MovieLens 20M Dataset

資料分析實作

盧慶原

Source Code : https://github.com/duge03705022/MovieLens_Data_Analysis

Question

基於使用者對於每部電影的評分，
我們可以從中獲取那些資訊？

Preprocess

- 將movies.csv的title拆分為name以及year

	movieId	title	genres	name	year
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	Toy Story	1995
1	2	Jumanji (1995)	Adventure|Children|Fantasy	Jumanji	1995
2	3	Grumpier Old Men (1995)	Comedy|Romance	Grumpier Old Men	1995

Preprocess

- 將movies.csv的genres以dummy variable呈現

Adventure	Animation	Children	Comedy	...	Horror	Mystery	Sci-Fi	IMAX	Documentary	War	Musical	Weste
1	1	1	1	...	0	0	0	0	0	0	0	
1	0	1	0	...	0	0	0	0	0	0	0	
0	0	0	1	...	0	0	0	0	0	0	0	

Preprocess

- 將genome_scores對應到movies表格

007	007 (series)	18th century	1920s	1930s	1950s	1960s	1970s	1980s	19th century	...
0.02500	0.02500	0.05775	0.09675	0.14675	0.21700	0.06700	0.26275	0.26200	0.03200	...
0.03975	0.04375	0.03775	0.04800	0.11025	0.07250	0.04775	0.10975	0.09925	0.02050	...
0.04350	0.05475	0.02800	0.07700	0.05400	0.06850	0.05600	0.18500	0.04925	0.02675	...

Preprocess

- 計算ratings.csv裡users對每部電影的平均評分
- 計算ratings.csv裡每部電影有多少位user的評分

rating_mean	rating_count
3.921240	49695
3.211977	22243
3.151040	12735

Preprocess

- 移除評分數小於50的電影
- 移除有空值的列

- 處理後共有 10524 筆電影資料

Top 5 Movies

依照平均評分進行排序

前五名的電影均為2000年以前的經典電影
而教父系列作佔了其中兩名

另外，前五名的電影分類主要為Drama(劇情片)
與Crime(犯罪片)

	title	rating_mean	genres
315	Shawshank Redemption, The (1994)	4.446990	Crime|Drama
843	Godfather, The (1972)	4.364732	Crime|Drama
49	Usual Suspects, The (1995)	4.334372	Crime|Mystery|Thriller
523	Schindler's List (1993)	4.310175	Drama|War
1195	Godfather: Part II, The (1974)	4.275641	Crime|Drama

Bottom 5 Movies

倒數五名的電影分類較為分散

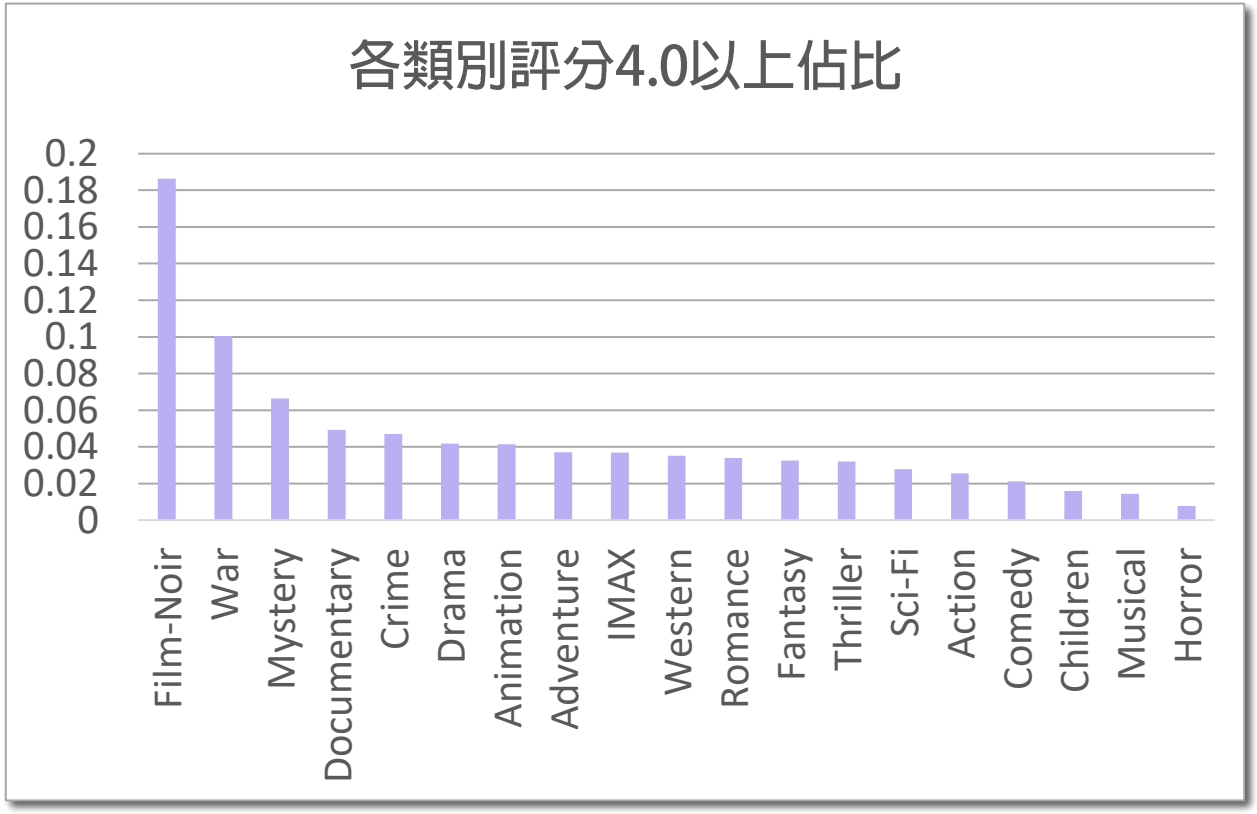
有Comedy(喜劇)、Musical(音樂劇)、Romance(文藝片)等

	title	rating_mean	genres
8176	SuperBabies: Baby Geniuses 2 (2004)	0.837321	Comedy
6373	From Justin to Kelly (2003)	0.973005	Musical|Romance
12013	Bratz: The Movie (2007)	1.105556	Comedy
4679	Glitter (2001)	1.124088	Drama|Musical|Romance
1746	Barney's Great Adventure (1998)	1.163484	Adventure|Children

Genres Analysis

篩選出評分為4.0以上的電影並分析各類別佔所有電影的比例

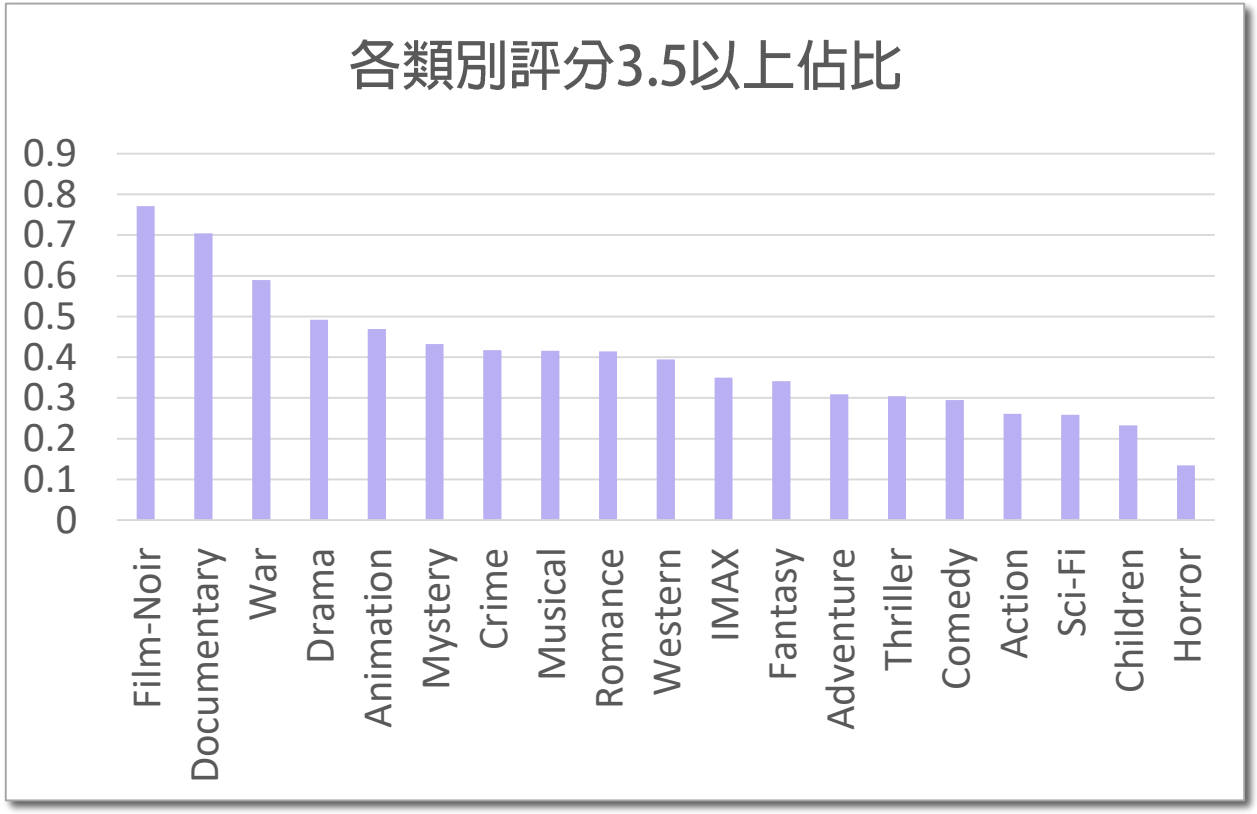
- 類別為Film-Noir(黑色電影)的電影有約18.6%其評分在4.0以上，該比例相對於其他類別來說相當高
- 而第二名為War(戰爭片)、第三名為Mystery(懸疑片)
- Horror(恐怖片)、Musical(音樂劇)與Children(兒童片)則鮮少有4.0以上的評分



Genres Analysis

篩選出評分為3.5以上的電影並分析各類別佔所有電影的比例

- 將範圍降低到評分3.5以上，可以發現Film-Noir(黑色電影)類別仍位居第一
- 有趣的是，原先位於第四名的Documentary(紀錄片)類別爬升到了第二名
- Musical(音樂劇)類別甚至從倒數第二爬升到第八名
- 而Horror(恐怖片)仍然墊底



Genres Analysis

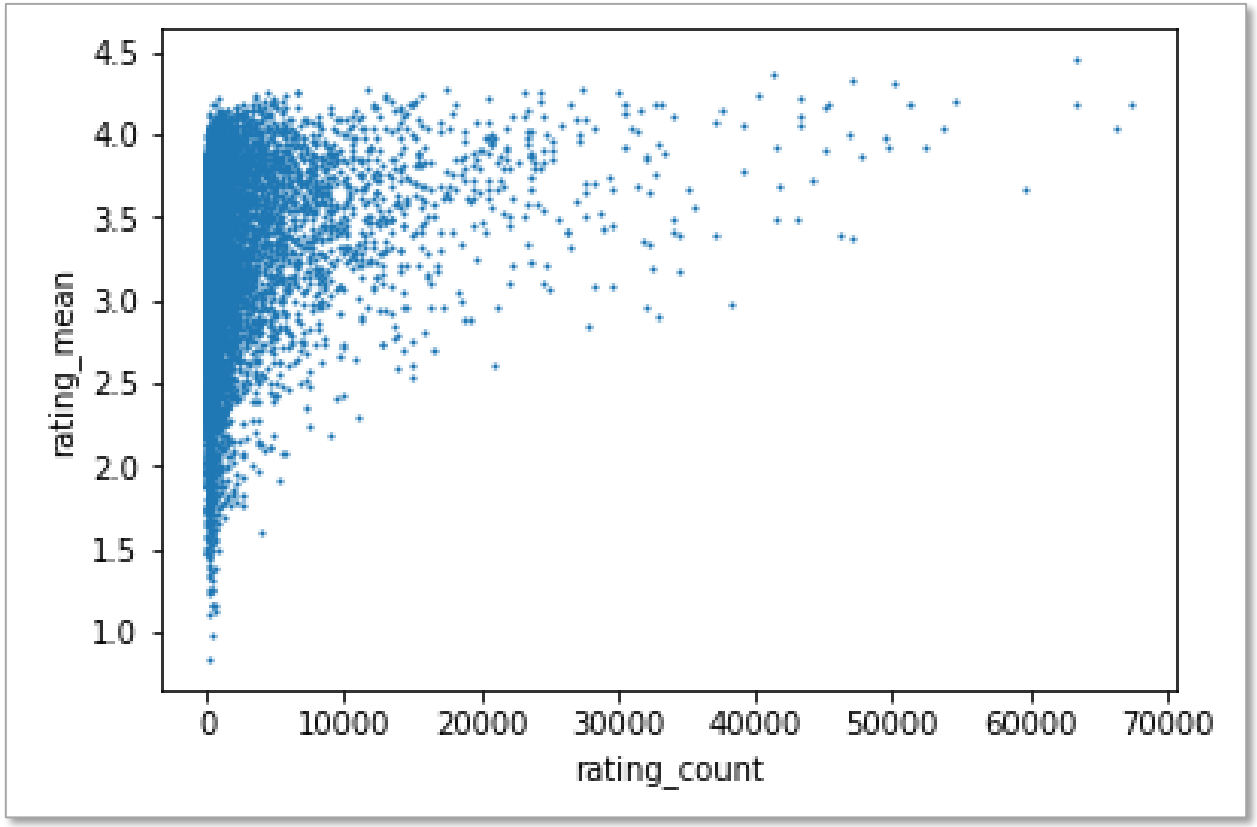
從以上的分析可知：

- Film-Noir(黑色電影)在MovieLens上普遍獲得好評，
- Musical(音樂劇)類別的表現中規中矩，不會有過高或過低的評價，而這與Bottom 5 Movies的觀察相悖
- Horror(恐怖片)容易被給低分

Rating Analysis

將每部電影的平均評分(rating_mean)與評分數(rating_count)畫成散佈圖

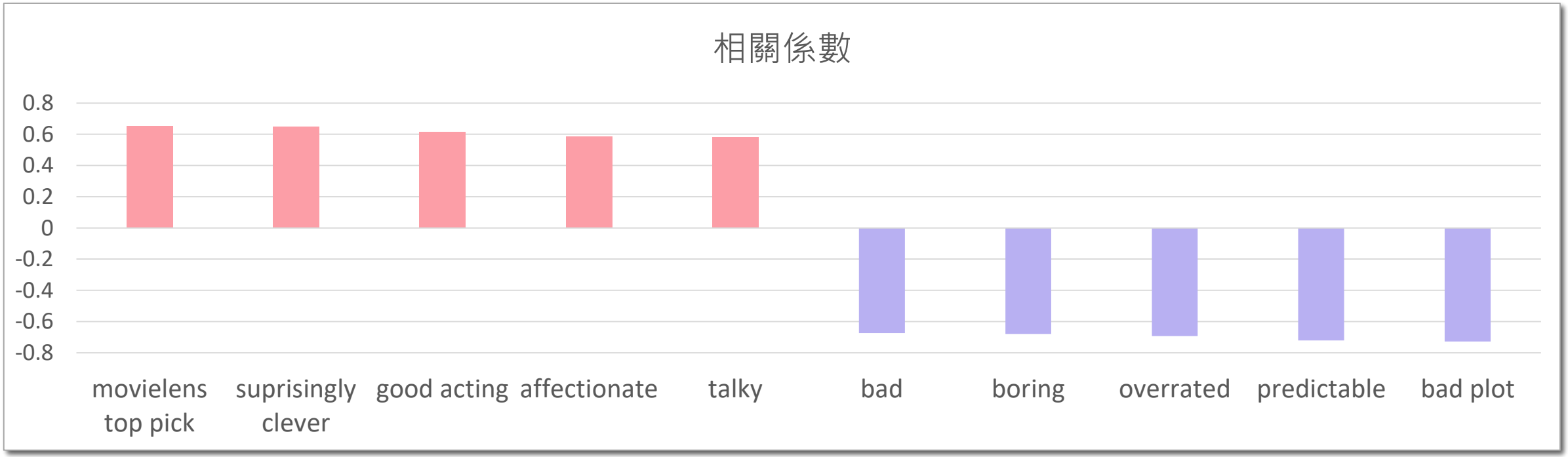
- 評分數10000以下的電影佔多數
- 隨著評分數量的增加，平均評分有上升的趨勢；可以推測出，評分數量高的電影，其平均評分通常較高



Correlation Coefficient

以Genome scores與平均分數計算出來的相關係數，圖中為前五高與前五低的標籤

- movielens top pick可能為該網站替高評分電影所加的標籤，因而才會有最高的相關係數
- 其中good acting、affectionate、talky等要素都與評分有高度正相關
- 而負相關中比較特別的則是overrated與predictable。



Relevance Analysis

以電影的類別做關聯性分析，藉此向使用者推薦相關類別的電影

- Confidence最高的為
  [懸疑片]->[驚悚片]
  代表有超過半數的懸疑片同時具有驚悚片的元素。
- 而Lift最高的組合
  [兒童片]->[動畫片]
  擁有最高的相關性

Rule	Support	Confidence	Lift
Mystery -> Thriller	0.029144365	0.525099075	3.428351502
Crime -> Action	0.014480534	0.514993481	3.362372469
Drama -> Mystery	0.012684214	0.479889043	3.133176954
Children -> Animation	0.017230002	0.457643622	10.96014287
Adventure -> Action	0.035633111	0.417346501	3.234198251
Children -> Adventure	0.016496811	0.395083406	4.627344424
Drama -> Crime	0.024965173	0.364561028	3.383632432
Fantasy -> Adventure	0.018733045	0.361898017	4.238666427
Adventure -> Animation	0.012904172	0.342745862	4.0143502

Conclusion

- 評分前五名的電影分類主要為Drama(劇情片)與Crime(犯罪片)
- 類別為Film-Noir的電影其為高評分的比率較高，而Horror類別的電影大多評分不佳
- 評分數量多的電影其評分通常也較高
- 利用關聯性分析結果(Ex.[懸疑片]->[驚悚片]等)可以向使用者推薦其他相關性高的電影