SAS® GLOBAL FORUM 2019

USERS PROGRAM

APRIL 28 – MAY 1, 2019 | DALLAS, TX

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S), Statistical Analysis Cell (SAC)

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Many surveys use open-ended questions to allow participants to express their opinions. When conducting statistical analyses of results, the text responses are often omitted or not analyzed properly. Existing text-mining software may not always be available or suitable for use. Therefore, the Statistical Analysis Cell (SAC), AMEDDC&S HRCoE, JBSA-Fort Sam Houston, TX, propose a practical alternative to performing text analysis utilizing Base SAS 9.4.

We start the text mining process with counting the number of keyword(s) utilized in respondents' statements, then capturing the entire sentence associated with the keyword(s) of interest. This step provides us the context in which the particular keyword(s) were used and helps to eliminate logical errors related to possible misunderstanding of the respondent's intention. Based on the initial results, we propose the next step to define a more precise search. The number of consecutive steps in the data mining process depends on the level of granularity required. This data mining process can be conducted in a stratified environment and will allow statistical comparisons of results. The output statistics provide information about the percent of respondents who used particular keyword(s) in each open-ended question, frequency profile (number of keywords per respondent), univariate analysis of text strings, and number and percentage of respondents with multiple keywords. We illustrate the process with the visual reference and results from survey data.

The views expressed in this article are those of the author and do not reflect the official policy or position of the Department of the Army, Department of Defense, or the U.S. Government.

AC
Statistical Analysis Cell

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S), Statistical Analysis Cell (SAC)

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Introduction

When analyzing survey data, closed-ended questions are typically designed using well-established structures, such as Likert scale, dichotomous scale, or nominal scale. However, text-mining through large amounts of data in open-ended survey responses is a challenge for data analysts and programmers. Quite often, the information included in open-ended responses is not analyzed or analyzed incorrectly. Though, the importance of obtaining data through these types of questions is vital. Exploring only keyword(s) of interest in text data is plausible, but expanding on how the keyword(s) are mentioned in the context of the sentence is immensely important in capturing and fully analyzing text data correctly.

Although text-mining software is available through various companies, there is the need for straightforward text-mining algorithms to be developed within the SAS community to allow for prompt, effective, and efficient utilization of text-mining techniques.

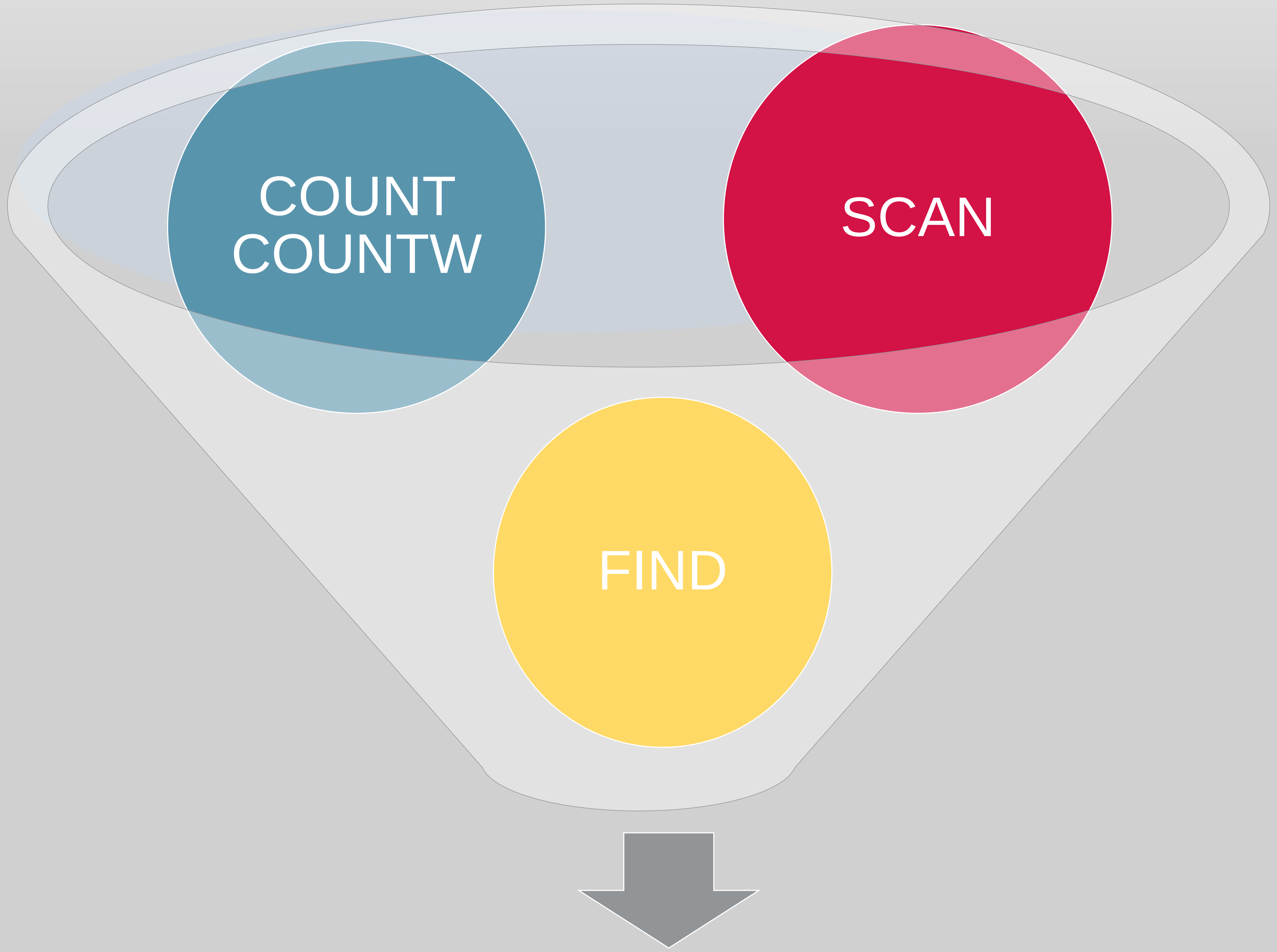
Objective

1. To present detailed classification methodology utilizing inexpensive, readily available Functions that are already in Base SAS® .
2. Reduce time and accurately capture 100% of the keyword(s) of interest.
3. Utilizing SAS Functions; COUNT, COUNTW, SCAN, and FIND, our approach provides a text-mining process capturing the entire sentence associated with the keyword(s) of interest and also provides the number of words used in a string.
4. Provide the context in which the particular keyword(s) were used and help eliminate logical errors related to possible misunderstanding of the respondent's intention.

Methods

Responses to the Medical Corps Engagement/Satisfaction Survey 2018 (MC Survey2018), submitted by 2,050 U.S. Army physicians were examined. The MC Survey 2018 consisted of three sections. Part I and II contained of the questions with answers defined in a Likert scale format or a possible set of other response options. Part III; however, included 4 open-ended questions. The Statistical Analysis Cell (SAC) concentrated on the free-text question involving what improvements could be made for Army Physicians (Q32) and focused on the keyword "pay" to illustrate our methodology. We used a random sample subset of 100 responses. All data analyses were performed using SAS version 9.4.

Visual Reference



Keyword(s) and Context Within Sentence

AC Statistical Analysis Cell

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S), Statistical Analysis Cell (SAC)

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

SAS Syntax

1) Each observation in the survey represented a different respondent's statement. For each observation, the number of keywords used were first counted in each string.

```
Data scan ;
 Set print100 ;
   N_Pay = COUNT(Q32, "pay", "i") ;  * The “i” modifier ignores character case ;
Run ;

Proc Print
 Data=scan ;
 Var N_Pay;
 Sum N_Pay;
Run;
```

Quick Tip: Keyword Concepts

In the **COUNT** function, one substring can be listed in an argument. While “pay” will find variations, such as “paying” and “payment,” the code below shows how multiple keywords can be accepted and eventually summed in a **PRINT** statement. This method allows multiple forms of the word to be recognized for an increased level of concept expansion in open-ended questions.

```
Data scan ;
  Set print100 ;

  PayCncpt =  Count (Q32, 'pay', 'i')
                    +Count(Q32, 'salary', 'i') ;
Run ;
```

SAS Syntax (cont.)

2) Frequency tables were conducted on the amount of times a respondent did or did not use the keyword, ”pay.” The number of times a respondent uses a keyword can provide an indication of the level of concern or appreciation the respondent has on the topic.

```
Proc Freq
 Data=scan_noresponse  ;
 Tables N_Pay ;
Run;

Proc Freq
 Data=scan_noresponse  order=formatted;
 Tables N_Pay * Q40 ; * Q40 is variable for Army Military Rank ;
Run;
```

3) The previous steps will provide a reference point of how many strings with the keyword you can expect. Extracting the actual statements will provide the intention of the respondent's use of the keyword. We utilized a **DO LOOP** with **SCAN** to parse long character string into variable “OUT.” By default, the **SCAN** function checks for any default delimiters: **blank ! $ % & ( ) * + , - . / ; < ^ :.** Since these were the common delimiters in the statements, the entire response was parsed out.

```
Data scan ;
 Set Print100 ;
 Format out $2000. ;
   n = countw(q32) ;
     DO i=1 to n ;
      Out = Scan(q32, i,   ) ;
    If FIND(OUT, "pay", "i") Then   * The “i” modifier ignores character case ;
   Output;
 End;
 Drop i ;
Run;
```

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S), Statistical Analysis Cell (SAC)

AC Statistical Analysis Cell

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Number of Keywords used in Statement

Respondent	N_Pay
1	2
2	0
3	0
4	0
5	0
6	1
7	0
8	0
9	1
10	1
50	0
51	1
52	0
53	0
54	0
55	0
56	0
57	0
58	0
59	0
60	0
90	0
91	1
92	0
93	0
94	1
95	0
96	0
97	1
98	0
99	2
100	0
	32

Overall Respondents Percentage of Keywords

N_Pay	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	61	61.00	61	61.00
1	22	22.00	83	83.00
2	5	5.00	88	88.00
No Written Response	12	12.00	100	100.00

This figure show that 27 respondents used the keyword; however, there were 32 keywords used because some respondents used the keyword twice.

Stratified Percentage of Keywords

Stratification can be used to analyze distributions of keywords by categorical variable.

N_Pay	Rank of Army Physician					
Frequency Percent Row Pct Col Pct	CPT	MAJ	LTC	COL	Missing Rank	Total
0	20 20.00 32.79 66.67	18 18.00 29.51 52.94	15 15.00 24.59 68.18	7 7.00 11.48 53.85	1 1.00 1.64 100.00	61 61.00
1	7 7.00 31.82 23.33	11 11.00 50.00 32.35	3 3.00 13.64 13.64	1 1.00 4.55 7.69	0 0.00 0.00 0.00	22 22.00
2	0 0.00 0.00 0.00	3 3.00 60.00 8.82	1 1.00 20.00 4.55	1 1.00 20.00 7.69	0 0.00 0.00 0.00	5 5.00
No Written Response	3 3.00 25.00 10.00	2 2.00 16.67 5.88	3 3.00 25.00 13.64	4 4.00 33.33 30.77	0 0.00 0.00 0.00	12 12.00
Total	30 30.00	34 34.00	22 22.00	13 13.00	1 1.00	100 100.00

AC Statistical Analysis Cell

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S), Statistical Analysis Cell (SAC)

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Number of Words Calculated

In step 3 of the **SAS Syntax**, this embedded code was used to count the number of words a respondent wrote. This can be beneficial in knowing the level of commitment a respondent has regarding a topic or topics.
It is also useful in the DO LOOP to not guesstimate the number of words until the end-of-statement condition was met.

```
n  =  countw(=Q32)  ;
```

```
DO i=1 to n ;
```

Respondents	n
1	20
2	61
3	72
4	45
5	120
6	45
7	44
8	18
9	119
10	91
11	59
12	210
13	45
14	68
15	45
16	11
17	6
18	18
19	36
20	97
21	50
22	8
23	18
24	63
25	247
26	12
27	284

Descriptive Statistics on Words Calculated for All Respondents vs. Respondents Classified by Rank

```
/* Number of Words in Total Strings */
 Proc Univariate
   Data=scan ;
   Var n ;
 Run;
```

Basic Statistical Measures - ALL

Location		Variability	
Mean	70.8	Range	278.0
Median	45.0	Interquartile Range	73.0

```
/* Number of Words in Total Strings by Rank*/
 Proc Univariate
   Data=scan ;
   Var n ;
   Class Q40 ; * Q40 is variable for Army Military Rank ;
 Run;
```

Basic Statistical Measures - CPT

Location		Variability	
Mean	50.7	Range	204.0
Median	12.0	Interquartile Range	55.0

Basic Statistical Measures - MAJ

Location		Variability	
Mean	93.2	Range	266.0
Median	63.5	Interquartile Range	74.0

Basic Statistical Measures - LTC

Location		Variability	
Mean	49.0	Range	54.0
Median	53.0	Interquartile Range	35.0

Basic Statistical Measures - COL

Location		Variability	
Mean	28.0	Range	16.0
Median	28.0	Interquartile Range	16.0

Text Mining of Open-Ended Survey Data

Brandon J. Hosek, MA, Barbara E. Wojcik, PhD

U.S. Army Medical Department Center and School, Health Readiness Center of Excellence (AMEDDC&S),
Statistical Analysis Cell (SAC)

Abstract
Introduction
Methods
Results 1
Results 2
Conclusion

Conclusion

Analyzing open-ended questions to enhance the information collected through surveys and questionnaires is arguably one of the most challenging and time consuming task for a data analyst, statistician, or programmer. Creating statistical results from and an unstructured form of data to a structured analytical concept leaves most of text-driven responses to be under analyzed or ignored.

Our methodological goal in this presentation was to provide users with available and effective **FUNCTIONS** that exist in BASE SAS® 9.4 for a cost efficient means of finding/counting keywords used in text data and also extracting the full statement to know the respondent's intent, providing the full acquisition of analytical data to produce sound and reliable results. With open-ended questions in big data, the task and time of reviewing responses has significantly left a gap in the full analytic composition of a study. Our contribution to text mining has shown that the ability is there to reduce analytical time and decrease the percentage of responses having to be read. In "Results 1," out of 100 statements, it was reflected that the statements having to be reviewed was significantly reduced from 88% to only 27%. Further stratification also reduces the amount of responses which have to be read if there is a particular topic of interest. As a result, final data analysis is more efficient and effective.

A prospective next step to continue with this methodology is to introduce an explanatory (or construct) variable which consists of several keywords, all of them describing the same concept. For instance, the variable "pay" may consist of "pay", "salary", "raise", 'financial", etc. Create a new categorical variable which would capture the respondent's intent (positive or negative) to use statistical methods to find the level of association between various levels of explanatory variables.

References

General Functions Syntax taken from SAS® 9.4 and SAS® Viya® 3.3 Programming Documentation/SAS Programming Documentation Copyright © 2018 SAS Institute Inc., Cary, NC, USA. Reprinted with permission. All rights reserved.

Suraweera W., Weerasooriya, J., Fernando N. 2018. A simple approach to text analysis using SAS functions. In SAS Global Forum 2018 Conference S. I. Inc, ed. SAS Institute Inc.

Contact Info

Your comments and questions are valued and encouraged. Contact the author at:
Name: Brandon J. Hosek
E-mail: brandon.j.hosek.civ@mail.mil

#SASGF

SAS® GLOBAL FORUM 2019

APRIL 28 – MAY 1, 2019 | DALLAS, TX

Kay Bailey Hutchison Convention Center