Text Analytics and Brand Topic Maps

Nick Evangelopoulos, University of North Texas

ABSTRACT

The use of Text Analytics as a tool for strategic analysis of an organization, a leader, a brand, or a process, is examined. Software solutions SAS® Enterprise Miner™ and Base SAS are used to extract topics from text and create visualizations that identify the relative placement of the topics next to various business entities. Two case studies that identify and visualize brand-topic relationships in the context of branding, leadership, and service quality, are presented.

INTRODUCTION

The recent rise of big data has increased the interest in visualizations that can help the analyst explore and understand data in large volumes and a variety of formats. One of the big data challenges is the summarization of data sets in unstructured text format. Unstructured text appears from a variety of sources, including consumer comments on consumer report web sites, customer feedback in online surveys, reader comments on news sites, periodical publications, news stories, social media postings, and many others. In the present paper we advocate the use of correspondence maps as a tool that visualizes associations between topics extracted from unstructured text and brands, product lines, or other business actors. We focus on the association between topics and brands, and dub such maps as brand topic maps.

Correspondence analysis emerged in the 1930s, as a method related to Singular Value Decomposition (SVD) and principal component analysis (Hirschfeld, 1935). It was later popularized in France by Benzecri (1973). The method projects a set of contingency table categories into a smaller number of orthogonal categories, the principal components. The original categories are then represented in the principal component space as points on a plot (map). Points that are close to each other signify rows of the contingency table with similar conditional distributions across the columns, or, equivalently, columns with similar distributions across the rows.

In the 1980s and 1990s, correspondence analysis became popular in marketing research, where consumers' preferences in product attribute and brand are recorded in contingency tables, and then analyzed, in search of ideas for improving a brand's competitive position. An important challenge for managers, who are trying to interpret the results of such an analysis, is the difficulty of interpreting the axes (principal components). When they see attributes or brands being close to each other on the correspondence analysis map they perceive them as being similar but, without clearly defined axes, they have trouble understanding the map and the meaning of this proximity (Whitlark and Smith 2001). Map interpretability may increase when both sets of rows and columns of the contingency table (i.e., brands and attributes) are presented on the same map.

This paper is organized as follows. Since Brand Topic Maps start with topic extraction, we start with an illustrative example that walks the reader through topic extraction from a very small collection of documents. We continue with two case studies that address topic extraction and brand topic map visualization of consumer comments and annual report content. We conclude with some remarks and recommendations. The illustrative example is presented in the next section.

TOPIC EXTRACTION: AN ILLUSTRATION

In this section we illustrate topic extraction with the help of a very small vignette. Topic extraction is performed by performing Latent Semantic Analysis (LSA), the main analytic method used in SAS® Enterprise Miner™.

Consider a very small collection of customer comments that consists of six comments in free text form, as presented in Table 1. A variety of text analytic methods start the processing of this text collection with the tabulation of a term frequency matrix, where the frequency of each term in each document is recorded. Trivial English words listed on a stop list are excluded and frequencies are typically transformed and

normalized. This is shown in Table 2, where term frequencies have been subjected to a term-frequency, inverse-document-frequency (TF-IDF) transformation. For more information on this and other similar types of transformation, see Chakraborty, Pagolu and Garla (2013, pp. 99–100). We will now process this data set using LSA, which is one of the text analytic technologies that power the Text Miner module of SAS Enterprise Miner. The next step in topic extraction is the SVD of this matrix. The decomposition produces the six eigenvalues shown on the scree plot in Figure 1. Based on this scree plot, a two-topic solution seems to be appropriate.

ID	TEXT
P001	The car is fun to drive
P002	Fun to drive, I love it!
P003	I love driving this car
P004	Great car that comes at an affordable price
P005	I love the price, very affordable
P006	Great price

Table 1. Six customer comments

TERM	P001	P002	P003	P004	P005	P006
afford	0	0	0	0.598	0.746	0
car	0.471	0	0.577	0.377	0	0
drive	0.471	0.471	0.577	0	0	0
fun	0.746	0.746	0	0	0	0
great	0	0	0	0.598	0	0.846
love	0	0.471	0.577	0	0.471	0
price	0	0	0	0.377	0.471	0.534

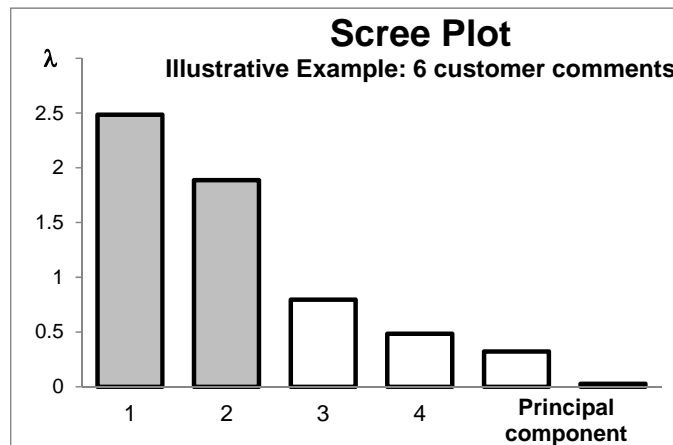**Table 2. Term Frequency Matrix after TF-IDF Transformation**

Figure 1. Scree Plot of Eigenvalues Produced by Singular Value Decomposition

The final outcomes of the two-topic solution include a matrix of term loadings, presented in Table 3, and a matrix of document loadings, presented in Table 4. After some examination of these two tables, we are able to label topic T2.1 as *the car is fun to drive, I love it*, and topic T2.2 as *great price/affordable*.

The two-topic solution presented above seems to be quite appropriate in (i) helping the analyst to gain a conceptual understanding of the document collection without the need for manual reading of the documents, and (ii) representing the six documents in a new, two-dimensional space. Nevertheless, we now proceed to extract a three-topic solution. This is done in order to illustrate LSA's powerful ability to recast the collection of documents in alternative concept spaces that correspond to different levels of semantic granularity.

The outcomes of the three-topic solution include, again, a matrix of term loadings, presented in Table 5, and a matrix of document loadings, presented in Table 6. In the three-topic solution, the first topic, *the car is fun to drive, I love it*, remains intact and reappears as T3.1. However, topic T2.2 now splits into two

sub-topics, T3.2, *great price*, and T3.3, *affordable/I love it*. For some additional discussion on LSA topic extraction and related recommendations, see Evangelopoulos, Zhang and Prybutok (2012), and Evangelopoulos, Ashton, Winson-Geideman and Roulac (2015).

TERM	T2.1	T2.2
afford	0	0.7512
car	0.6024	0
drive	0.8588	0
fun	0.8874	0
great	0	0.8716
love	0.6429	0
price	0	0.778

Table 3. Two-Topic Solution: Association of Extracted Topics with Terms Using Term Loadings

TEXT	T2.1	T2.2
The car is fun to drive	0.8908	0
Fun to drive, I love it!	0.9038	0
I love driving this car	0.7919	0
Great car that comes at an affordable price	0	0.9366
I love the price, very affordable	0	0.7036
Great price	0	0.8088

Table 4. Two-Topic Solution: Association of Extracted Topics with Documents Using Document Loadings

TERM	T3.1	T3.2	T3.3
afford	0	0	0.8144
car	0.5829	0	0
drive	0.851	0	0
fun	0.9385	0	0
great	0	1.0204	0
love	0.5142	0	0.6136
price	0	0.6585	0

Table 5. Three-Topic Solution: Association of Extracted Topics with Terms Using Term Loadings

TEXT	T3.1	T3.2	T3.3
The car is fun to drive	0.9397	0	0
Fun to drive, I love it!	0.8951	0	0
I love driving this car	0.7163	0	0
Great car that comes at an affordable price	0	0.8187	0.4756
I love the price, very affordable	0	0	0.9343
Great price	0	0.9575	0

Table 6. Three-Topic Solution: Association of Extracted Topics with Documents Using Document Loadings

CASE STUDY 1: CONSUMER COMMENTS ON THREE SMALL SEDANS

BACKGROUND AND DATA COLLECTION

The Kelley Blue Book Co.® web site (Kelley Blue Book Co. 2016) allows consumers to post reviews on different vehicle models. Consumers can rate their vehicle's model, list pros and cons, and write their own review. In this case study I examine consumer reviews on three popular small sedans, dubbed here as Car Model A, Car Model B, and Car Model C, respectively, for my personal use as a potential buyer. The questions I am trying to answer are:

- What high-level concepts are consumer review contributors trying to communicate to me, as a potential buyer, in their posted comments?

- How can I visualize such concepts in a way that allows me to compare the three car models against

each other?

The KBB web site offers the capability to browse through posted consumer reviews for personal use. For commercial purposes, KBB can enter into an agreement with third parties, in which case data collection on consumer reviews can be automated. For the purposes of this case study, I browsed through consumer reviews related to the three selected vehicle models, and took notes in plain text, organized into individual paragraphs. Personal information from the original review contributors was completely disregarded. Data collection took place in late 2015. Table 7 summarizes the collected data, listing the three car models, their parent companies, and the corresponding total number of paragraphs of unstructured text generated from the consumer reviews.

Car Model	Parent Company	Total paragraphs in my notes on consumer reviews
Car Model A	Motor Company X	438
Car Model B	Motor Company Y	491
Car Model C	Motor Company Z	300
Total		1229

Table 7. Summary of the Original Source of Collected Data for Case Study 1

TOPIC EXTRACTION IN SAS TEXT MINER

We now perform topic extraction in SAS Text Miner. Start by creating a new analysis project in SAS Enterprise Miner. Create a new SAS Library, import the consumer reviews data file as a Data Source, open a new analysis diagram canvas, select the Text Mining tab on the toolset space and build the diagram shown on Display 1, where Enterprise Miner/Text Miner version 13.2 is used. The analysis diagram consists of a Data Source node, a Text Parsing node, a Text Filter node, a Text Topic node, and a Save Data node. For details on the use of the Text Mining nodes, see Chakraborty et al. (2013, pp. 122–130). In the Text Parsing node settings, a list of synonyms that can be customized, is included. An example use of a customized synonym list is the situation where the documents include mixed American and British English spelling. The text Parsing node includes a list of terms to be excluded from further analysis, or a stop list. This can also be customized. For example, the stop list can include the names of different car models, so that topic extraction results disregard language use patterns that revolve around specific car models, and focus on overarching high-level concepts.

In the Weightings section of the Text Filter node settings, select Frequency Weighting = None, and Term Weight = Inverse Document Frequency. In the Term Filters section of the Text Filter node settings, set Minimum Number of Documents = 4, so that terms that appear in less than four documents in the collection are filtered out. The scree plot of eigenvalues (see Figure 2) suggests 7, 12, 18, 28, etc., as candidate high-level dimensionalities. For the purpose of this illustration, we select $k = 12$. Therefore, in the Learned Topics section of the Text Topic node settings, set Number of Multi-term Topics = 12.
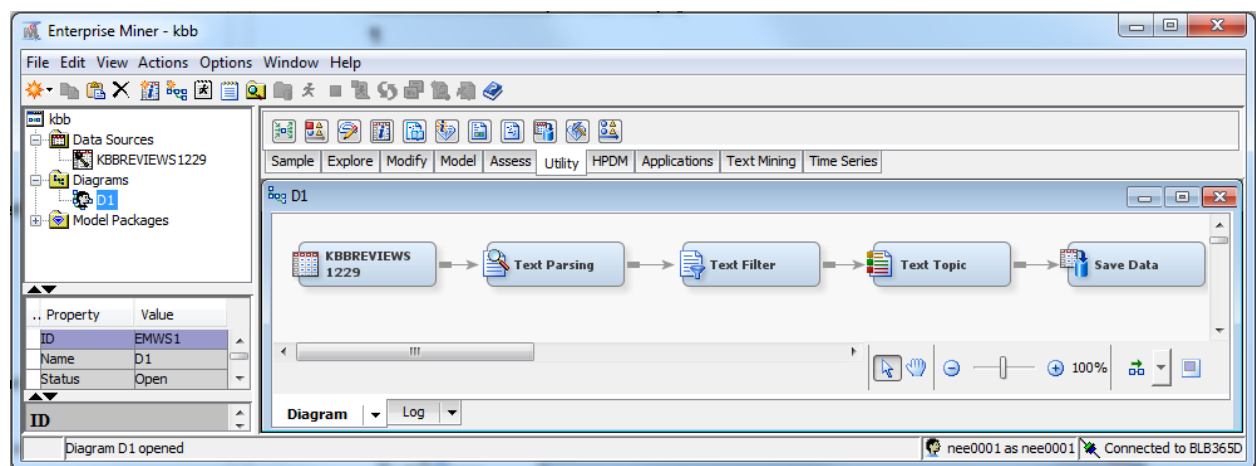
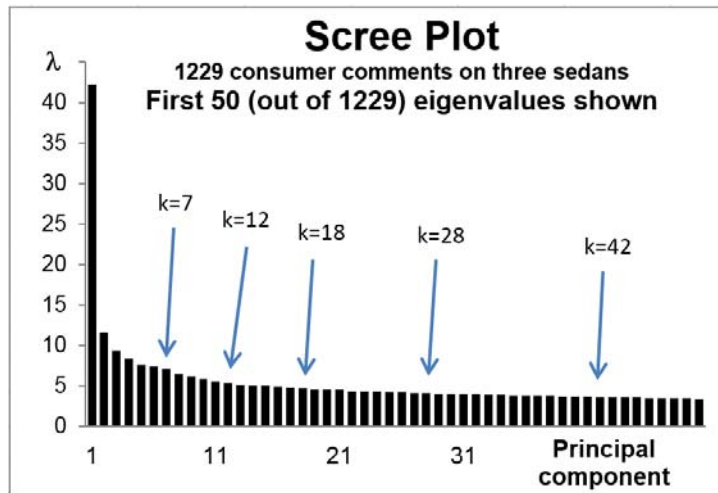**Display 1. Analysis Diagram for Topic Extraction in SAS Text Miner**

Figure 2. Scree Plot of Eigenvalues Produced by Singular Value Decomposition

Run the analysis diagram from the Text Topic and view the Results. Text Miner produces a summary of the extracted topics that lists, for each topic, characteristic terms that help identify the topic, and number of related documents. With the help of the characteristic terms and some browsing through associated documents, the 12 topics are labeled as shown in Table 8.

Topic ID	Topic Label	Topic Terms	# Docs
1	Space and comfort	+seat,room,back,+back seat,+leg	142
2	Transmission	+transmission,manual,+shift,automatic,focus	166
3	Highway MPG	mpg,+highway,+mile,+drive,fuel	170
4	Power steering system	+system,power,+package,+steer,+wheel	121
5	Great gas mileage	+mileage,gas,gas mileage,+good,great gas mileage	216
6	Good value	+vehicle,+purchase,value,+month,good	149
7	Fuel economy	+good,+car,fuel,economy,fuel economy	175
8	Great car	+great,great car,+car,+love,great gas mileage	171
9	Customer review text	text review,text,customer,+review,+read	40
10	Average MPG	mpg,+mile,+average,driving,+highway	132
11	I love this car	+car,+love,+drive,+buy,+look	200
12	Car problems	+problem,+replace,+dealer,+mile,+car	158

Table 8. Extracted Topics and Corresponding Topic Labels

CORRESPONDENCE ANALYSIS IN SAS/STAT®

After extracting topics from unstructured text, the next step in the creation of a Brand Topic Map involves corresponding analysis. In preparation for correspondence analysis, the data set saved by the Save Data node needs to be restructured. Table 9 shows the original structure of the file, where an indicator variable is created for each extracted topic. Only the first three rows of data are shown. The original format needs to be converted into a long categorical data format, where a categorical variable *Topic* holds the topic value of each observation. This is shown in Table 10. Again, only the first three rows of data are shown. Documents 1, 8, and 15, are all associated with topic 1, *Space and Comfort*.

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	_DOC_	Car
1	0	0	0	0	0	0	1	0	1	0	0	1	Car C
0	0	0	0	0	1	0	0	0	0	0	0	2	Car C
0	0	0	0	0	0	1	0	0	0	0	0	3	Car C

**Table 9. Format of the Data Set Produced by the Save Data Node**

Topic	Document	Car
Space and comfort	1	Car C
Space and comfort	8	Car C
Space and comfort	15	Car C

Table 10. Long Categorical Data Format

The data set in long categorical data format is now ready for correspondence analysis. The Base SAS code shown below assumes the presence of a data file called ConsumerTopics.csv in location d:\temp, and uses PROC CORRESP to perform correspondence analysis:

```
proc import out= work.topics
    datafile= "d:\temp\ConsumerTopics.csv"
    dbms=csv replace;
    getnames=Yes;
proc corresp all data=topics outc=Coord;
    tables Topic, Car;
run;
```

Execution of the above code produces a contingency table, chi-square information, profiles, and all results of the correspondence analysis. The OUTC= option creates an output coordinate data set. The TABLES statement specifies *Topic* and *Car* as the row and column categorical variables, respectively. Display 2 presents the correspondence analysis map, which is a Brand Topic Map. The three-by-twelve contingency table is projected onto a two-dimensional space. Since the contingency table has three rows, the complete information gets projected on two dimensions, therefore the two principal components account for 100% of the variability in the contingency table. Based on the 12 topics, Dimension 1 seems to be a contrast between negative features (transmission, car problems) and positive features (space and comfort, fuel economy, good value.) Dimension 2 seems to be a contrast between *space and comfort* and *I love this car*, versus *fuel economy* and *good value*. Then, Car A is strongly associated with *car problems* and *transmission*, Car C is associated with fuel economy, and Car B with *space and comfort*.
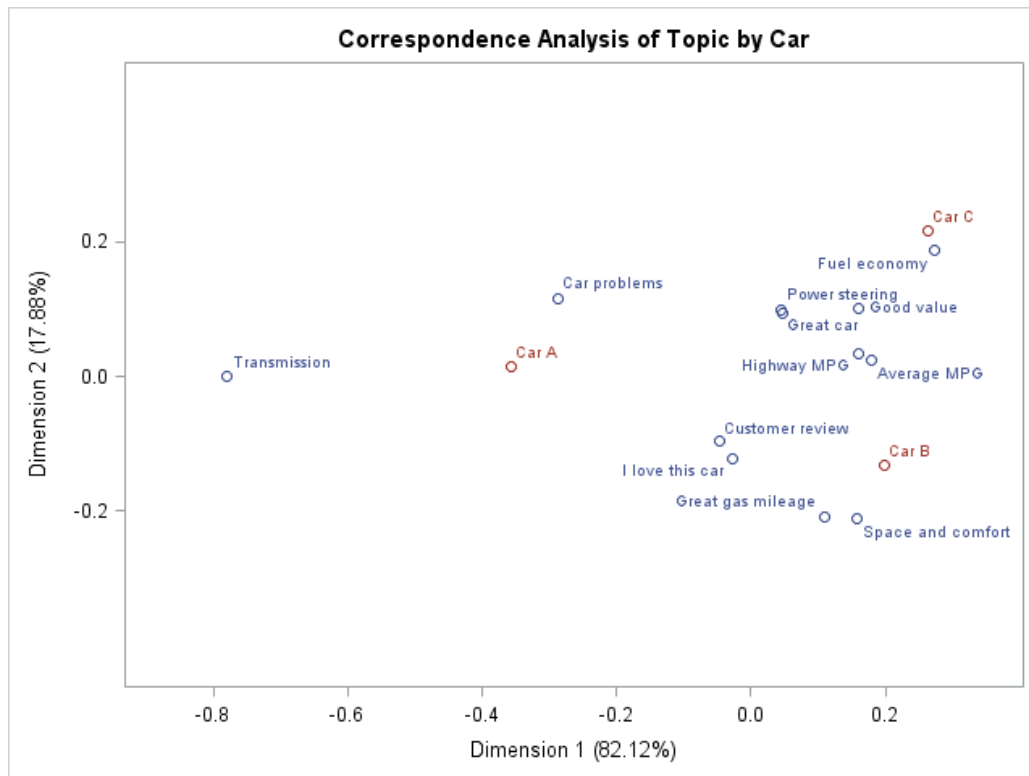
Display 2. Brand Topic Map for the 12-Topic Solution in Case Study 1

CASE STUDY 2: ANNUAL REPORTS OF THREE AUTOMOBILE MANUFACTURERS

Each year, publically traded corporations publish an annual report that contains important information for stock owners, and the investing public. Near the beginning of each annual report, there is typically a letter from the Chief Executive Officer to the shareholders. In these letters, CEOs present their case on why anybody would want to invest in their company. In this case study I examine annual reports from three large companies in the automotive industry, dubbed here Motor Company X, Motor Company Y, and Motor Company Z, for my personal use as a potential investor. The questions I am trying to answer are:

- What high-level concepts are CEOs from the three companies trying to communicate to me, as a potential investor, in their annual report letters?

- How can I visualize such concepts in a way that allows me to compare the three companies against each other, and also examine their dynamic behavior over time?

Data collection covers the years 2003-2014. For the purposes of this case study, from each annual report I read the CEO's letter and took notes in plain text, organized into individual paragraphs. In doing so, I may have introduced a subjectivity bias. However, such bias should be mitigated by the fact that I followed the same procedure for each one of the three document sources. Table 11 summarizes the collected data, listing the three automobile companies, the total number of annual reports available for each one of them with reference to the observation period, the number of different Chief Executive Officers (CEOs) that worked for them, and the corresponding total number of paragraphs of unstructured text generated from the CEO letters included with the annual reports.

Company	Annual reports	CEO count	Total paragraphs in my notes on CEO letters
Motor Company X	12	3	297
Motor Company Y	9*	3	366
Motor Company Z	12	2	105

Table 11. Data Collection Summary for Case Study 2. (*) Motor Company Y did not publish annual reports in years 2007, 2008, and 2009

TOPIC EXTRACTION IN SAS TEXT MINER

Following very similar steps to those in Case Study 1, we now perform topic extraction in SAS Text Miner. The analysis diagram is very similar to the diagram shown on Display 1. The scree plot of eigenvalues, produced outside Text Miner, is shown in Figure 3. There seems to be an elbow point at $k = 13$, suggesting 13 as a potential number of topics to be extracted. Also, a very steep step at $k = 3$ indicates that three may be an appropriate number of very high-level topics. We continue the analysis by producing both solutions, summarized in Table 12 and Table 13.
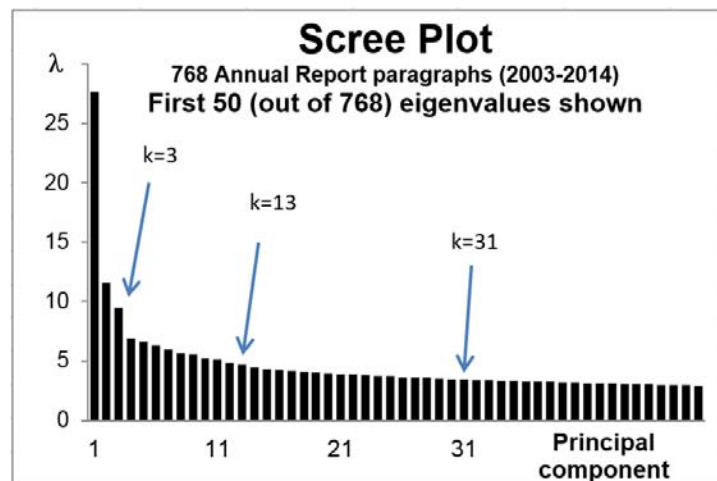

Figure 3. Scree Plot of Eigenvalues Produced by Singular Value Decomposition

Topic ID	Topic Label	Topic Terms	# Docs
1	Cost reduction	unitedstates,+reduce,+cost,+improve,+retiree	86
2	Net income	+billion,net,income,net income,+loss	75
3	Sales	+sale,million,+unit,percent,growth	81
4	Fuel efficiency	+vehicle,+fuel,+technology,+hybrid,fuel economy	79
5	World markets	america,europe,north,africa,asia	86
6	What customers want	+customer,+want,balance,sheet,balance sheet	49
7	Shares & dividends	+share,+prefer,+dividend,ordinary,+ordinary share	50
8	Volunteers & community	+employee,+community,+retiree,+volunteer,+work	44
9	Best-selling brand	+car,+truck,+brand,best-selling,unitedstates	97
10	Growth	growth,+continue,profitable,progress,+company	103
11	Building a better world	+good,+world,+business,+great,+build	71
12	Pre-tax profits	+profit,pre-tax,+operate,+exclude,+special item	62
13	Good product	global,+product,+good,+development,+economy	99

Table 12. Extracted Topics and Corresponding Topic Labels for the 13-Topic Solution

Topic ID	Topic Label	Topic Terms	# Docs
1	Profits and cash	+billion,+profit,cash,automotive,+operate	78
2	Building a good product	+product,+good,+customer,+world,+build	111
3	Sales and market share	+sale,+market,+share,percent,million	102

Table 13. Extracted Topics and Corresponding Topic Labels for the Three-Topic Solution

CORRESPONDENCE ANALYSIS

As with case study 1, the results of the topic extraction performed by the Text Topic node in Text Miner were saved with the help of a Save Data node, and the data set was converted to the long categorical data format as shown in Table 14. Then, assuming that the data file was saved in folder d:\temp as ReportTopics.csv, PROC CORRESP was used once again to run correspondence analysis:

```
proc import out= work.topics
    datafile= "d:\temp\ReportTopics.csv"
    dbms=csv replace;
    getnames=Yes;
proc corresp all data=topics outc=Corr;
    tables Topic, CEO;
run;
```

Topic	Document	ParagraphID	Company	CEO	Year
Profits & cash	2	P0002	X	CEO X1	2003
Profits & cash	3	P0003	X	CEO X1	2003
Profits & cash	6	P0006	X	CEO X1	2003

Table 14. Long Categorical Data Format after Extracting 3 Topics

The resulting Brand Topic Map is shown in Display 3. The contingency table has now a dimensionality of three topics by eight CEOs. Since the CEOs are ordered in time within each company, the map allows for the examination of the dynamic behavior in the way CEOs position their discourse in the two-dimensional projection space. Since the topics are now three, the map accounts again for 100% of the variance in the contingency table. Interestingly, the three CEOs in company X took turns in repositioning their discourse from a focus on *building a good product* to a focus on *sales and market share*. In the meanwhile, the change of CEO in company Z was accompanied with a discourse migration away from *profits and cash* and closer to *sales and market share*. This concludes the presentation of Case Study 2.
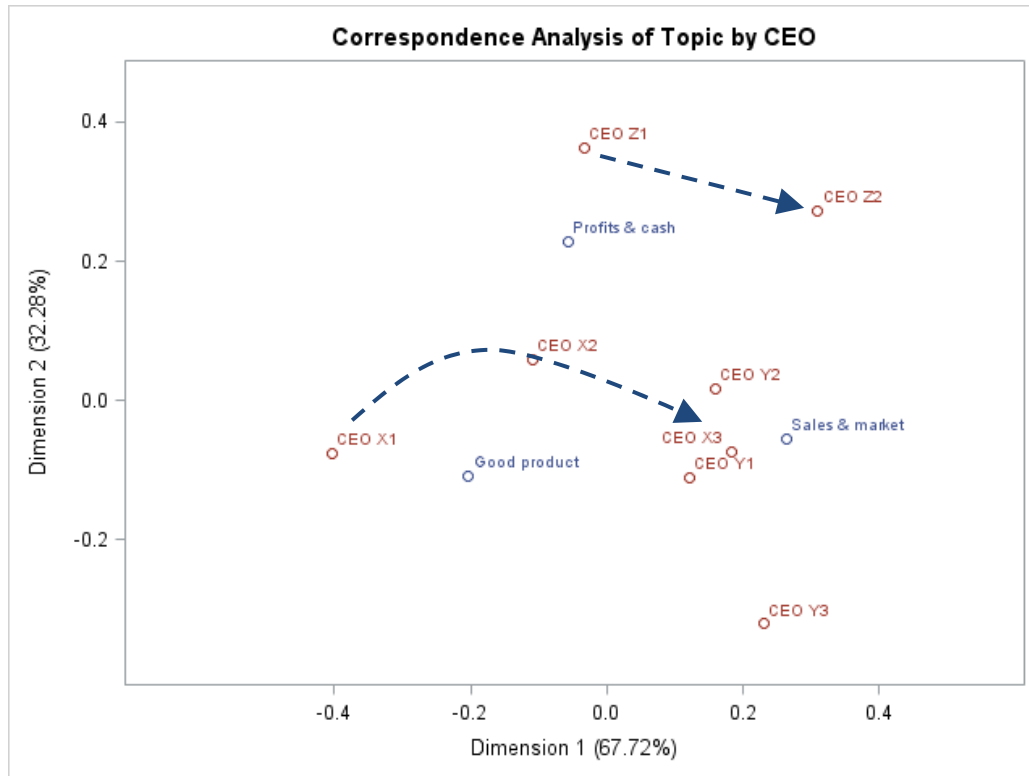
Display 3. Brand Topic Map for the 3-Topic Solution in Case Study 2

DISCUSSION

Referring to the Brand Topic Maps in Display 2 and Display 3, we observe that the maps can help as make associations between Brands and Topics, not only in a static manner, as was done in Case Study 1, but also in a dynamic manner, as was done in Case Study 2. Figure 4 presents a Brand Topic Map abstraction depicting the repositioning of Brand A from a position mostly associated with Topic 1, to a position mostly associated with Topic 3. Brand Topic Maps can be a useful tool in strategic marketing. They can help the strategist plan for conceptual change and the analyst tell a nice story (and show it on the map.) This analytic and visualization approach can be extended to work with associations between topics and a variety of actors that declare the origin or the purpose of the documents in the data set. For example, if the documents are news stories, the actors can be different news sources. If the documents describe leadership initiatives, the actors can be leaders in tan organizational arena or a community. If the documents describe customer reaction, the actors can be products, services, or promotions. For some additional discussion on the use of correspondence analysis maps as a visualization tool for concept spaces, see Evangelopoulos (2016).

As a concluding remark, please note that the purpose of this paper is to demonstrate the extraction of high-level concepts from collections of documents and their visualization using Brand Topic Maps, in a way that facilitates comparisons across different document sources. Note that the purpose is neither to actually compare existing brands, nor to make any claims regarding existing brands. The main goal of this paper is the demonstration of a methodological approach. In order to add a touch of realism, the two case studies presented here reflect on my experiences from researching three existing companies from the automobile industry for personal use as a potential investor or potential customer. Readers should refrain from drawing any conclusions on the case study companies, their executives, their products, their services, or their customers. However, readers are encouraged to set out on their own studies, following steps similar to those presented in this paper.
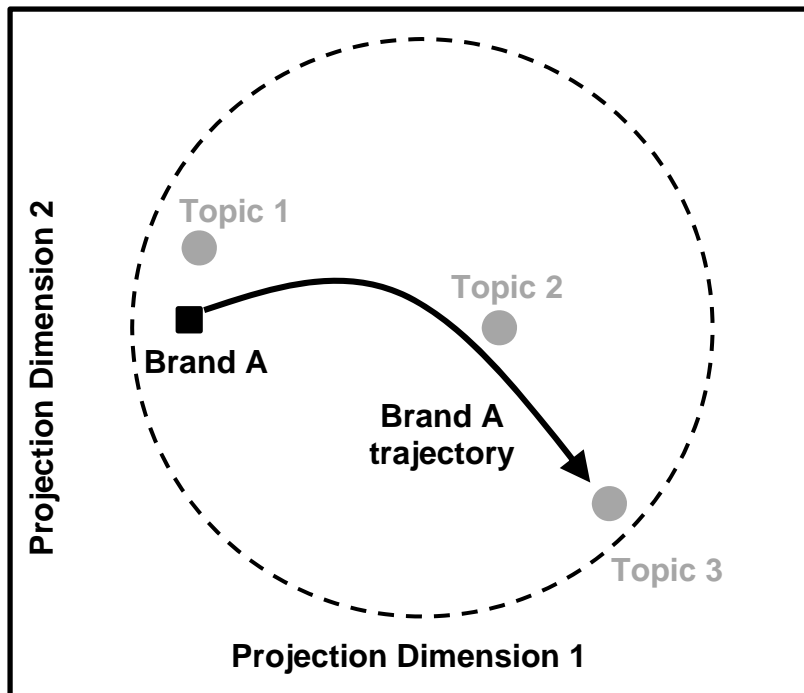
Figure 4. Layout of a Brand Topic Map

CONCLUSION

The two case studies presented in this paper extract high-level concepts from collections of documents and visualize them using Brand Topic Maps. Overall, the two case studies provide support for the following conclusions:

- SAS Enterprise Miner is an effective tool in extracting high-level concepts from collections of unstructured text documents.

- Extracting a small number of topics does not reproduce the first few topics in a topic solution that includes a larger number of topics. Rather, topics are recast and concepts are re-organized at a higher level of semantic abstraction.

- Correspondence analysis provides a way to produce informative output that facilitates comparison among document sources that can be linked to different brands, providers, leaders, products, services, or time periods

- Caution should be exercised in mixing document sources from different cultural or language use background in the same analysis

REFERENCES

Benzecri, J. P. 1973. *L' Analysé des Données, vol. 2, L' Analysé des Correspondances.* Paris: Dunod.

Chakraborty, G., M. Pagolu and S. Garla. 2013. *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®.* Cary, NC: SAS Institute.

Evangelopoulos, N. 2016. "Thematic orientation of the ISJ within a semantic space of IS research." *Information Systems Journal*, 26:39–46.

Evangelopoulos, N., X. Zhang and V. Prybutok. 2012. "Latent semantic analysis: five methodological recommendations. " *European Journal of Information Systems*, 21:70–86.

Evangelopoulos, N., T. Ashton, K. Winson-Geideman and S. Roulac. 2015. "Latent Semantic Analysis and Real Estate Research: Methods and Applications. " *Journal of Real Estate Literature*, 23(2):355–380.

Hirschfeld, H. O. 1935. "A connection between correlation and contingency. " *Cambridge Philosophical Society Proceedings (Math. Proc.)*, 31:520–524.

Kelley Blue Book Co. 2016. "Kelley Blue Book." Accessed March 18, 2016. http://www.kbb.com/.

Whitlark, D. and S. Smith. 2001. "Using correspondence analysis to map relationships. " *Marketing Research*, 13(3):22–27.

RECOMMENDED READING

- *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS®*

- *Getting Started with SAS® Text Miner 13.2*

- *Getting Started with SAS® Enterprise Miner™ 14.1*

- *SAS/STAT® 13.1 User's Guide: The CORRESP Procedure*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Nick Evangelopoulos
University of North Texas
940-565-3056
Nick.Evangelopoulos@unt.edu
http://www.cob.unt.edu/itds/faculty/evangelopoulos/evangelopoulos.htm