Paper 099-2013

Replace Manual Coding of Customer Survey Comments with Text Mining: A Story of Discovery with Text as Data in the Public Sector

Jared Prins, Alberta Tourism, Parks and Recreation

ABSTRACT

A common approach to analyzing open-ended customer survey data is to manually assign codes to text observations. Basic descriptive statistics of the codes are then calculated. Subsequent reporting is an attempt to explain customer opinions numerically. While this approach provides numbers and percentages, it offers little insight. In fact, this method is tedious and time-consuming and can even misinform decision makers. As part of the Alberta Government's continual efforts to improve its responsiveness to the public, the Alberta Parks division transitioned from manual categorization of customer comments to a more automated method that uses SAS® Text Miner™. This switch allows for faster analysis of unstructured data, and results become more reliable through the consistent application of text mining.

INTRODUCTION

Alberta's provincial parks system protects more than 27,600 square kilometres, or approximately 4.2 per cent of the province. This area is larger than Hawaii.

The 478 sites in the Alberta Parks system offer a rich diversity of opportunities and uses. Some parks are designed for recreation, but many others support both conservation and recreation activities. There are 250 campgrounds with nearly 14,000 campsites in the Alberta Parks system, utilized by approximately 1.5 million campers annually.

Since 2002, a province-wide Camper Satisfaction survey has been conducted at parks that offer camping. The purpose of this survey is to gain an understanding of visitors' satisfaction with services, facilities, opportunities and overall satisfaction for evaluating program performance.

Text mining was introduced in 2008 to analyze the unstructured data from customer comments on the survey.

"THERE HAS GOT TO BE A BETTER WAY"

When faced with manually summarizing and making sense of up to 2,000 observations of open-ended customer comments (i.e., unstructured data or qualitative data), a common response is "There has got to be a better way". Typically the approach to analyze unstructured data is to manually read each record and assign category codes (i.e., coding). This is a labour intensive and frustrating task. It begs the question if there is software to treat text as data.

SAS Text Miner[1] can automate text categorization. It not only replaces the task of manual text categorization but also supports insight discovery through predictive and descriptive models. Equally important, the SAS solution improves data driven decision-making.

SURVEY DETAILS

OVERVIEW

An annual "Camper Satisfaction" survey is randomly distributed to campers during their stay at select parks between June 1 and early September. It is a 3-panel, brochure style, paper-based questionnaire. *A sample of the survey instrument is available online by downloading the 2008 Camper Satisfaction survey report. Follow the link provided in the References section* [1].

The survey collects both quantitative and qualitative data. The quantitative portions are "fill in the bubble" responses and "simple" hand-written responses (Figure 1).

[1] SAS Text Miner is an add-on to Enterprise Miner™, the data mining solution from SAS.

Figure 1. Examples of "fill in the bubble" and simple hand written responses

In 2008, there were 2,027 surveys completed of which 1,118 surveys included comments. Completed surveys are returned to staff, dropped off at any check-in station, self-registration vault or visitor comment box, or by mail. All surveys eventually end up in the head office for processing. Processing involves converting the paper surveys into an electronic format and preparing the data for analysis (i.e., data cleaning and conversion to a SAS data set).

The survey is an electronically scannable form. The quantitative questions on the survey are "read" using Optical Mark Recognition (OMR) and Optical Character Recognition (OCR) scanning technology.

Using electronically scannable forms supports rapid creation of the electronic data set and minimizes potential technical non-sampling errors that might occur during recording, coding, tabulating, or analyzing data. Over the years the conversion of quantitative data to electronic has been streamlined. The challenge has always been converting open-ended comments (Figure 2) to an electronic format (i.e., creating the corpus[2] data set) and manually assigning codes to those comments.

Figure 2. Example of an open-ended comment area

CREATING THE CORPUS

The open-ended customer comments are manually typed into an electronic spreadsheet (e.g., Microsoft Excel). The spreadsheet is used for its ease of use with data entry and is simply an interim step. In the past, an attempt of OCR on handwritten comments was found unreliable due to extreme variation in handwriting. In 2010, speech recognition software was implemented. A quality microphone, aptitude for learning to use the software and clear voice are recommended. Otherwise, there is greater accuracy and ease of use through manual transcription.

Manual transcription will always be a challenge for paper based surveys. Although the internet is becoming a more viable approach for surveying (i.e., comments are provided electronically), it carries its own set of shortcomings for researchers.

[2] A corpus is a collection of written text (also referred to as documents). Examples of a document include an email, letter, survey comment or a database record containing text.

Table 1 highlights a sample of transcribed comments received in response to the survey question: "*What could we have done to make your visit better?*"

Table 1: Examples of verbatim customer comments

Serial number	Comments
1	Come by more often with wood.
2	Provide sites with electricity.
3	No map on directions to amphitheatre for nightly program.
4	Put a flush toilet system in place.
5	The campground was very clean and well layed out. Your staff was very friendly, and courteous. We enjoyed our visit to Canada and this campground.
6	Free firewood.
7	Firewood included in price of camp site. Contractor could deliver free fire wood too.
8	Would like to see some sort of playground for kids.
9	Signage (directional) is important to us.

Once all the comments have been typed (or dictated) into a spreadsheet, it is converted to a SAS data set. The comments range in character length, so care is taken to ensure comment text is not accidentally truncated.

The TEXTSIZE= option of the IMPORT procedure avoids truncation. The length of the variable will be set to the longest comment present in the data. Although TEXTSIZE is set to 3000, the resulting variable was a length of $1052, corresponding to the length of the longest comment in the 2008 data.

```
PROC IMPORT OUT=qual_data
    DATAFILE= "C:\Comments2008.xls" REPLACE;
    RANGE="Comments$"; /*Sheet name*/
    GETNAMES=YES; /*Variable name in header row*/
    SCANTEXT=YES; /*See SAS Usage Note:13386 before using this option*/
    TEXTSIZE=3000; /*Set a generous length value*/
RUN;
```

This process and the technology described so far expedite data preparation. The manual text categorization method described is anything but expedient.

MANUALLY CATEGORIZING COMMENTS (THE OLD METHOD)

DEVELOPING THE TAXONOMY[3]

The majority of our business taxonomy was created during the first iteration of the survey. This taxonomy was used to classify comments. Refinements or minor additions to the taxonomy were made each year. For example, with the rise of internet connected mobile devices, so too did requests for Wi-Fi or cell phone coverage in and around parks. The analyst decides when enough occurences require a new category. By 2008, there were 28 general categories such as:

Washrooms	*Information Services*	*Policy*	*Firewood*
Pest Control	*Trails*	*Roads*	*Playgrounds*
Showers	*Reservation System*	*Fee / Value*	*Security*
Camping Preferences	*Fishing*	*Noise*	

[3] Taxonomy, in the context of business, is a defined catalog of classifications pertaining to the business.

Within each general category are sub-categories. An extract of the Firewood and Positive Comments general categories with sub-categories are presented in Table 2. In total, there were 187 sub-categories!

Table 2: Coded general categories (firewood, positive comments) with sub-categories

3	**Firewood**
a	Too expensive
b	Should be free
c	Better access (location, timing of wood access)
d	Quantity (not enough / no wood)
e	Poor quality (wet, too long, type)
f	Delivery services needed
g	Need chopping blocks
h	Other
i	Should be included in fee
j	Firewood shelter needed / upgraded

100	**Positive Comments**
a	General (e.g., nice time, nothing, enjoyed stay
b	Nice area
c	Facilities – campground / campsites
d	Good staff / host / operator
e	Wood free / good quality
f	Other
g	Clean / well run
h	Good amphitheatre programs
i	Good road / facility improvements

To support analysis, each general category and sub-category was assigned an identifier as in Table 2. The combination of a general category with one of its sub-categories produces a unique identifier (i.e., code) used to represent text coments.

ASSIGNING CODES TO COMMENTS

Manual categorization (i.e., assigning codes) of unstructured data can be described with one word: Painful. Anyone who has endured the process might wish to skip over this section for health reasons.

The analyst reads all comments and assigns a code for each distinct phrase. Consider the following two comments in Table 3. The first respondent has written 4 phrases. The second respondent expresses only 2 responses. Each phrase is assigned a single code.

Table 3: Example process of assigning codes to comments

Respondent #1	"We had a great time and enjoyed our stay. But firewood was too expensive and it was wet."	We had a great time = 100a and enjoyed our stay = 100a But firewood was too expensive = 3a and it was wet = 3e
Respondent #2	"Firewood is too expensive, it should be free."	Firewood is too expensive = 3a it should be free = 3b

It is entirely valid that respondent #1 is assigned a duplicate code (i.e., 100a). The goal of assigning codes is to match what is written with as little human interpretation as possible. An example of how these codes look in a spreadsheet is presented in Figure 3 (actual comments removed to fit the page).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		code1	code2	code3	code4	code5	code6	code7	code8	code9	code10	code11	code12	code13
2	comment text	16e	16e	100d	100f									
3	comment text	7a	5c	3d	5h	3d	7a							
4	comment text	4a												
5	comment text	12c	12c	16a	24f									
6	comment text	12c	12d	3a	3d	21d	16a							
7	comment text	3i	3i	3e	3a									
8	comment text	100d	11a	3a										
9	comment text	4a	24f	4a	4a	3a	10e	19h	2d	19h	100r	11c	11e	24f
10	comment text	100a	100b	100a	100c									
11	comment text	4a	10e	3b										

Figure 3: Coded comments, highlighting the longest comment

Thirteen new variables (code1 - code13) are present in the final data set. The number of codeX variables correspond to the comment with the highest count of phrases. The first comment contained only 4 phrases, the second comment contained 6 phrases and so forth.

Once all comments are represented as codes, the spreadsheet is converted to a SAS data set and ready for analysis.

ANALYSIS OF CODED COMMENTS

The TABULATE procedure was used to generate frequencies for general categories and sub-categories (Figure 4).

General Category	Sub-Category	# of Comments	% of Category	% of All Comments	% of ALL Surveys Represented
Firewood	Too expensive	96	30.1	3.6	8.6
	Firewood Quantity (not enough/no wood)	61	19.1	2.3	5.5
	Poor Quality (too long, wet)	48	15.0	1.8	4.3
	Poor Access (location, timing)	47	14.7	1.8	4.2
	Should be free	40	12.5	1.5	3.6
	Firewood Delivery Needed and other	13	4.1	0.5	1.2
	Firewood Should be Included in Fees	12	3.8	0.4	1.1
	Firewood Shelter Needed/Upgraded	2	0.6	0.1	0.2
	Subtotal	319	100.0	11.9	28.5
Hook-ups/Dump stations/Water	Additional power campsites	86	34.8	3.2	7.7
	Full Power-Water-Sewer Hook-ups Needed	31	12.6	1.2	2.8
	Other (specific amperage, water filling station needed)	26	10.5	1.0	2.3
	More Taps / Water Locations	24	9.7	0.9	2.1
	Poor Drinking Water Quality / Need Potable Water	21	8.5	0.8	1.9
	Install power campsites	20	8.1	0.7	1.8
	Sewage Dump-stations Needed / Dirty / Full	18	7.3	0.7	1.6
	Water Hook-ups Needed	11	4.5	0.4	1.0
	Running Water Needed (not washroom related)	10	4.0	0.4	0.9
	Subtotal	247	100.0	9.2	22.1

Figure 4: PROC TABULATE output of coded comments

Only a portion of the PROC TABULATE output is presented in Figure 4 because the large taxonomy of 187 categories produces pages of output. *The full table is available online by downloading the 2008 Camper Satisfaction survey report. Follow the link provided in the References section. Sample PROC TABULATE code is included in the Appendix.*

Roughly 2,000 comments takes one full-time employee approximately 3 to 4 weeks to convert to electronic text and manually assign codes. The time it takes for quantitative and qualitative data preparation and analysis leaves little

time to pursue other forms of statistical exploration or modeling. SAS Text Miner recovers this time by replacing the tedious task of coding comments.

TEXT MINING (NEW METHOD)

In 2008, SAS Text Miner replaced the manual coding approach.

REPLICATING THE TAXONOMY

The Text Parsing node and Text Filter node will automatically generate taxonomy similar to the business defined taxonomy. These two nodes are the first used in a typical text mining process flow within SAS Enterprise Miner (Figure 5).

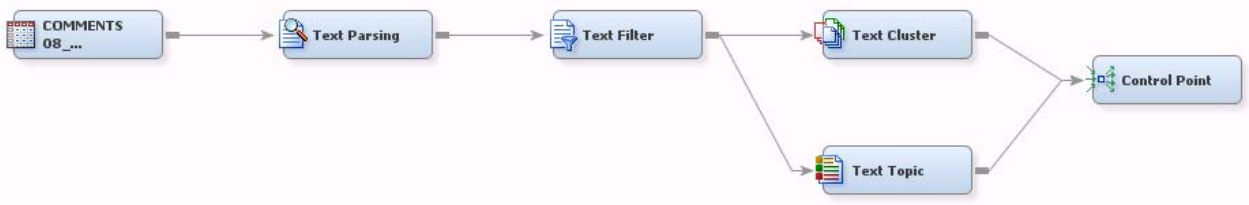


Figure 5: A typical text mining process flow in Enterprise Miner

(Tip: A "Control Point" node will run all nodes preceding it).

The first step is importing the comment data into Enterprise Miner where it will assign a default role to each variable. These defaults often suffice and do not need to be changed by the analyst. Ensure the comment variable is assigned a role of "Text" (Figure 6).

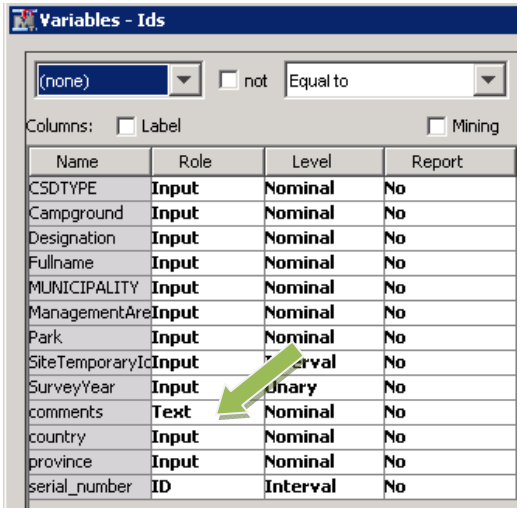


Figure 6: Importing the survey comment data as "Text" role

Tip: The Text role will be assigned to all text (character) variables present in the data. By default, SAS Text Miner will use the largest of these variables in a text mining process unless otherwise selected by the analyst. Avoid processing the wrong variable by setting the role of unwanted variables as "Rejected". Later, when variables are wanted, change the role by using a Metadata node. Rejecting a variable excludes it from node processing. It is retained in the data.

Once the data is imported, a Text Parsing node is used to generate frequencies of terms within documents. This creates a "term-by-document matrix" which is well described by Albright [2]. Consider the term-by-document matrix an extremely large data set with many observations and variables.

The Text Parsing node and Text Filter node have features that help condense the term-by-document matrix. Not unlike standard data cleansing techniques applied to quantitative data, both of these nodes contain features to clean text data (i.e., reduce noise). A given corpus can react differently to these node options. It is important to experiment to find optimal settings.

In the Text Parsing node property sheet (Figure 7), the following features help reduce noise in the survey data:

- "Different Parts of Speech" will treat nouns, verbs, adverbs separately. For example, this will distinguish homonyms – terms with the same spelling and pronunciation, but different meanings. For example, River **bank** vs. Institutional **bank**.
- "Noun Groups" will treat frequent term sequences as a single term, such as the bigrams[4] "Provincial Park".
- "Stem Terms" will consider terms as their root form (stems, stemmed, stemmming all become stem).
- The "Stop List" excludes unwanted terms and a "Start List" includes wanted terms during analysis. Low frequency terms or high frequency terms (such as "and", "the", "is") are candidates for the stop list.
- "Entity Extraction" identifies tokens[5] such as phone numbers, names, and dates.

Entity extraction is a valuable feature. Although personal contact information is not asked of customers, some still provide it in survey comment boxes. These records should be identified since we have an obligation to reply. Furthermore, this feature supports removal of personally identifiable information to comply with privacy legislation requirements.

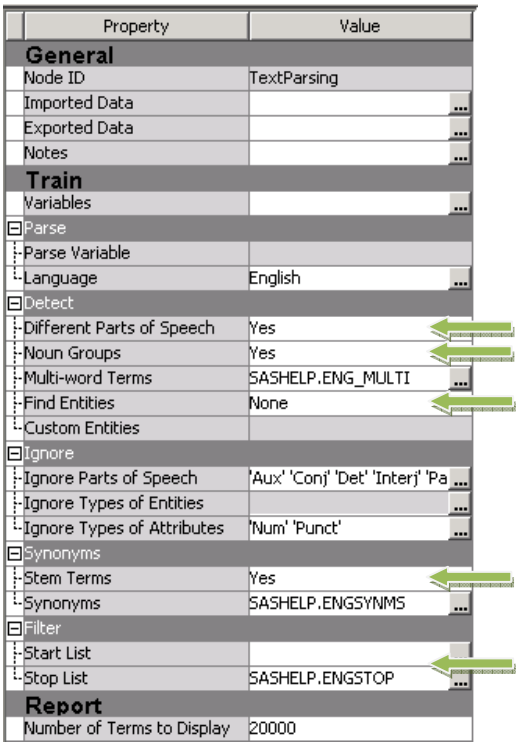


Figure 7: Text Parsing node properties

A Text Filter node normally follows parsing. The goal of this node is data cleaning, term exploration and querying, but more importantly, to weight terms based on their importance in the corpus. For example, frequent terms are not helpful in discriminating documents and will receive a very low weighting. These weights prepare the data for more

[4] A bigram is every sequence of adjacent elements in a string of tokens.

[5] A token is a contiguous sequence of characters.

efficient dimension reduction in subsequent nodes. Dimension reduction is a way of reducing noise while keeping enough data as to represent the original data. SAS Text Miner uses a mathematical technique called Singular Value Decomposition (SVD) [2].

Common settings to change in the Text Filter node property sheet are shown in Figure 8. Recall in Figure 4 that code frequencies are calculated *within* a general category and *across* all surveys. SAS Text Miner takes a similar approach:

- Term Weighting (Global Weight) accounts for how a term is spread across the corpus. A number of term weighting methods are available such as Entropy, Inverse Document Frequency, Mutual Information and None. For the survey comments, Inverse Document Frequency is preferred, but the default Entropy weighting method also produces acceptable results.
- Frequency Weighting (Local Weight) accounts for how terms relate within a document. A number of frequency weights are available such as Log, Binary and None. The survey comment data responds well to Log weighting and in some cases, None.

Discovering which weighting methods to use involves some trial and error.

Other features of the node include checking for spelling and the Filter Viewer. Although spell checking is computer CPU intensive, it is recommended for user-generated data. Examples of such data include online surveys or social media content where misspellings are often the rule rather than the exception.

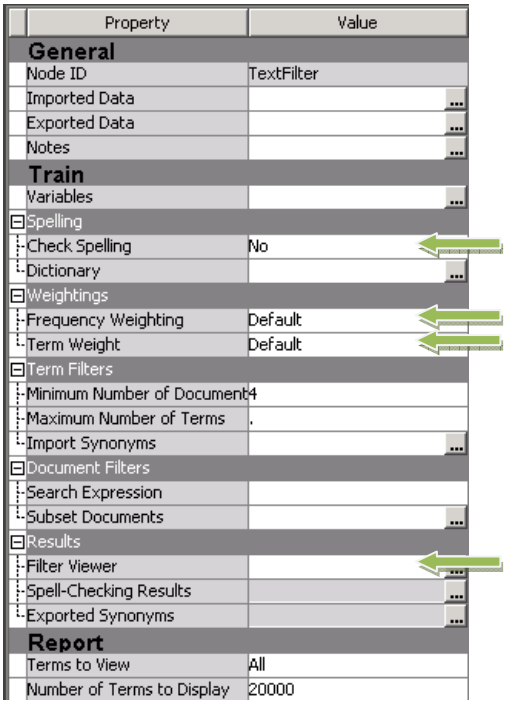


Figure 8: Text Filter node properties

In many ways, the results of the Text Filter node are where the taxonomy begins to reveal itself. The Filter Viewer (accessed from the node's property sheet) lists terms and their frequencies (Figure 9). The top terms are similar to the original general categories.

TERM	FREQ	# DOCS	KEEP	WEIGHT
campsite	1098	606	☑	2.471
campground	970	598	☑	2.49
visitor	810	486	☑	2.789
information	485	300	☑	3.485
noise	336	233	☑	3.85
day	295	208	☑	4.014
fee	425	202	☑	4.056
time	246	197	☑	4.092
night	250	189	☑	4.152
great	236	185	☑	4.183
staff	208	164	☑	4.357
reservation	309	158	☑	4.41
washroom	212	156	☑	4.429
showerhouse	209	141	☑	4.575

Figure 9: The Filter Viewer stop list and weights

Sub-categories are discovered with a feature called “Concept Linking”. Use the pointer to right click on any term in the Filter Viewer list and select “View Concept Links”. Algorithms are used to define the strength of association between terms which form the concept link (Figure 10) [3].

The term “Firewood”, for example, reveals sub-categories such as “free firewood”, “include firewood [with fee]”, and “dry firewood”. These concepts as seen in Figure 10 are strikingly similar to the sub-categories of Table 2. Hovering the pointer over these related terms will show a tooltip for which the second number represents the total number of documents in the corpus containing the term.

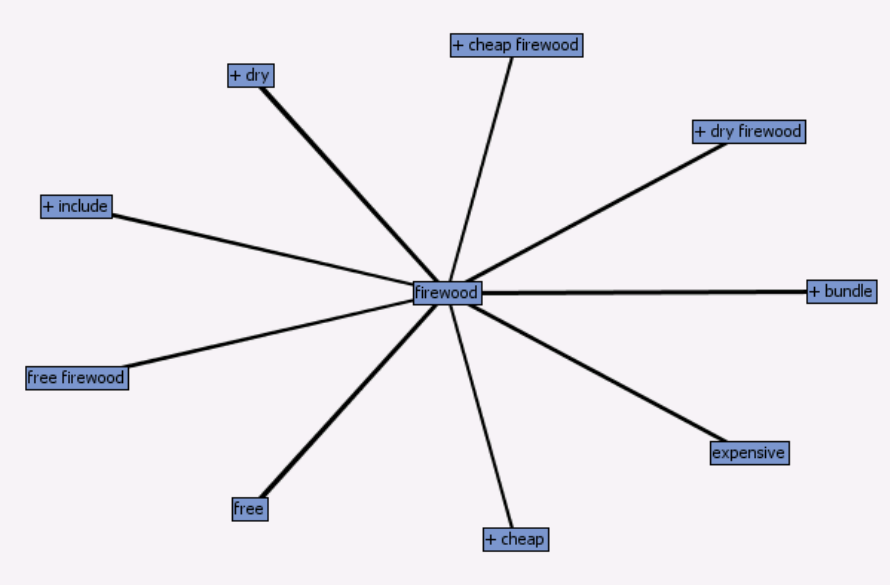


Figure 10: Concept Linking for term “Firewood”

Continuing with the example of firewood, the agreement between text mining and manual coding can be explored. Figure 11 shows the frequency of the terms “firewood” (234) and “wood” (100) for 334 total responses (across 290 documents). This calculated to roughly 25% of all surveys. In 2008, the manual coding counted 319 comments related to firewood, representing 28.5% of all surveys (Figure 12).

Terms

	TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE
⊞	site	347	248	☑	0.255	Noun
	firewood	234	209	☑	0.263	Noun
⊞	campground	248	203	☑	0.273	Noun
⊞	good	231	202	☑	0.269	Adj
⊞	park	242	191	☑	0.282	Noun
⊞	shower	219	190	☑	0.278	Noun
⊞	camp	144	124	☑	0.338	Verb
⊞	campsite	143	121	☑	0.342	Noun
⊞	area	141	115	☑	0.352	Noun
⊞	nice	119	112	☑	0.348	Adj
	power	127	107	☑	0.361	Noun
⊞	facility	114	105	☑	0.358	Noun
⊞	great	114	102	☑	0.363	Adj
⊞	washroom	113	95	☑	0.376	Noun
⊞	bathroom	104	92	☑	0.378	Noun
⊞	wood	100	81	☑	0.399	Noun
	nice	85	81	☑	0.392	Adv

Figure 11: Frequencies for terms "firewood" and "wood"

General Category	Sub-Category	# of Comments	% of Category	% of All Comments	% of ALL Surveys Represented
Firewood	Too expensive	96	30.1	3.6	8.6
	Firewood Quantity (not enough/no wood)	61	19.1	2.3	5.5
	Poor Quality (too long, wet)	48	15.0	1.8	4.3
	Poor Access (location, timing)	47	14.7	1.8	4.2
	Should be free	40	12.5	1.5	3.6
	Firewood Delivery Needed and other	13	4.1	0.5	1.2
	Firewood Should be Included in Fees	12	3.8	0.4	1.1
	Firewood Shelter Needed/Upgraded	2	0.6	0.1	0.2
	Subtotal	319	100.0	11.9	28.5
Hook-ups/Dump stations/Water	Additional power campsites	86	34.8	3.2	7.7
	Full Power-Water-Sewer Hook-ups Needed	31	12.6	1.2	2.8
	Other (specific amperage, water filling station needed)	26	10.5	1.0	2.3
	More Taps / Water Locations	24	9.7	0.9	2.1
	Poor Drinking Water Quality / Need Potable Water	21	8.5	0.8	1.9
	Install power campsites	20	8.1	0.7	1.8
	Sewage Dump-stations Needed / Dirty / Full	18	7.3	0.7	1.6
	Water Hook-ups Needed	11	4.5	0.4	1.0
	Running Water Needed (not washroom related)	10	4.0	0.4	0.9
	Subtotal	247	100.0	9.2	22.1

Figure 12: Manual coding results for general category "firewood"

The similarity of these results validates that SAS Text Miner is working as anticipated. Any discrepancy is likely due to human error in manual coding or data quality issues which affect text mining.

To be fair, firewood might not be the best example because even a word count would produce a similar result. The power of the Text Parsing and Text Filter nodes are their ability to process natural language[6]. A better example would be the term "Park". This term has multiple meanings. Take for example the following two comments:

"There was nowhere to park my car"

"This is a fantastic park to visit".

A word count cannot differentiate between term meanings.

The ability to detect "parts of speech" (e.g., noun, verb) addresses this issue. This is vital for the comment corpus where respondents use both meanings of the term and use them often. Other corpora sometimes benefit from turning off "parts of speech", thus leaning more toward a "bag of words" approach to text mining. The analyst should experiment with both to see which works best.

With the firewood example, synonyms[7] can be created. Figure 13 shows how a synonym is created in the Filter Viewer – by right clicking with the pointer on the two terms and selecting "Treat as Synonyms".

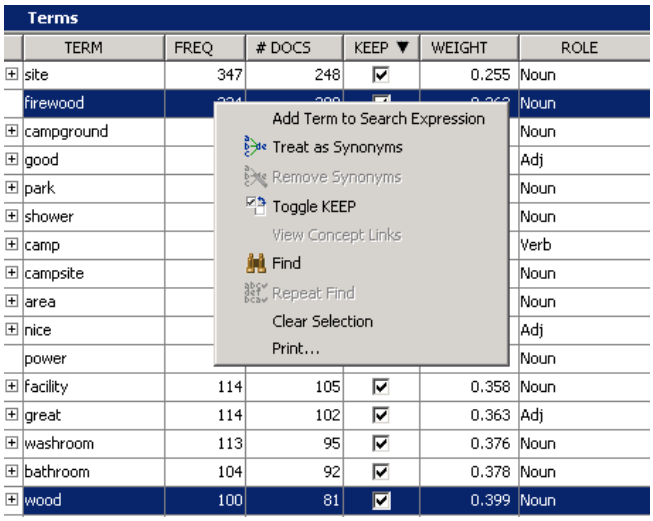


Figure 13: Setting term "wood" as a synonym of "firewood"

Creating synonyms further reduces noise in the data. To illustrate what happens when reducing noise, notice a new term appears (Figure 14) that is different from the original concept link (Figure 10) after setting "wood" as a synonym of "firewood".

[6] A natural language is a human written or spoken language.

[7] Synonyms are different terms with the same meaning. Example: Firewood and Wood.

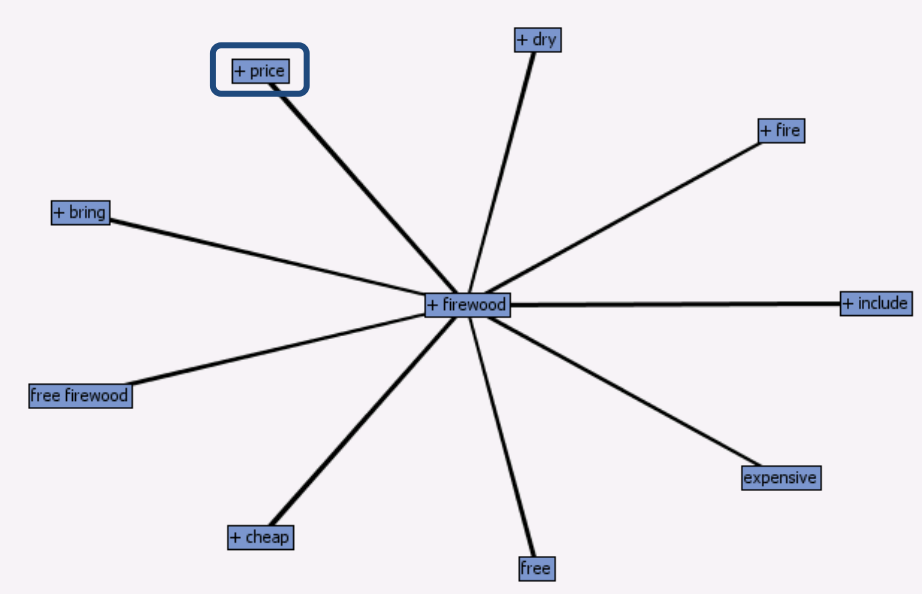


Figure 14: Noise reduced concept link shows new term "price"

Incidentally, "price" could be set as a synonym of "expensive" (seen in the lower right corner of Figure 14). This process of curating synonyms refines data which in turn improves analysis. While synonyms are a powerful feature, it is easy to overuse it (or abuse it). Use the synonym feature where obvious, as in the case of "wood & firewood" or, to name another example, "garbage & trash".

As demonstrated, the Text Parsing and Text Filter node begin to produce similar outputs as the manual coding approach. Additionally, SAS Text Miner avoids two pitfalls of the manual coding method:

- Manual categorization is a subjective exercise. To illustrate with a personal account, when I took over as analyst for the survey program in 2005, my code assigning produced different results than my predecessor – some categories were up to a 20% difference! Manual coding is inevitably a subjective exercise. SAS Text Miner is objective and results can be reproduced from analyst to analyst.
- Manual coding is time consuming. Thousands of comments can take weeks to assign codes. SAS Text Miner can take as little as minutes.

Clearly, SAS Text Miner is worth its *weight* in gold (terrible pun intended)!

Producing numbers and percentages of terms is helpful, but to better mirror the categories in the original taxonomy, SAS Text Miner can summarize data by grouping related terms using the Text Cluster and Text Topic nodes.

CLUSTERS AND TOPICS

After the Text Filter node has run, the next node is a Text Cluster node or Text Topic node (or both!). The Text Cluster and Text Topic nodes use algorithms to group comments into themes.

Since the results of both nodes are similar, it can be challenging to explain the difference between them. SAS Enterprise Miner documentation addresses this.

> *"The **Text Topic** node enables you to explore the document collection by automatically associating terms and documents according to both discovered and user-defined topics. Topics are collections of terms that describe and characterize a main theme or idea...The approach is different from clustering because clustering assigns each document to a unique group while the **Text Topic** node assigns a score for each document and term to each topic. Then thresholds are used to determine if the association is strong enough to consider that the document or term belongs to the topic. As a result, documents and terms may belong to more than one topic or to none at all"* [4].

For the analyst, the Text Cluster node creates a new nominal variable in the data (Figure 15) and each comment is assigned a value that corresponds to a theme (i.e. cluster). For the non-analyst, consider it a "single bucket" approach. The software takes a "pragmatic" view of the comments and identifies the predominant cluster for each comment.

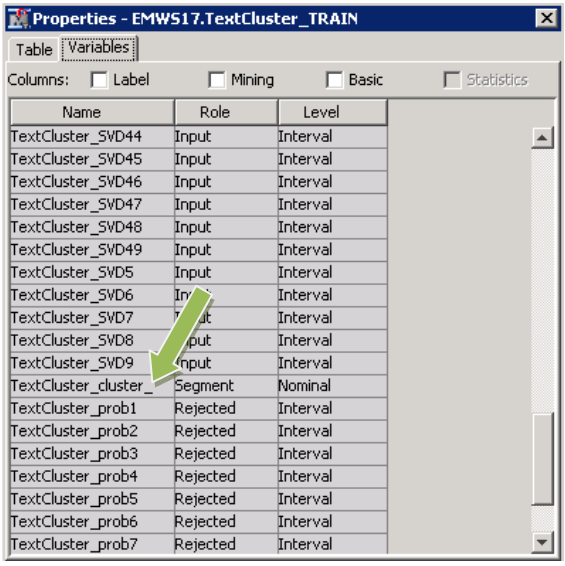


Figure 15: Cluster node produces a new variable

For the analyst, the Text Topic node creates a new binary variable for each topic (Figure 16). For the non-analyst, consider it a "multi bucket" approach. Each comment can fall into multiple topics (including none at all). In Figure 16, the Text Topic node automatically discovered 10 topics.

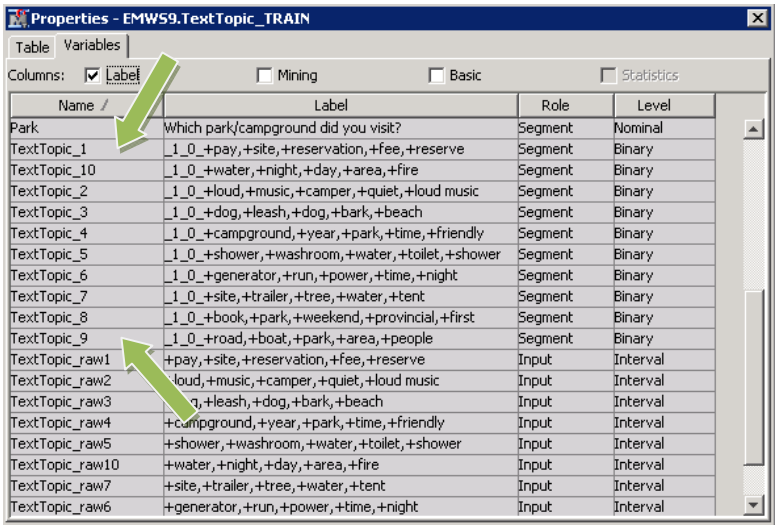


Figure 16: Text Topic node produces multiple new variables

In Figure 16, the label column shows a group of terms that tend to occur together throughout the document collection (e.g., TextTopic_2 →+loud, +music, +camper, +quiet, +loud music). These are SAS Text Miner's descriptions of topics it discovered (clusters are similarly labeled).

Assigning *meaningful* labels to topics and clusters is important because "Clusters are meaningful if they are plausible, and one sign of their plausibility is the ability to write meaningful labels" [3]. If a meaningful label cannot be applied, it

might be a sign that the text mining process needs further refinement or there are challenges with the underlying data.

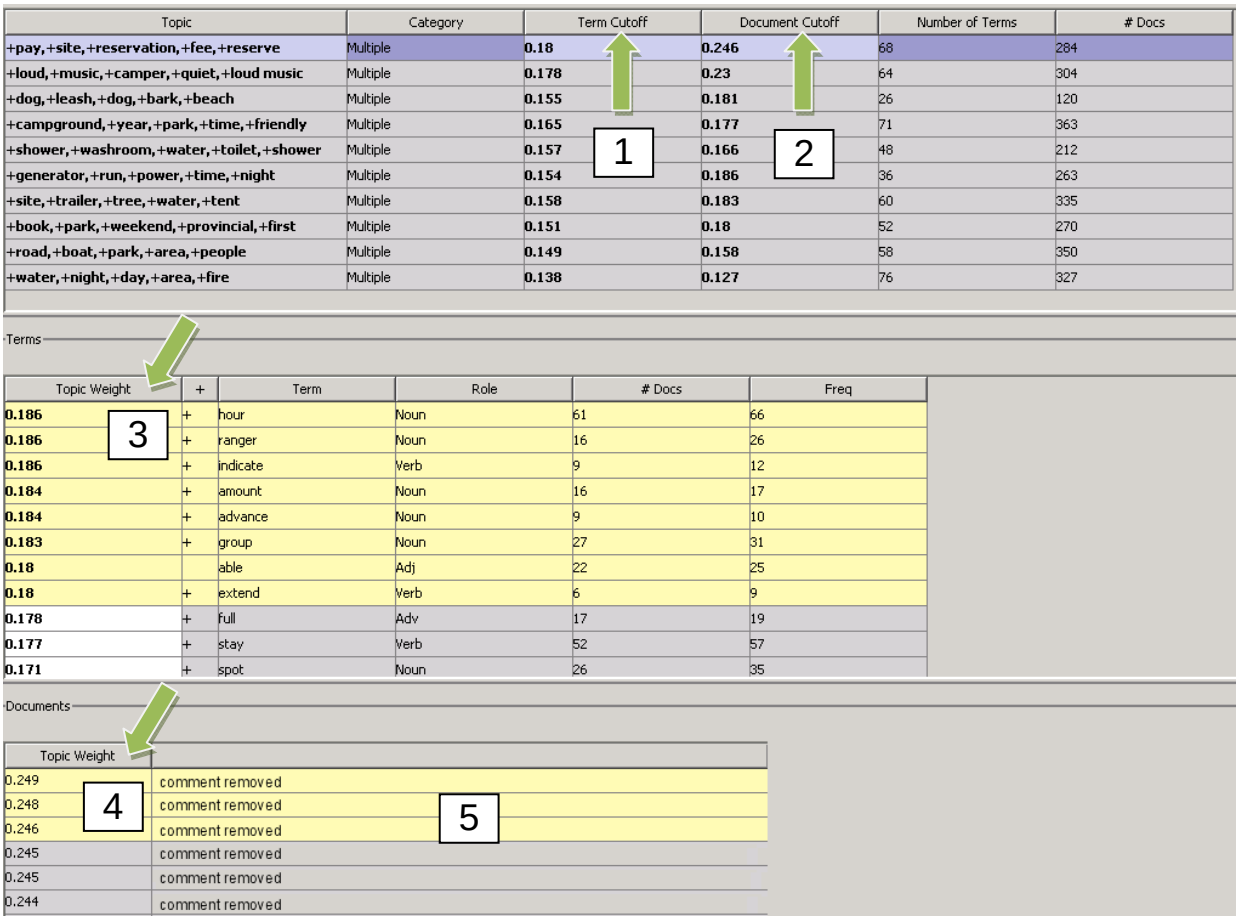
The step of labeling these term groups (Table 4) is performed by the analyst. Though not always necessary for text mining, this is where subject matter expertise is helpful.

Table 4: Applying meaningful labels to topics

Topic ID	Topic Terms	Meaningful Label
1	+pay,+site,+reservation,+fee,+reserve	Reservation fee complaints
2	+loud,+music,+camper,+quiet,+loud music	Loud music at night
3	+dog,+leash,+dog,+bark,+beach	Enforcing dog rules
4	+campground,+year,+park,+time,+friendly	Kudos and positive comments
5	+shower,+washroom,+water,+toilet,+shower	Shower and washroom maintenance
6	+generator,+run,+power,+time,+night	Loud generators at night
7	+site,+trailer,+tree,+water,+tent	Campsite preferences
8	+book,+park,+weekend,+provincial,+first	Campsite booking issues
9	+road,+boat,+park,+area,+people	Road and parking complaints
10	+water,+night,+day,+area,+fire	Water tap access and maintenance

Some themes can be challenging to label meaningfully. In such cases, use the Topic Viewer (accessed from the Text Topic node properties) to explore the raw text (i.e., verbatim comments) associated with themes. This should provide a sense of the overall theme. When doing this, be sure to avoid reading comments that fall below the cutoffs. In Figure 17, ignore terms with topic weights (box 3) that fall below the term cutoff (box 1) and ignore comments with topic weights (boxes 4 and 5) that fall below the document cutoff (box 2). These are shaded grey rather than yellow.

(Note: Figure 17 comments removed for privacy).



Topic	Category	Term Cutoff	Document Cutoff	Number of Terms	# Docs
+pay,+site,+reservation,+fee,+reserve	Multiple	0.18	0.246	68	284
+loud,+music,+camper,+quiet,+loud music	Multiple	0.178	0.23	64	304
+dog,+leash,+dog,+bark,+beach	Multiple	0.155	0.181	26	120
+campground,+year,+park,+time,+friendly	Multiple	0.165	0.177	71	363
+shower,+washroom,+water,+toilet,+shower	Multiple	0.157	0.166	48	212
+generator,+run,+power,+time,+night	Multiple	0.154	0.186	36	263
+site,+trailer,+tree,+water,+tent	Multiple	0.158	0.183	60	335
+book,+park,+weekend,+provincial,+first	Multiple	0.151	0.18	52	270
+road,+boat,+park,+area,+people	Multiple	0.149	0.158	58	350
+water,+night,+day,+area,+fire	Multiple	0.138	0.127	76	327

Terms

Topic Weight	+	Term	Role	# Docs	Freq
0.186	+	hour	Noun	61	66
0.186	+	ranger	Noun	16	26
0.186	+	indicate	Verb	9	12
0.184	+	amount	Noun	16	17
0.184	+	advance	Noun	9	10
0.183	+	group	Noun	27	31
0.18		able	Adj	22	25
0.18	+	extend	Verb	6	9
0.178	+	full	Adv	17	19
0.177	+	stay	Verb	52	57
0.171	+	spot	Noun	26	35

Documents

Topic Weight	
0.249	comment removed
0.248	comment removed
0.246	comment removed
0.245	comment removed
0.245	comment removed
0.244	comment removed

Figure 17: The Text Topic node's "Topic Viewer"

Having both the Text Cluster and Text Topic nodes begs the question, which node is best? The answer is a matter of translating business objectives to analysis objectives. It is left to the analyst to choose from a variety of multivariate and univariate techniques for the available data.

When exploring patterns in data, it is helpful to have all available data in a single data set. The Cluster and Topic nodes can run in series (Figure 18) to generate both sets of variables so that they are available if needed.

Figure 18: Process flow using both the Text Topic node and Text Cluster node

Furthermore, knowledge discovery is enhanced by combining existing organizational data (e.g., park attributes such as type of park, flush toilets, showers and available services). Customer origin data such as city, province and latitude/longitude derived from the respondents' Postal Code become additional variables. All of these variables can (and should) be used for segmenting or data mining.

At this point, the data have been transformed from free text to a form that is ready for insight discovery. New quantitative variables are derived from the qualitative data and the original comments are no longer required for analysis. The analyst is ready to look for relationships and patterns in the data.

FROM TEXT MINING INTO DATA MINING

SAS Enterprise Miner allows for a seamless transition from text mining to data mining and statistical analysis. It is beyond the scope of this paper to describe these in detail. However, two suggestions would be the Regression node and Decision Tree node. The Decision Tree node is one of the more popular features in Enterprise Miner. Both of these nodes are used to see how cluster or topic themes load up on overall customer satisfaction (Figure 1).

Below are some insights revealed by text mining the Camper Satisfaction survey comments. There is no easy way to discover insights like these from the outputs of manual categorization.

- Generator noise complaints are often after quiet hours. Enforcement patrol times could be modified to address this.
- It is important to address the noise issue because it contributes to the perception of safety.
- There are a number of suggestions to improve the campsite reservation process, but because no single issue stands out in text miner, it suggests inconsistency with reservation operations or information.
- The same issue holds true for information services (e.g., signage, campground maps). Consistency is a key component to improving the visitor experience.
- Maintained washrooms correlate to fewer comments about needing new or more washroom facilities.

The list above is province-wide in scope. Similar outcomes and priorities can be identified for each park.

Predictive models and inferential statistical procedures are the difference between outputs and outcomes. Yet having both is important. The first provides context. The second provides guidance. However, decision making is not just a numbers game. For example, a topic related to customer safety should be addressed regardless of statistical significance. Even more, that the topic exists should warrant action. This is one reason for Alberta Parks' latest evolutionary step – to combine the analysis process with leveraging corporate knowledge.

Help your team "find it" with you

When SAS Enterprise Miner is referred to as a discovery tool, it's not just refering to the software, but a process. SAS Enterprise Miner empowers teams to work together to discover patterns and trends in data. The job of the analyst is not to generate reports, disseminate information, and expect clients to carry on. In Alberta Parks, data "reporting" is transitioning to an iterative and collaborative process (affectionately refered to as "huddles").

Analysts meet staff whose experience and knowledge guide analysis. This data-driven dialogue focuses analysis on current and upcoming challenges. As we like to put it, "SAS Enterprise Miner helps your team find insight with you". The huddle approach has been met with positive reviews and a desire to expand huddles across the organization.

Expanding the use of Text Mining

For the public sector, text mining public correspondence (e.g., letters, emails, consultations) is strongly encouraged. This unstructured data can be summarized quickly to improve responsiveness to public concerns.

The word "correspondence" is used intentionally. There is no reason why text mining could not also apply to outgoing responses to ensure message consistency. In testing this, we ran text mining on letters pertaining to a specific hot-topic. It was discovered that part way through events, the outgoing message had changed. Why did this happen? Should this be a cause for concern?

Text Analytics is also often used to listen to social media channels. Lag time with traditional forms of communication can bring topics to a boil. Text mining can identify topics before they become hot-topics, affording the opportunity to be pro-active.

CONCLUSION

SAS Text Miner is a tool for discovering themes and relationships. Text is now a valid datasource for statistical processing. Organizations can leverage exisiting data and use text mining or text analytics tools to improve current processes. This is particularly true now more than ever when organizations are producing text data at accelerating rates.

"Big Data" is a big buzzword, but it is a real problem. Big data problems exist when an organization is under capacity to handle the volume, velocity and variety of data. This data can be consumer or machine generated. If data is

generated at a rate faster than capacity can handle, a big data problem results. Conversely, if an organization is under-capacity, a big data problem exists. Many public sector entities are under-capacity for handling even small data! To address either scenario means investing in these new technologies and the people to use them.

As a researcher, it seems fitting to end this paper with a reference. "If you're not dealing with this data in an intelligent fashion, then it is likely that you are falling behind the business intelligence curve and are probably losing valuable information you might not even know about" [5].

REFERENCES

1. *Alberta Parks 2008 Camper Satisfaction Survey (Provincial Summary): http://www.albertaparks.ca/albertaparksca/library.aspx (keyword: statistics)*
2. Albright, Russell. *Taming Text with the SVD.* January 7, 2004: 72 paragraphs. Available ftp://ftp.sas.com/techsup/download/EMiner/TamingTextwiththeSVD.pdf
3. Cerrito, Patricia B. 2006. *Introduction to Data Mining Using SAS® Enterprise Miner™*. Cary, NC: SAS Institute Inc.
4. SAS Institute Inc. 2012. *Getting Started with SAS® Text Miner 12.1*. Cary, NC: SAS Institute Inc. Available http://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf
5. Collica, Randall S. 2011. *Customer Segmentation and Clustering Using SAS Enterprise Miner, Second Edition*. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

I want to sincerely thank my supervisor, my mentor and my friend, Roy Finzel, who has always supported me in my endeavors and puts up with my shenanigans. I would also like to thank Brian Kelly for providing support for innovative technology and learning opportunities.

Lastly, to all the staff in Alberta Tourism, Parks and Recreation, whose constant dedication is a source of inspiration.

RECOMMENDED READING

- Sanders, Annette and Devault, Craig. (2004) *"Using SAS® at SAS: The Mining of SAS Technical Support." Proceedings of SUGI, 010-29.*

CONTACT INFORMATION

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

Name: Jared Prins
Enterprise: Alberta Tourism, Parks and Recreation – Parks Division
Address: Oxbridge Place 2nd Floor, 9820-106 Street
City, State ZIP: Edmonton, Alberta T5K2J6
Work Phone: 1-780-427-6313
Fax: 1-780-427-5980
E-mail: jared.prins@gov.ab.ca
Web: www.albertaparks.ca

APPENDIX

Sample SAS program used to produce the original frequency output of manually categorized comments.

```
DATA survey;
     INPUT serial_number code1 $ code2 $ code3 $;
     DATALINES;
     1 1a 1b 2a
     2 1b .  .
     3 3a 3b .
     4 1a .  .
     5 1a 3b .
     6 1b .  .
     7 2a 2b 2a
     8 2b 1a 1a
     ;
RUN;


PROC FORMAT;
     VALUE $commentrecode
            "1a","1b" = 1
            "2a", "2b" = 2
            "3a", "3b", "3c" = 3
     ;
     VALUE category
            1="cat 1"
            2="cat 2"
            3="cat 3"
     ;
     VALUE $subcategory
            "1a" = "subcat 1a"
            "1b" = "subcat 1b"
            "2a" = "subcat 2a"
            "2b" = "subcat 2b"
            "3a" = "subcat 3a"
            "3b" = "subcat 3b"
            "3c" = "subcat 3c"
     ;
QUIT;


DATA survey_reshaped (drop = i); SET survey;
     LENGTH code $10.;
     respondent=1; responses=1;

     ARRAY codes(3) code1 - code3;
     DO i = 1 to 3;
            general = input(put(codes(i), $commentrecode.), best8.);
            code = codes(i);

            IF not missing(general) THEN output;
            respondent=.;
     END;
RUN;


DATA codes; SET survey_reshaped;
     BY serial_number;
     IF first.serial_number THEN respondent=1;
     totalrespondent+respondent;
RUN;


PROC TABULATE DATA=codes ORDER=freq;
     CLASS general code;
     VAR responses respondent;
     TABLE   general=' '*(code=' ' all="Subtotal") all='Total',
            responses='Category Totals'*(n='# of Comments'*f=8.0 pctn<code*responses
            all*responses>='% of Category'*f=8.1
```

```
            colpctn='% of All Comments'*f=8.1
            pctn<general*code*respondent general*all*respondent all*respondent>='% Surveys
            Represented'*f=11.1)
            /BOX='General Category / Sub-Category' rts=55;
    FORMAT general category. code $subcategory.;
    TITLE "General and Sub-Category Comments";
RUN;
```