MOVIE RATING ANALYSIS

Danny Dean
Webster University
Data Mining Foundations, Fall 2009

Outline

- Data Set
- Research Description
- Analysis Technique
- Results

Data Set

- Netflix Prize Data Set
  - 17,000+ Movies
  - 480,000+ Customers
  - 100,000,000+ Ratings 

Research Description

- Can the number of words in a movie's title be linked to its average rating?

Analysis Technique

- Collect a random sample of 1,537 movies
- Group movies by the number of words in their title
  - “Paula Abdul's Get Up & Dance” = 6 words
- Calculate the average rating for each group

Results (1st Iteration)

Rank	Word Count	Number of Movies	Minimum Average Rating	Average Rating	Maximum Average Rating	Minimum Standard Deviation of Ratings	Average Standard Deviation of Ratings	Maximum Standard Deviation of Ratings
1	13	3	3.0531	3.6326	3.93752	0.98831	1.0732	1.20149
2	7	69	2.18548	3.47063	4.2	0.89694	1.15593	1.58683
3	9	23	2.20238	3.46036	4.46482	0.91425	1.16488	1.46528
4	10	9	2.78	3.43411	3.95552	0.98249	1.09407	1.16816
5	6	108	1.4	3.41791	4.25	0.76168	1.14549	1.54909
6	4	221	1.6699	3.32801	4.4839	0.84119	1.10333	1.43974
7	5	195	1.74194	3.31637	4.42814	0.85603	1.13473	1.54905
8	12	8	2.12821	3.2826	4.09787	1.00122	1.16659	1.34932
9	8	41	1.53801	3.28169	4.46073	0.79859	1.18768	1.42835
10	3	302	1.49624	3.22392	4.31399	0.74492	1.09592	1.54804
11	2	354	1.72414	3.09514	4.1426	0.83197	1.06617	1.47119
12	1	201	1.65546	3.08871	4.14223	0.83805	1.08277	1.40304
13	11	3	2.41975	2.71712	2.97059	1.09389	1.14149	1.16629

Results (2nd Iteration)

Rank	Word Count	Number of Movies	Minimum Average Rating	Average Rating	Maximum Average Rating	Minimum Standard Deviation of Ratings	Average Standard Deviation of Ratings	Maximum Standard Deviation of Ratings
1	13	3	3.43694	3.61706	3.90718	1.02981	1.17496	1.35625
2	18	1	3.60377	3.60377	3.60377	1.09789	1.09789	1.09789
3	7	58	1.92754	3.40083	4.25401	0.885	1.17033	1.47626
4	11	6	2.97403	3.38544	4.2735	0.97656	1.17269	1.41325
5	6	111	1.90244	3.38472	4.38008	0.82449	1.15591	1.54909
6	9	28	1.9469	3.34991	4.14086	0.99594	1.126	1.34626
7	4	231	1.95122	3.29863	4.59551	0.77926	1.12546	1.4762
8	5	173	1.81034	3.29477	4.51601	0.71354	1.12616	1.45519
9	8	53	1.53801	3.27189	4.00284	0.88949	1.18858	1.55593
10	3	316	1.6622	3.23906	4.67099	0.68251	1.08072	1.45177
11	10	11	2.22196	3.16551	3.5998	1.03038	1.09758	1.23482
12	15	1	3.16412	3.16412	3.16412	1.18733	1.18733	1.18733
13	2	351	1.6	3.12298	4.25188	0.80883	1.06803	1.42905
14	1	188	1.76829	3.11148	4.44671	0.8135	1.07134	1.37667

Results (3rd Iteration)

Rank	Word Count	Number of Movies	Minimum Average Rating	Average Rating	Maximum Average Rating	Minimum Standard Deviation of Ratings	Average Standard Deviation of Ratings	Maximum Standard Deviation of Ratings
1	13	3	2.75862	3.50678	3.93752	0.98831	1.07593	1.22063
2	12	4	3.25	3.46214	3.60117	1.05637	1.19329	1.36744
3	7	61	2.0102	3.45853	4.28295	0.885	1.15672	1.52658
4	8	40	2.36842	3.41979	4.26608	0.93875	1.14182	1.54406
5	6	112	2.16296	3.38283	4.33596	0.91927	1.15293	1.48393
6	14	2	3.01887	3.38055	3.74222	1.03262	1.15277	1.27292
7	5	189	1.76596	3.36516	4.6	0.71354	1.13672	1.58659
8	4	244	1.49776	3.29792	4.44833	0.82299	1.11032	1.42609
9	9	19	2.20238	3.29497	3.82423	1.00256	1.1556	1.31072
10	10	16	2.33884	3.2944	3.74044	0.98626	1.15836	1.41747
11	3	296	1.90576	3.25164	4.40801	0.85114	1.09642	1.3996
12	11	6	2.41975	3.21774	3.92373	1.0251	1.1662	1.38752
13	15	1	3.16412	3.16412	3.16412	1.18733	1.18733	1.18733
14	1	193	1.76829	3.15569	4.52261	0.8135	1.0801	1.34483
15	2	351	1.72115	3.08266	4.40408	0.81677	1.07193	1.5013

Comprehensive Results

- Movies with 1, 2, or 3 words in their title consistently showed up at the bottom of the table with regard to average rating.
- Movies with 4 or 5 words in their title maintained a position in the middle of the list (50^{th} percentile) over each iteration.
- Group of movies with 13 words in their title topped the charts over each iteration.

Comprehensive Results (cont)

1st Iteration				2nd Iteration				3rd Iteration		
Rank	Word Count	Average Rating		Rank	Word Count	Average Rating		Rank	Word Count	Average Rating
1	13	3.6326		1	13	3.61706		1	13	3.50678
2	7	3.47063		2	18	3.60377		2	12	3.46214
3	9	3.46036		3	7	3.40083		3	7	3.45853
4	10	3.43411		4	11	3.38544		4	8	3.41979
5	6	3.41791		5	6	3.38472		5	6	3.38283
6	4	3.32801		6	9	3.34991		6	14	3.38055
7	5	3.31637		7	4	3.29863		7	5	3.36516
8	12	3.2826		8	5	3.29477		8	4	3.29792
9	8	3.28169		9	8	3.27189		9	9	3.29497
10	3	3.22392		10	3	3.23906		10	10	3.2944
11	2	3.09514		11	10	3.16551		11	3	3.25164
12	1	3.08871		12	15	3.16412		12	11	3.21774
13	11	2.71712		13	2	3.12298		13	15	3.16412
14	-	-		14	1	3.11148		14	1	3.15569
15	-	-		15	-	-		15	2	3.08266

Citations

- McClave, S. (2009). A First Course in Statistics. Person Education, Inc.
- Microsoft Corporation. (n.d.). *MSDN*. Retrieved December 05, 2009, from Random Class: http://msdn.microsoft.com/en-us/library/system.random(VS.71).aspx
- UCI. (n.d.). *UCI Machine Learning Repository*. Retrieved December 12, 2009, from Netflix Prize Data Set: http://archive.ics.uci.edu/ml/datasets/Netflix+Prize

Thank you!

QUESTIONS?