wiki擴充兩次的結果最佳。
- DT方法的最佳結果是在不加入任何意圖分析的情況下。
- RF方法，Wiki以及Unigram擴充一次的結果較好。

整體而言，在北美票房模型中，擴充兩次的結果優於只擴充一次。

表 43 支援向量回歸(SVR)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 2	Phrases 1	Unigram 1
CC	0.5164	0.4938	0.5706	0.5139	0.5396	0.5665	0.5391
MAE	42,486,717	44,460,244	40,259,614	42,264,520	41,896,969	40,425,268	40,775,646
RMSE	82,489,611	84,083,149	79,134,340	82,650,665	81,021,383	79,7108	81,024,761
RAE	80.6226	84.3675	76.3964	80.2009	79.5035	76.7108	77.3757
RRSE	84.8321	86.4709	81.3815	84.9977	83.3222	81.5732	83.3256

表 44 線性回歸(LR)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 2	Phrases 2	Unigram 1
CC	0.534	0.5352	0.5626	0.5109	0.5751	0.5407	0.5621
MAE	50,751,931	46,156,884	45,865,400	46,937,876	44,480,208	49,269,982	44,209,923
RMSE	86,144,997	85,014,952	82,239,384	84,679,332	79,556,070	84,975,987	80,144,086
RAE	96.3066	87.5871	87.034	89.0691	84.4054	93.4945	83.8925
RRSE	88.5913	87.4291	84.5748	87.084	81.8152	87.3891	82.42

表 45 決策樹(DT)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 2	Google 1	Wiki 1	Phrases 2	Unigram 2
CC	0.4996	0.4799	0.4887	0.4898	0.4847	0.4765	0.4693
MAE	45,087,034	45,143,222	46,682,285	45,992,936	46,086,101	46,962,031	48,013,819
RMSE	86,015,703	87,041,844	86,789,860	86,394,930	87,738,573	88,211,228	88,534,640
RAE	85.5569	85.6635	88.5841	87.276	87.4528	89.1149	91.1108
RRSE	88.4583	89.5136	89.2545	88.8483	90.2301	90.7162	91.0488

表 46 隨機森林(RF)模型評估結果彙整表(北美票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	0.4257	0.4136	0.3968	0.4144	0.4389	0.4152	0.4366
MAE	38,426,739	36,997,435	38,775,659	37,583,284	36,716,882	37,616,309	36,538,197
RMSE	87,308,541	87,830,747	88,939,719	87,884,217	86,450,060	87,814,959	86,503,271
RAE	72.9184	70.2061	73.5805	71.3178	69.6738	71.3805	69.3347
RRSE	89.7879	90.3249	91.4654	90.3799	88.905	90.3087	88.9597

敏感度分析- 北美首周周末票房

本小節以北美首周周末票房作為目標變數進行分析。

- SVR方法在所有方法中表現最佳，特別是在使用Phrases以及Unigram擴充一次的情況下有較好的結果。
- LR和DT方法，使用預訓練模型的效果較佳。
- RF方法中，不使用意圖分析的結果最好。

整體而言，在北美首周周末票房模型中，擴充兩次的效果優於只擴充一次。

表 47 支援向量回歸(SVR)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 1	Google 2	Wiki 2	Phrases 1	Unigram 1
CC	0.611	0.5846	0.6273	0.5882	0.6166	0.6378	0.6283
MAE	15,033,651	15,532,974	14,889,710	15,498,112	15,004,708	14,772,440	14,698,834
RMSE	25,925,845	26,498,415	25,511,530	26,437,033	25,806,001	25,285,855	25,535,264
RAE	78.6924	81.3061	77.939	81.1236	78.5409	77.3251	76.9398
RRSE	78.8502	80.5916	77.5901	80.4049	78.4857	76.9037	77.6623

表 48 線性回歸(LR)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 2
CC	0.6077	0.544	0.5947	0.5876	0.5523	0.589	0.5406
MAE	16,407,707	16,913,331	16,940,752	16,046,358	15,640,006	16,992,644	17,887,245
RMSE	26,682,864	28,959,060	27,034,514	27,889,725	26,258,315	27,642,260	30,200,478
RAE	85.8848	88.5315	88,6750	83.9934	81.8663	88.9466	93.6293
RRSE	81.1526	88.0753	82,2221	84.8231	79.8614	84.0705	91.851

表 49 決策樹(DT)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 2	Unigram 1
CC	0.5198	0.5924	0.559	0.6065	0.5714	0.5545	0.5576
MAE	15,605,522	14,975,100	15,399,591	14,263,166	14,690,946	14,555,463	14,856,864
RMSE	28,864,125	26,927,699	27,905,112	26,025,331	27,296,533	27,651,080	27,160,797
RAE	81.6858	78.3859	80.6079	74.6594	76.8986	76.1894	77.767
RRSE	87.7866	81.8972	84.8699	79.1528	83.019	84.0973	82.6061

表 50 隨機森林(RF)模型評估結果彙整表(北美首周周末票房)

	No Intention	Seed	Expansion 2	Google 2	Wiki 1	Phrases 1	Unigram 1
CC	0.4706	0.4308	0.4333	0.4506	0.4239	0.4285	0.4085
MAE	13,806,495	14,220,663	14,181,670	13,672,495	14,188,381	14,241,745	14,272,064
RMSE	28,667,291	29,539,970	29,412,806	29,040,562	29,727,615	29,674,514	29,969,918
RAE	72.269	74.4369	74.2328	71.5676	74.2679	74.5472	74.706
RRSE	87.188	89.8421	89.4553	88.3232	90.4128	90.2513	91.1497

結論與 未來研究方向

結論

- Word2vec詞嵌入方法確實可以有效地針對Bigram進行擴充，而如果為傳統字典法或是沒有特別設計過的模型，則是只能擴充單詞。
- 觀察彙整表的結果，對於同一個模型，意圖辭典擴充兩次的結果通常會優於只擴充一次，因為兩次擴充可以更廣泛地捕捉相關的意圖關鍵字。
- 在各種機器學習方法中，支援向量回歸(SVR)總是獲得最佳的模型評估結果。Unigram模型通常優於Bigram模型。
- 本實驗中刪除過度離散的票房值，建立過濾後票房的實驗集，其模型評估結果普遍優於原始數據集。
- 比較原始數據的8個種子字和敏感度分析的28個種子字，兩者的評估結果差異不大，且大多是8個種子字優於28個，但在決策樹模型中則有相反的結果。
- 對比原始數據的全球票房和敏感度分析的北美票房和北美首周周末票房，原始數據的結果優於敏感度分析的結果，可能是因為電影評論概括了多國使用者的評論，與北美票房關係存在一定差異。

未來研究方向

- **數據量不足**。建議增加實驗的電影筆數和Word2vec訓練文本數量，以提升結果準確率。
- **Bigram組成方法不夠彈性**。本研究原先判斷 Bigram 的條件必需為緊鄰的兩個詞。建議可以放寬組成Bigram 的條件，例如: “I will watch it again soon.”。
- **Unigram 的意圖種子字不夠多元**。有些 Unigram 的模型即便只能針對 Recommend 進行擴充，但是結果卻很好。因此建議可以提升 Unigram 做為意圖種子字的比例。
- **建議參考更多電影相關變數**。如: 導演、演員、製片商等，考慮它們對票房的影響，雖然這些類別變數在數據筆數較少時較難處理，但對觀眾消費動機可能有重要影響。