Paper 279-2011

Creating word cloud in SAS®: A DSGI approach

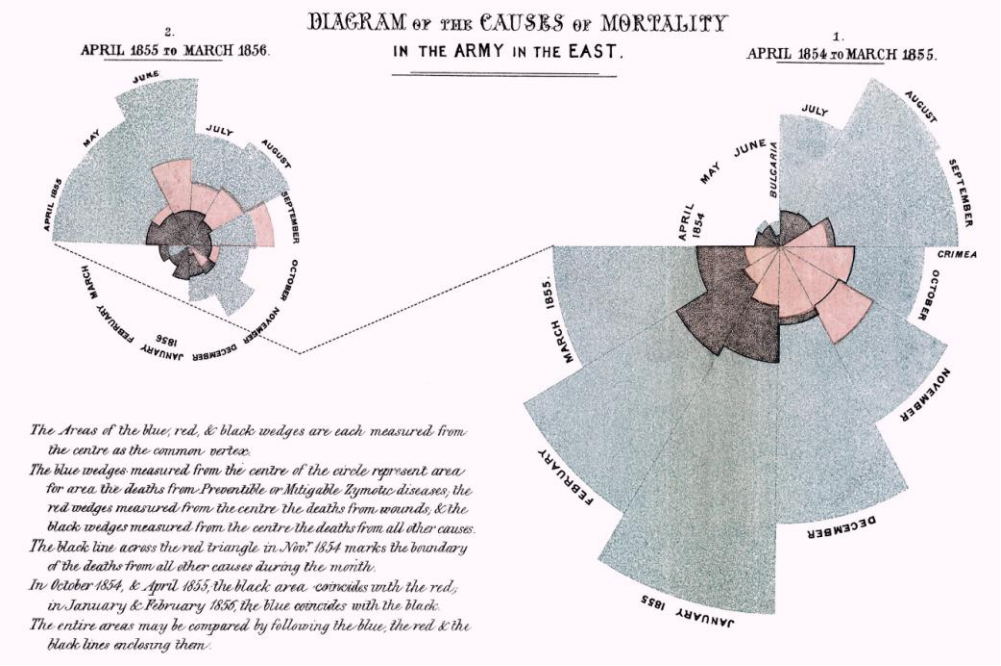
Murphy Choy, School of Information System, SMU, Singapore

ABSTRACT

With an increasing amount of unstructured information on the internet, data visualization of such information is fast becoming an important part of data analysis. This evolution of graphical approaches to handle unstructured data has led to a new field of analytics known as visual analytics. Newer approaches towards visualization have led to a wide variety of graphical techniques such as word clouds, linkage graphs and geo-visualizations. Many of these techniques have not been incorporated into SAS®. However, innovative use of SAS® Macro in conjunction with the use of SAS® DSGI can enable users to exploit the power of visual analytics.

INTRODUCTION

Data visualization has always been an important aspect of statistical inference, especially when numbers cannot convey the full picture. Many famous problems have been uncovered and solved using visual analytics. Florence Nightingale produced an excellent statistical diagram detailing the causes of death in the Crimean War (Tufte, 1990) which is subsequent named Nightingale Rose Diagram (See Below).



This particular graph convinced the politicians of the era of the importance of hygiene in saving the lives of people. While this graph may have been useful in that era, the modern age has brought forth a need for more sophisticated displays of statistics.

WHAT IS WORD CLOUD?

While the Nightingale Rose Diagram might be the best way to display statistical information that is numeric in nature, it is of little practical use to the world of unstructured data. New demands require new solutions. One of the most frequently cited solutions to the world of increasing textual data is the word cloud chart.

THEORY BEHIND WORD CLOUD

Word cloud charts are extremely simple representations of textual information created by randomly assigning positions in a chart area. Each word is plotted in the font that is most representative of the impact

of the word, which is typically exemplified by how often it occurs in the data itself. By having varying word size, we can visualize the various impacts of the words in the article which is typically unstructured data.

The approach has been formalized using the equation below, kindly extracted from Wikipedia (2010).

$$s_i = \left\lceil \frac{f_{\max} \cdot (t_i - t_{\min})}{t_{\max} - t_{\min}} \right\rceil$$ for ti > tmin; else si = 1

- si: display fontsize
- fmax: max. fontsize
- ti: count
- tmin: min. count
- tmax: max. count

Below is an example of word cloud created using SAS®.

From the data, we can see that the word Singaporeans, HDB, flats and housing are the key words that are highly emphasized in a particular speech by an important politician in Singapore. From the word cloud, we can easily see what the main themes in the speech are and where the speaker is heading. Given such usefulness in analyzing speeches, why is this so rarely found in SAS® papers?

PROBLEMS IN CREATING WORD CLOUDS

The main barrier to the creation of word clouds in SAS® is that there is no pre-defined SAS® Procedures to generate these graphs. To create word clouds, one requires a customized approach to building the chart which is not necessarily the simplest of tasks for ordinary SAS® programmers or analysts. At the same time, most of the existing SAS® graphing tools do not offer the extended capability to inscribe words easily in a chart form. However, with the new experimental DSGI, it has become easier to draw such charts.

DSGI is the SAS® Data Step Graphical Interface which is experimental in SAS 9.1.3. DSGI provides a variety of tools to draw charts or word in a plot with the flexibility of chart placement. Given these functionality, the DSGI is perfect for creating the word cloud. By using a SAS® Macro (Appendix A), we can plot these clouds with ease.

ANALYZING THE WORD CLOUD

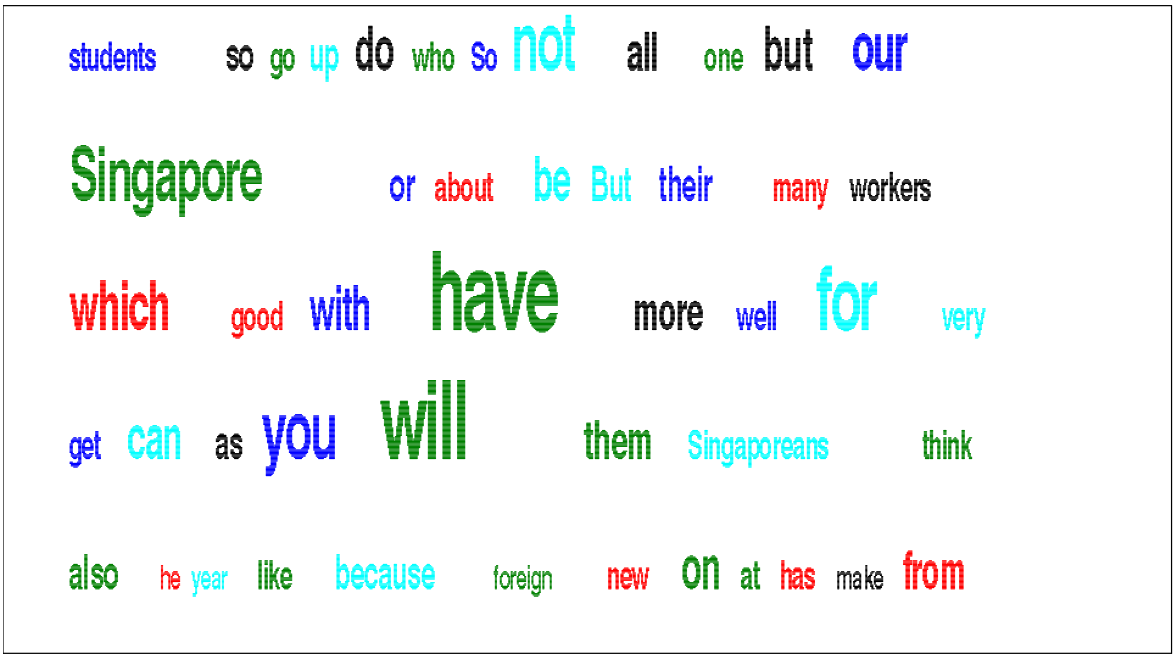
There are several types of word clouds currently used by analysts. The most common is the one showcased above—the size-frequency approach. The frequency of word appearance dictates the size of that word in the final output. This type of word cloud is very useful in the analysis of speeches or unstructured text.

The fundamental advantage of word cloud in analyzing data lies in the graphical representation of the words. The random positioning allows us to ignore the syntactical structure of sentences and instead, we focus on words with high frequency, and hence focus on the overall theme. By changing the size of the words with respect to the frequency, we can observe the important words and subsequently link them up.

Let us examine a few examples and see how word cloud can be used for speech analysis.

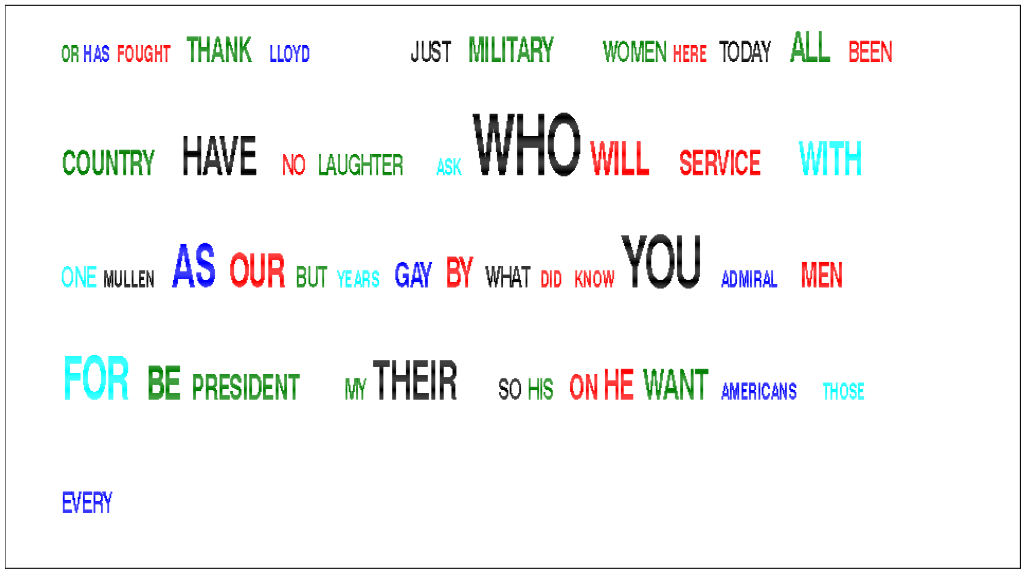
EXAMPLE SPEECH 1: NATIONAL DAY RALLY 2010

Below is a word cloud of the speech from the Prime Minister of Singapore in his annual address and rally to the people of Singapore. Some of the key themes are Singapore, workers, students, will, Singaporeans. Another interesting thing to note from the word cloud is the presence of negative words such as not and but. There are not too many positive words here which indicate an overall gloomy environment.



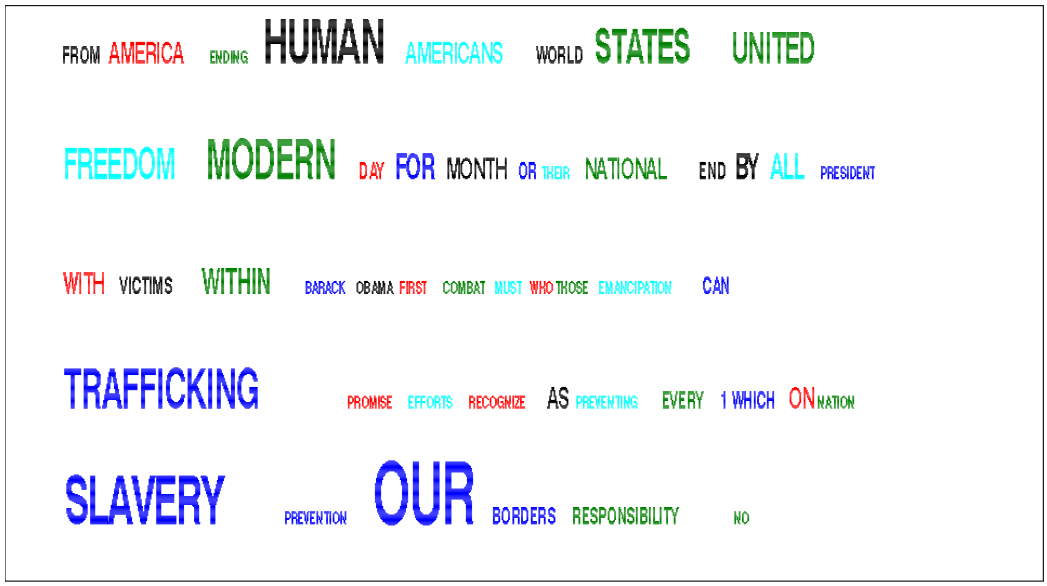
EXAMPLE 2: PRESIDENT OBAMA'S REMARK ON REPEALING OF "DON'T ASK, DON'T TELL".

Below is the tag graph of President Obama's remark on the action of repealing of the "Don't ask, don't tell" rule. This is perhaps one of the longest speeches analyzed so far. The main theme is as expected to be focused on the issue of men and women in the military. However, interestingly, the word you is heavily used in the speech and can be interpreted as an acknowledgement that the achievement is by the hands of many people which the President did not named.



EXAMPLE 3: PRESIDENT OBAMA'S PROCLAMATION OF ANTI SLAVERY MONTH

From the President's speech, there were strong underlying themes in the proclamation with words like Slavery, Modern, Trafficking, United States and Freedom. One very interesting word that is highlighted in the speech is the word Our. The particular use of this word seems to indicate a heavy focus on such problems within the United States. The presence of words such as borders, Americans, freedom and human verifies the position of the heavy focus on the internal aspects of these problems.



CONCLUSION

Word clouds provide excellent visualization as a way of understanding data. With the use of DSGI, we can create such charts in a simplified manner. DSGI is a flexible approach to solving some of the graphing needs of the SAS® users.

ACKNOWLEDGEMENTS

Much credit has to be given to Richard DeVenezia's work in both Word Cloud and DSGI. His earlier paper on this topic has generated widespread interest in word clouds. I will also like to acknowledge the support and editorial effort from Phil and my wife in editing my paper.

BIBLIOGRAPHY

Tag Clouds - A list of tokens, sized by relative frequency, Paper SIB-096, SESUG 2008, Richard A. DeVenezia

CONTACT INFORMATION

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

Name: Murphy Choy
Enterprise: School of Information Systems, Singapore Management University
Address: 80 Stamford Road
City, State ZIP: Singapore 178902
Work Phone: +65-92384058
E-mail: goladin@gmail.com/murphychoy@smu.edu.sg

APPENDIX A:

```
/*************************************************************************
-------------------------------------------------------------------------
TAG CLOUD CREATION MACRO
-------------------------------------------------------------------------
THIS MACRO USES THE DSGI APPROACH TO CREATE THE TAG CLOUD IMAGES WHICH ARE
WIDELY USED AS A DISPLAY OF THE MOST CRITICAL INFORMATION FROM A PIECE OF
TEXT. THIS GRAPHICAL APPROACH HAS BEEN IN USE FOR SOME TIME AND HAS RECENTLY
GAINED SOME MOMENTUM IN TEXT ANALYSIS.
-------------------------------------------------------------------------
INPUTS                  DESCRIPTION
-------------------------------------------------------------------------
INPUT                   TEXT FILES DATA PREPARED IN A PARTICULAR WAY
MAXFONT                 MAXIMUM FONT SIZE
-------------------------------------------------------------------------
*************************************************************************/


/*************************************************************************
MAXIMUM FONT SIZE
*************************************************************************/

%MACRO TEXT_CLOUD(INPUT);

/*CLEANING THE DATA UP*/

DATA &INPUT;
SET &INPUT;

/*REMOVING SYMBOLS*/
WORD = TRANSLATE(WORD,'                          ',',.?!:&*$%@#_+-=/{}[]|()');

/*REMOVING UNNECESSARY WORDS*/
IF UPCASE(WORD) IN ('IS','ARE','WAS','WERE','THE','I','WE','THEY','THERE',
'THESE','IN','A','OF','TO','AND','IT','THAT','THIS') THEN DELETE;

RUN;

/*DATA PRESUMMARIZATION*/

PROC SQL NOPRINT;

    CREATE TABLE WORDLIST AS SELECT WORD AS WORD, COUNT(*) AS COUNT FROM
&INPUT GROUP BY WORD ORDER BY COUNT DESCENDING;

QUIT;

/*DOW LOOP TO ACHIEVE MIN AND MAX CALCULATIONS*/

DATA WORDLIST;

/*DECLARING THE MIN MAX VALUES*/
MIN = 100;MAX = 0;

    /*ESTIMATING THE DATASET MIN AND MAX COUNTS*/
    DO UNTIL (EOF);
           SET WORDLIST END = EOF;

           IF COUNT < MIN THEN MIN = COUNT;

           IF COUNT > MAX THEN MAX = COUNT;
```

```
    END;

    /*CALCULATING THE FONTSIZE*/
    DO UNTIL (EOF2);
           SET WORDLIST END = EOF2;

                  FONTSIZE = CEIL(14*(COUNT - MIN)/(MAX - MIN))+2;
                  OUTPUT;
    END;


RUN;

/*SORTING THE DATA AND CREATING THE NEEDED DATA FOR PRINTING*/

PROC SORT DATA = WORDLIST;
BY DESCENDING FONTSIZE;
RUN;

DATA WORDLIST;
SET WORDLIST(OBS = 49);
RANDOM = RANUNI(0);
RUN;

PROC SORT DATA = WORDLIST;
BY RANDOM;
RUN;

/*DECLARING THE MACRO VARIABLES*/

DATA _NULL_;
SET WORDLIST;

CALL SYMPUT("WORD"||TRIM(LEFT(_N_)),WORD);
CALL SYMPUT("FONT"||TRIM(LEFT(_N_)),FONTSIZE);
CALL SYMPUT("COUNTS",_N_);

RUN;

/*DECLARING THE IMAGE SPECIFICATIONS*/

GOPTIONS RESET=GLOBAL GUNIT=PCT BORDER
     HSIZE=10 IN VSIZE=6 IN;


/* EXECUTE DATA STEP WITH DSGI */
DATA _NULL_;

    /*CREATION A LOCATION MEMORY SPOT*/
    RETAIN X Y;

    /*INITIAL SPOT*/
    X = 10;Y = 90;

    /* PREPARE SAS/GRAPH SOFTWARE */
    /* TO ACCEPT DSGI STATEMENTS */
    RC=GINIT();
    RC=GRAPH("CLEAR");

    /*SETTING THE FONT*/
    RC=GSET('TEXFONT', 'SWISSB');

    /*CREATION OF COUNTER*/
```

```
    %LET BASE = 0;

    /*CREATING THE LOOP*/
    %DO I = 1 %TO 7;

           %DO J = 1 %TO 7;

                  /*EVALUATE THE COUNTERS*/
                  %LET BASE = %EVAL(&BASE+1);

                  /*DECLARING THE TEXT COLORS*/
                  RC=GSET('TEXCOLOR', FLOOR(RANUNI(0)*5+1));

                  /*CALCULATING THE NEXT POSITION*/
                  IF X + 0.5*(&&FONT&BASE)*LENGTH("&&WORD&BASE") > 160 THEN
DO;
                         X = 10;
                         Y = Y - 20;
                  END;

                  /* DEFINE HEIGHT OF TEXT */
                  RC=GSET("TEXHEIGHT", &&FONT&BASE);
                  RC=GDRAW("TEXT", X, Y, "&&WORD&BASE");

                  /*CALCULATING THE NEXT POSITION*/
                  X = X + 0.5*(&&FONT&BASE)*LENGTH("&&WORD&BASE");


           %END;

    %END;

/* DISPLAY THE GRAPH AND END DSGI */
RC=GRAPH("UPDATE");
RC=GTERM();

RUN;

%MEND;

/*************************************************************************/
```