Life's a Song! Mining Country Music Topics Using SAS® Text Miner

Deovrat Kakde, Saurabh Ghanekar, and Neetha Sindhu
Kavi Associates, Barrington, IL 60010

Introduction

Rich lyrics, often with a message, are a hallmark of American country music. Typical song topics in American country music include family, marriage, divorce, cheating, finding love, losing love, heartbreak, happiness, drinking, children, men, women, honky tonk, religion, politics and love of country. This paper demonstrates the use of Text Mining capability of SAS® Enterprise Miner 12.1 to identify topics in American country music. The lyrics of Billboard top 20 songs for the last 25 years are analyzed. The prominent topics as identified by SAS Text Miner are compared against the tags of last.fm to develop a measure of accuracy.

Methodology

Data Collection

The documents contain lyrics of Billboard top 20 songs for the period 1990 to 2000[1].The corpus consists of 500 documents. The lyrics were collected using two online sources[2]. Every song's lyrics were copied into a text file and placed in a common folder.

Analysis

SAS® Enterprise Miner 12.1 provides nodes that can be used to perform text mining. Following nodes were used in the analysis -

Text Import

This node was used to import the lyrics from various text files into one dataset. It creates a dataset with one column for lyrics of all the songs. Each row in this column contains the lyrics data for one song. The dataset also has other columns to support the analysis based on uri, creation time, etc.

Text Parsing

The Text Parsing node was used to parse lyrics text and create a term-document matrix. Text parts of speech tagging was disabled. Stop list was used which consisted of words like lyric, chorus, and website address. The resultant dataset, after parsing contained about 10,000 terms along with their roles, frequency and the number of documents in which they appear.

Text Filter

The Text filter node is used to reduce the dimensionality of term-document matrix. It can eliminate terms that appear very frequently or rarely. This is done by setting the properties of the node and also by manually adding or removing certain terms. The terms which appear in less than four documents were excluded from the analysis. The terms that had a very high frequency in too many documents, for example, get, but, etc. were also dropped. The interactive filter viewer was used to override the keep/drop decisions made by SAS in order to preserve important information.

Figure 1. Most frequently used words.

Artist		Childhood	Country Life	Drinking	Family	Heartbreak	Love & Happiness	Men	Nature	Religion	Women	Memories	Misc
George Strait		.	5	1	2	7	6	2	4	2	.	1	.
Alan Jackson		7	10	3	9	7	7	5	1	5	3	.	.
Tim McGraw		7	8	1	9	5	8	5	3	5	2	2	.
Brooks & Dunn		4	6	4	4	11	2	1	2	1	.	2	1
Garth Brooks		3	9	5	3	3	1	1	6	1	1	3	.
Kenny Chesney		5	4	2	5	2	5	5	2	5	5	3	.
Toby Keith		7	4	2	8	3	4	6	1	6	5	.	.
Brad Paisley		4	4	5	6	2	2	5	.	5	3	3	.
Rascal Flatts		.	3	1	4	6	4	4	3	4	.	4	.
Clint Black		1	4	.	.	6	.	.	3	.	1	.	.
Reba McEntire		1	1	.	1	6	.	1	1	1	1	1	.
Carrie Underwood		3	5	.	3	2	.	1	2	1	1	1	1
Keith Urban		2	2	.	2	2	2	1	1	1	1	1	1
Lonestar		1	4	.	2	2	4	2	2	2	1	3	1
Faith Hill		2	1	1	.	3	3	.	1	.	2	.	1

Figure 2. Topics for Top 15 Artists

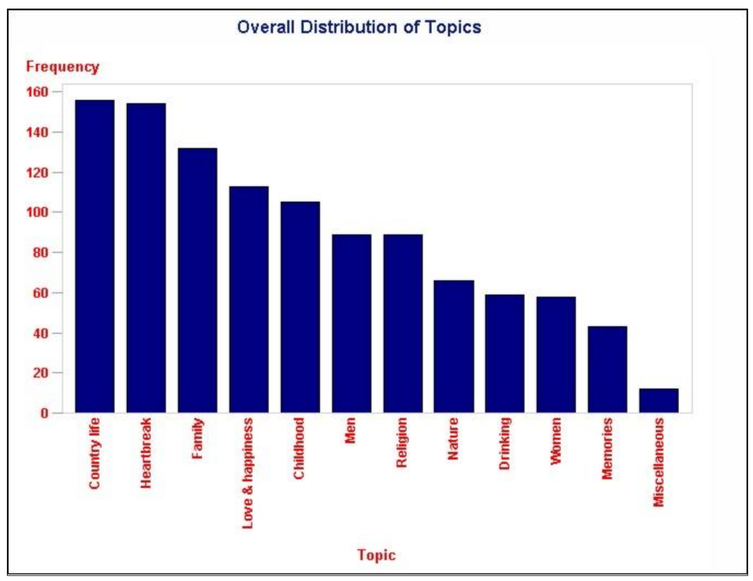

Figure 3. Overall Distribution of Topics

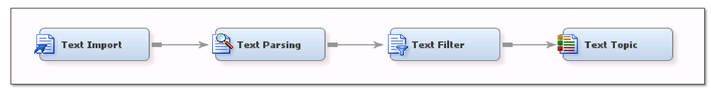

Figure 4. EM Process Flow.

Text Topic

The filtered and cleaned dataset consisting of relevant terms was used to assign the documents in the corpus to various topics. The Text Topic node classifies the documents into multiple topics based on the term weights. Based on the terms assigned to each topic, an appropriate label was selected.

Post Processing

A SAS dataset called *texttopic_train*, created as a standard result of the Text Topic node contained all the documents with scores assigned to each topic. Flag variables indicate whether or not a document belongs to a particular topic. The flags are set based on the document cut-off limits assigned in the Text Topic node. However, the order of relevance of topics for each document is not readily available from this data. This data was hence processed to identify the top three topics for each document. It was found that about 30% of the documents did not have any topic assigned to them. This was because the scores assigned to all topics for these documents were below the document cut-off limits. In such cases, the topic with the maximum score was considered as the most relevant topic for a document.

DISCUSSION

An exercise was performed by the authors to validate the assignment of topics to documents by SAS® Enterprise Miner. As the corpus consists of a manageable number of documents, it was possible to manually study each document to validate the topics and create a framework for future predictive modeling. The tags of Last.fm were also considered. Based on the validation exercise, it was found that 74% of the documents were correctly classified by SAS® Enterprise Miner.

Figure 3 shows the distribution of topics across the entire document corpus that was analyzed. It is important to note that the assignment of topics is not mutually exclusive. For example, a song may belong to the topic *Finding love* as well as *Country life* if it is about someone finding love in a small town. As seen in Figure 3, *Heartbreak* and *Country life* are the most popular topics, each with over 30% of the songs. It is not surprising that these are the most common topics, as any country music fan would agree! Other common topics include *Family*, *Love & happiness* and *Childhood*.

Country life	Heartbreak	Family	Love & happiness	Childhood	Religion	Men	Nature	Drinking	Women	Memories
play	love	man	love	country	man	man	free	dance	beautiful	miss
home	heart	god	feel	daddy	god	dad	blue	play	girl	alright
road	break	daddy	kiss	mama	pray	old	rain	beer	young	wish
song	lonely	old	baby	boy	heaven	boy	sun	bar	smile	feel
wild	tear	year	close	little	sunday	year	wild	country	love	home
country boy	night	boy	first	play	grace	grow	fall	song	look	time
truck	hurt	life	heart	alright	lord	fast	river	rock	baby	life
boot	leave	always	want	angel	bless	kid	day	down	hair	friend
american	long	live	girl	grow	preacher	wife	sail	band	honey	back
cowboy	cry	mama	hold	girl	faith	road	sea	music	sweet	old

Figure 5. Top Terms by Topic

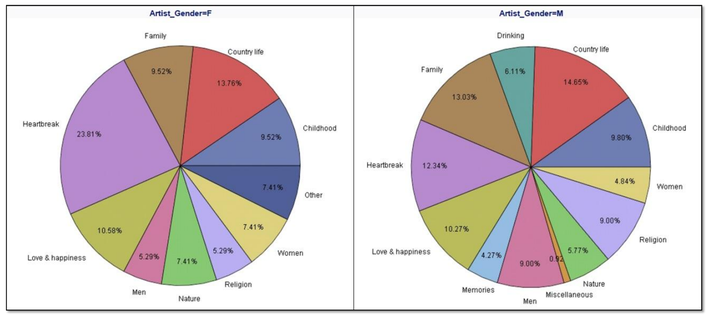

Figure 6. Topic Distribution by Gender

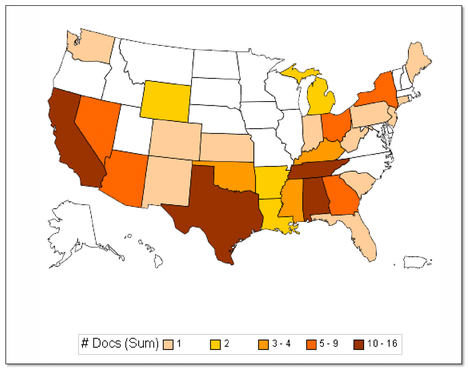

Figure 7. Country Artists love these States!

- Female artists are twice as likely to sing about heartbreak as compared to male artists
- Percentage of songs about *Country Life, Love & happiness* and *Childhood* are almost identical for both female and male singers

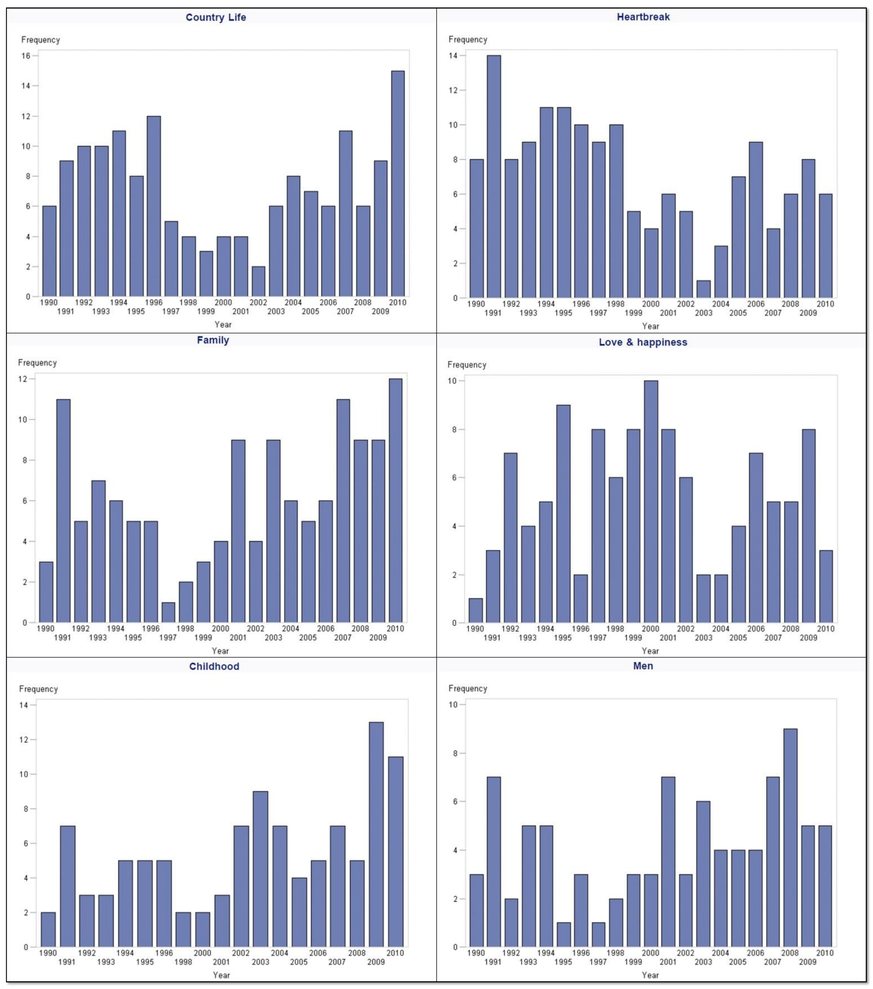

Figure 8. Yearly Trend for popular Topics

Countries they also talk about!

- Argentina
- Iran
- Iraq
- Mexico
- Panama
- France
- Italy
- Scotland
- Vietnam

References

[1] http://countrycharttalk.com/charts

[2] http://www.mldb.org/ ; http://easylyrics.org/

Kakde,
Ghanekar,
Sindhu

SAS® GLOBAL FORUM