Text Mining of Movie Synopsis by SAS Enterprise Miner

Yiyun Zhou - Ph.D. in Analytics and Data Science
Faculty Advisor: Jennifer Lewis Priestley, Professor of Statistics and Data Science
College of Science and Mathematics, Kennesaw State University

Abstract

This project described the method to classify movie genres based on synopses text data by two approaches：term frequency and inverse document frequency (tf-idf) and C4.5 decision tree. Using the performance comparison of the classifiers by manipulating the different parameters, the strength and improvement of this method in substantial text analysis were also interpreted. As the result, these two approaches are powerful to identify movie genres.

Data

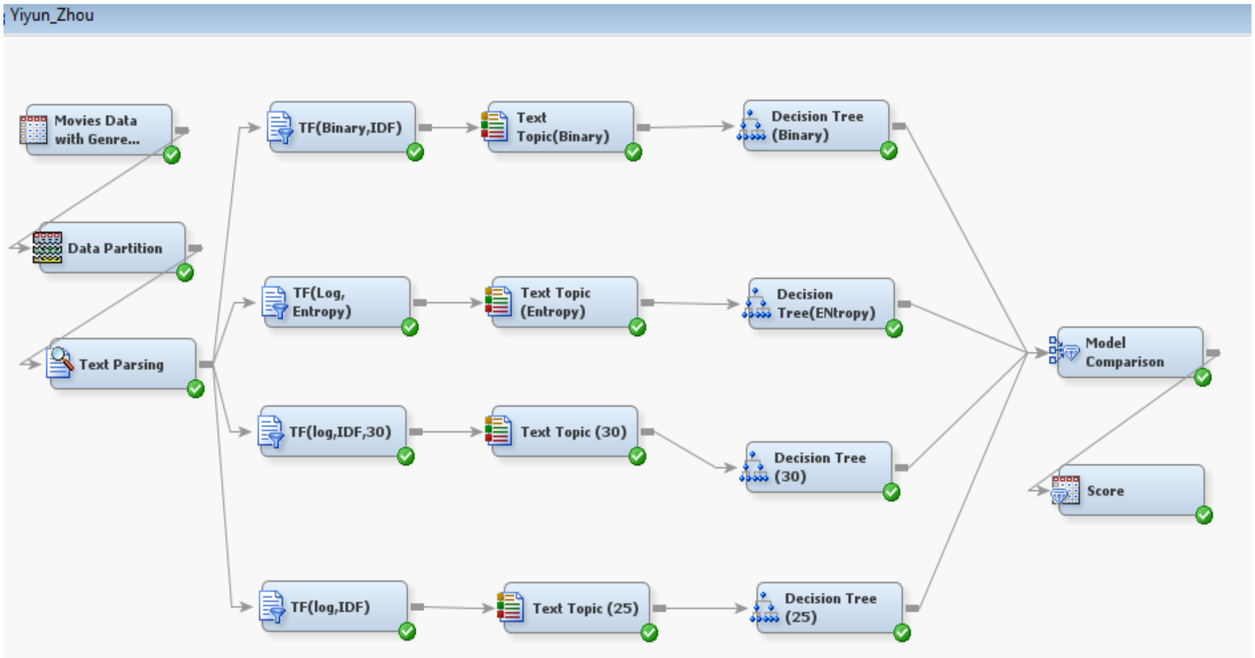
The dataset was downloaded from the Public IMDB dataset which contained 1527 synopses and 10 genres: action, comedy, documentary, drama, horror, kids/family, mystery, romance, scifi, and suspense. The dataset contained the synopsis of reviews on each movie, and the movies were assigned to a maximum of 5 genres based on relevancy of the movie's content. Data was split into two parts, 75 % for training, 25% for validation

Text Topic Modelling and Decision Tree

The quantitative data was reduced to reflect the number of parsed terms or documents to be analyzed and non-relevant data was eliminated (stop words and stemming). Then, term frequency and inverse document frequency (tf-idf) models produced a composite weight for each relevant term in each document.

Topical model	Weight	Term Weight	Min# Docs	Min# Topics
1	Binary	IDF	4	10
2	Log	Entropy	4	10
3	Log	IDF	30	10
4	Log	IDF	4	25

SAS Enterprise Miner Workflow



As we define that there were 25 topics for topical method 4, while others had 10 topics. After processing the text topic nodes, we can get strong information straight from the topics and infer the genres of each movie.

Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
1	0.057	0.018	+team,+coach,football,+player,+ga...	616	107
2	0.048	0.018	+jew,+family,+child,+mother,+death	720	119
3	0.067	0.018	+school,+mother,+parent,+girl,+stu...	623	177
4	0.062	0.018	+town,western,texas,+family,costner	702	167
5	0.120	0.017	+show,+recommend,acting,+sex,nu...	587	179
6	0.115	0.017	hollywood,+protagonist,+order,+cin...	654	120
7	0.082	0.018	+woman,york,principal,+husband,+f...	729	137
8	0.066	0.018	+comedy,+joke,+funny,humor,+laugh	616	166
9	0.090	0.018	+viewer,+moment,nearly,+minute,+r...	756	179
10	0.079	0.017	+bond,bond,connery,james,+agent	425	50
11	0.077	0.017	+alien,alien,earth,+crew,+special ef...	589	108
12	0.072	0.018	+war,+soldier,+battle,war,+army	565	120
13	0.069	0.017	granger,gauge,+revolve,tv,+writer	642	164
14	0.084	0.017	+kid,+age,jeffrey,+dog,+voice	550	119
15	0.078	0.018	acceptable,+language,+rate,+teena...	606	190
16	0.080	0.017	best,+win,+nominate,+oscar,suppo...	499	78
17	0.059	0.018	+president,+thriller,+murder,politica...	686	147
18	0.072	0.017	harry,+harry,dirty,dvd,san	508	74
19	0.054	0.018	+horror,+thriller,horror,suspense,+...	704	139
20	0.061	0.018	+crime,+heist,joe,max,+criminal	635	137
21	0.070	0.018	+love,+romance,romantic,ryan,charlie	752	180
22	0.065	0.018	+action,chan,martial,+stunt,+art	692	133
23	0.052	0.018	+woman,+sex,sexual,+relationship,...	720	162
24	0.048	0.018	+song,music,+musical,+sing,musi...	716	121
25	0.046	0.018	murphy,eddie,+child,police,family	680	114

Topics extracted from the topic node are groups of terms that create a formative definition of a document collection. For example, Topic ID 8 above includes: comedy, jokes, funny, humor, laugh, hilarious. These topics can then be used to form a definition for the genre "Comedy".

Mixed Genre Assessment

Model Node	Model Description	Valid: Misclassification Rate	Average Squared Error	Train: Misclassification Rate
Tree3	Decision Tree (30)	0.86744	.002552557	0.88220
Tree	Decision Tree(ENtropy)	0.87320	.002558212	0.88559
Tree4	Decision Tree (25)	0.87320	.002554867	0.88390
Tree2	Decision Tree (Binary)	0.87608	.002572575	0.89237

Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Tree3	Decision Tree (30)	0.42748	0.028821	0.46384	0.028447
Tree	Decision Tree(ENtropy)	0.43003	0.028385	0.45944	0.028407
Tree4	Decision Tree (25)	0.44020	0.028670	0.45679	0.028770
Tree2	Decision Tree (Binary)	0.44529	0.029143	0.46473	0.028958

The result of the training misclassification rates and the validating misclassification rates are similar among the 4 models. The main reason of the bad prediction is the mixed type of genre was difficult to fit, and most of the topics can only reflect one genre. From the dataset, 12.43% of movie had two kinds of genre, 34.21% of movie had three kinds of genre and the rest had four or five types of genre. After change the target variable to the main genre of movies, the training misclassification and validating misclassification rate decreased by less 50% and C4.5 decision tree performed better.

Conclusion

The topics which generated by tf-idf model present a promising text analysis outcome. Terms topics can map their meanings to genre. When topic was extracted for C4.5 decision tree model as input variable and genre was set as target variable, the validating misclassification rate was high for mixed genres targets and low for main genre targets.