Paper 1687-2018

SAS® Visual Analytics: Text Analytics Using Word Clouds

Jenine Milum, Atlanta, GA, USA

ABSTRACT

There exists a limit to employee skills and capacity to efficiently analyze volumes of textual unstructured data in a manner that provides actionable business insight. SAS provides several tools that provide the capacity to explore text; SAS® Text Miner, SAS® Visual Analytics , SAS® Contextual Analysis, and SAS® Visual Text Analysis. Utilizing Text Analytics & Sentiment Analysis within SAS Visual Analytics Word Clouds puts the availability and usefulness of these tools in the hands of a much broader audience. Visual Analytic Word Clouds deliver a point and click tool to explore volumes of unstructured data, identify patterns and create usable reports to aid in defining policies to improve business operations.

INTRODUCTION

This paper will utilize the text from Shakespeare's play "A Midsummer Night's Dream". It's an excellent source for unstructured text data providing ample opportunity for analysis and exploration for this exercise. Similar to text that may be captured through company surveys, the lines in the play show varying moods and topics that provide insight to each of its contributors.

We will discuss the terminology, tools used to perform text analytics with word clouds and sentiment analysis. We will explore considerations for preparing our data, moving it into the SAS Visual Analytics environment and explore the possibilities available to present the results in a World Cloud and other formats for further actions. The user will need to have available textual data (in our case we will be using a .csv that has been converted to a SAS data set), SAS Visual Analytics versions 7.1, 7.4 or 8.x (this presentation is using 7.4).

TERMINOLGY

TEXT ANALYTICS

Text Analytics is the process of converting unstructured text data into meaningful data for further analysis. SAS software applies sophisticated linguistic tools to consistently provide context, apply numerical representation of text and categorization for further investigation.

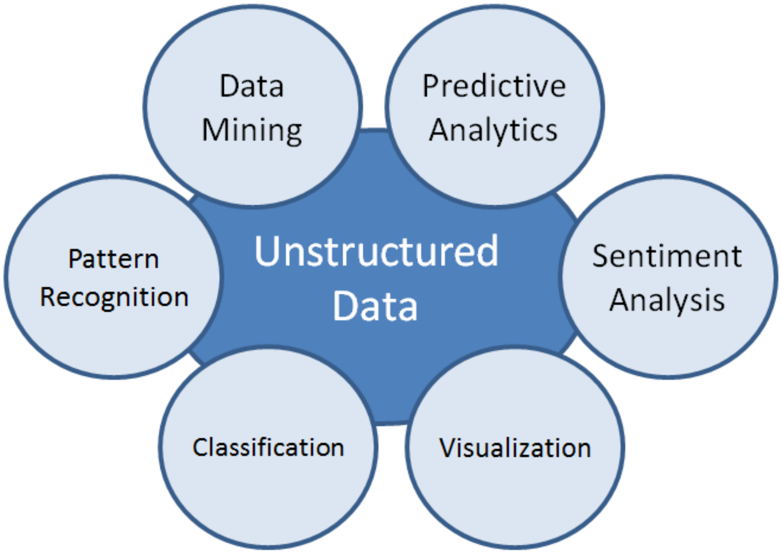


Figure 1 – Unstructured Data Uses

WORD CLOUD

Word clouds are a graphical representation of words from textual data. Depending on the type of word cloud designed, words appear more prominent based on the frequency or importance of that word or phrase. "If a picture paints a thousand words, then what can a thousand words paint?"

Caption 1 – Word Cloud

SAS® VISUAL ANALYTICS TERMS

Topic

A machine-generated category, the purpose of which is to indicate what documents are about. A topic identifies groupings of important terms in a document collection. A single document can contain one or more topics, or no topics.

Term

A representation of a single concept in one or more textual forms, as defined by rules or algorithms.

Term Role

A function that is performed by a term in a particular context. A term can function as a part of speech, entity type, or other purpose that is user defined.

Topic Term Weight

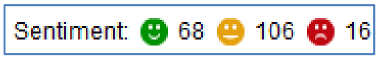
An indicator of the relative importance of a term in a topic as compared to other terms. A term with a value above a specified cutoff value contributes to the assignment of a document to the topic.

Relevance

A score that indicates how well a document satisfies a rule or model. The best match has a score of 1 and reflects a perfect (100%) match.

Sentiment

An attitude that is expressed about an item that is being analyzed, which can be a segment of text, a group of text segments, or a specific subject of interest.



Sentiment: ☺ 68 😐 106 ☹ 16

Caption 2 – Sentiment Distribution

Sentiment Analysis

The use of natural language processing computational linguistics and text analytics to determine the attitude of a speaker or writer with respect to a topic document or other item of analysis. Sentiment analysis results in a positive, negative, or neutral score on the target of analysis.

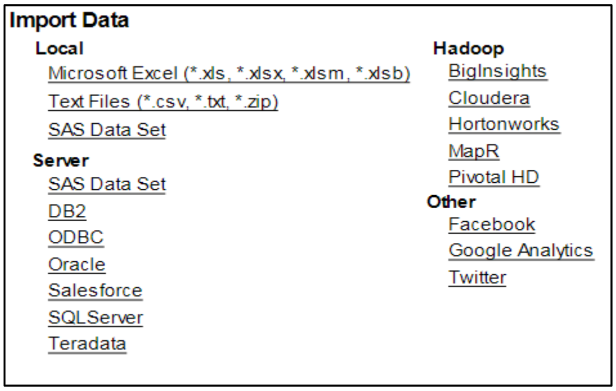
PREPARING DATA

All text data seems to come with flaws. Although it is not necessary, I find it best to prepare unstructured data prior to loading it into the Visual Analytic environment. Transforming all the text to a single case, either upper or lower case, makes the resulting word cloud more visually consistent. If there is a word or phrase which is used in a statistically significant manner as to be captured in the analysis for which you already know is not of use, consider removing it from the unstructured text prior to Visual Analytic loading. You may find that the text as a whole is worthy of contribution but want to avoid a word being a distraction. As an example, suppose a company's name is officially "Exciting". In the text analysis, the word exciting would be considered an adjective rather than a noun and would skew the sentiment categorization. It's best to look at the data for anything you determine would unfavorably alter or interfere with the statistical processes being applied.

To create a Word Cloud in Visual Analytics, a Unique Row Identifier is required. For example, a collection of survey data may contain a field for a survey code which is uniquely identified with a text field. This is quite sufficient for VA's Unique Row Identifier classification. If such a field is not available, it's as simple as creating an additional column with the row count with each row being unique from the others.

For the purpose of this paper which uses the play "A Midsummer Night's Dream", I have captured the following data points for analysis. The (Line) spoken by each (Character) and the (Act) the line was delivered. A row count (SEQ) was created to be used as a Unique Row Identifier.

Visual Analytics will accept data input in many different formats. For our purposes, we are using data that has been converted into a SAS dataset and loaded into memory using the LASR server as set up by the system administrator.



Import Data

Local
- Microsoft Excel (*.xls, *.xlsx, *.xlsm, *.xlsb)
- Text Files (*.csv, *.txt, *.zip)
- SAS Data Set

Server
- SAS Data Set
- DB2
- ODBC
- Oracle
- Salesforce
- SQLServer
- Teradata

Hadoop
- BigInsights
- Cloudera
- Hortonworks
- MapR
- Pivotal HD

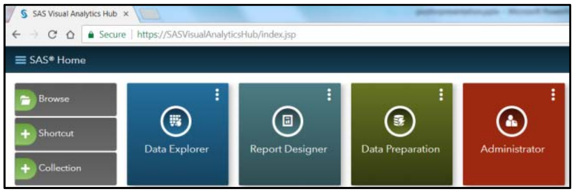
Other
- Facebook
- Google Analytics
- Twitter

Table 1 – Data Source Options

STEPS TO CREATE A TEXT ANALYTICS WORD CLOUD

GETTING STARTED IN VA

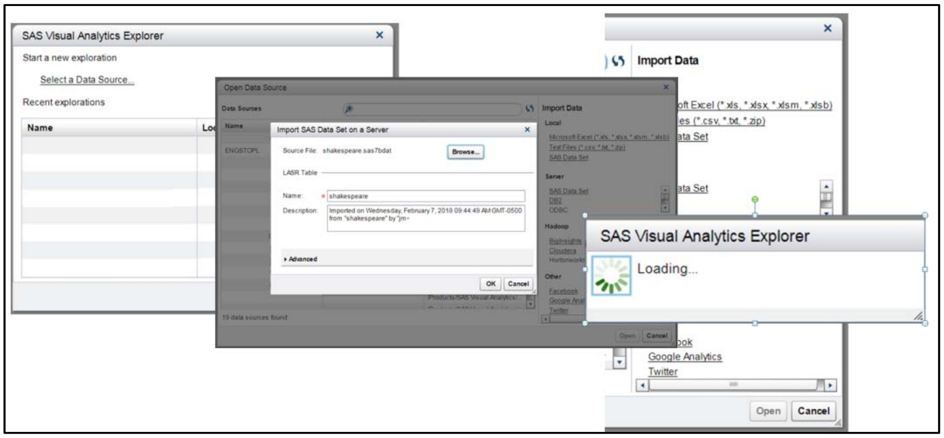
Having logged into your Visual Analytics environment; to begin the process of creating a Word Cloud using Text Analytics, select your Data Explorer. You can create a Category Word Cloud in both the Data Explorer and the Report Designer, but the Textual Analytics version is only available in the Data Explorer.



Caption 3 – Visual Analytics 7.4 Home Screen

The Visual Analytics wizard will step you through the process of selecting your data. For this paper's purpose, the input is a SAS dataset located on the server but has not been loaded into memory on the LASR for Visual Analtyics use. The wizard assists with the data load.

1. Select "Select a Data Source" if your data has not already been loaded.
2. Select the format of the data being loaded under "Import Data".
3. Browse the location of the data to be imported.
4. Select OK if the data is ready to load. If it had previously been loaded and an update is required, either rename the data or replace.



Source File: shakespeare.sas7bdat

Name: * shakespeare

Description: Imported on Wednesday, February 7, 2018 09:44:49 AM GMT-0500 from "shakespeare" by "jm~

19 data sources found

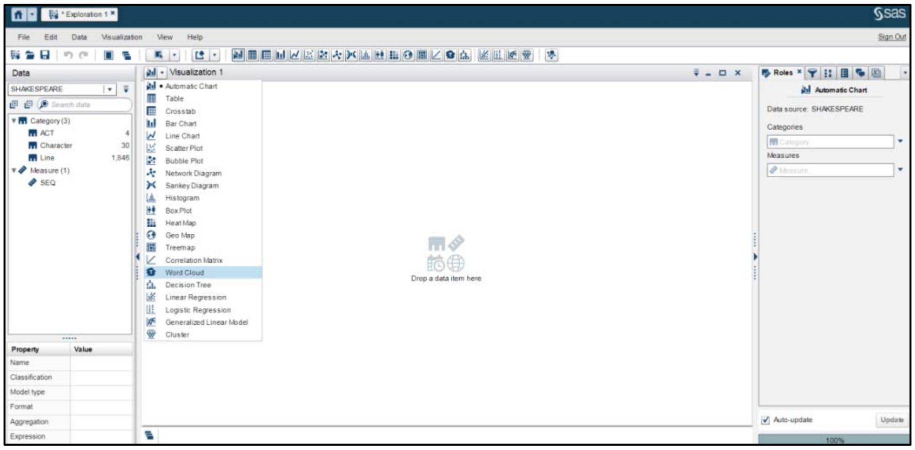
SAS Visual Analytics Explorer

Loading...

Caption 4 – Data Source selection

INITIATING A TEXT ANALYTICS WORD CLOUD

Having selected your data for the Data Explorer, you may now take the steps to set up the parameters that will result in a meaningful and descriptive Word Cloud. At this point, you have a data source and a blank work space. The Data Explorer provides a drop down list of all the data Explorations available.



SHAKESPEARE

Category (3)	
ACT	4
Character	30
Line	1,846
Measure (1)	
SEQ	

Data source: SHAKESPEARE

Caption 5 – Data Explorer Workspace & Visualization Selection Options

Now that Visual Analytics knows what you are initially intending to create, it needs to know more about your data. As discussed in the section, Preparing Data, one of the variables needs to be selected as the Unique Row Identifier.

5. Right click on the variable SEQ under the Data section located to the left of your workspace and select “Set As Unique Row Identifier”.

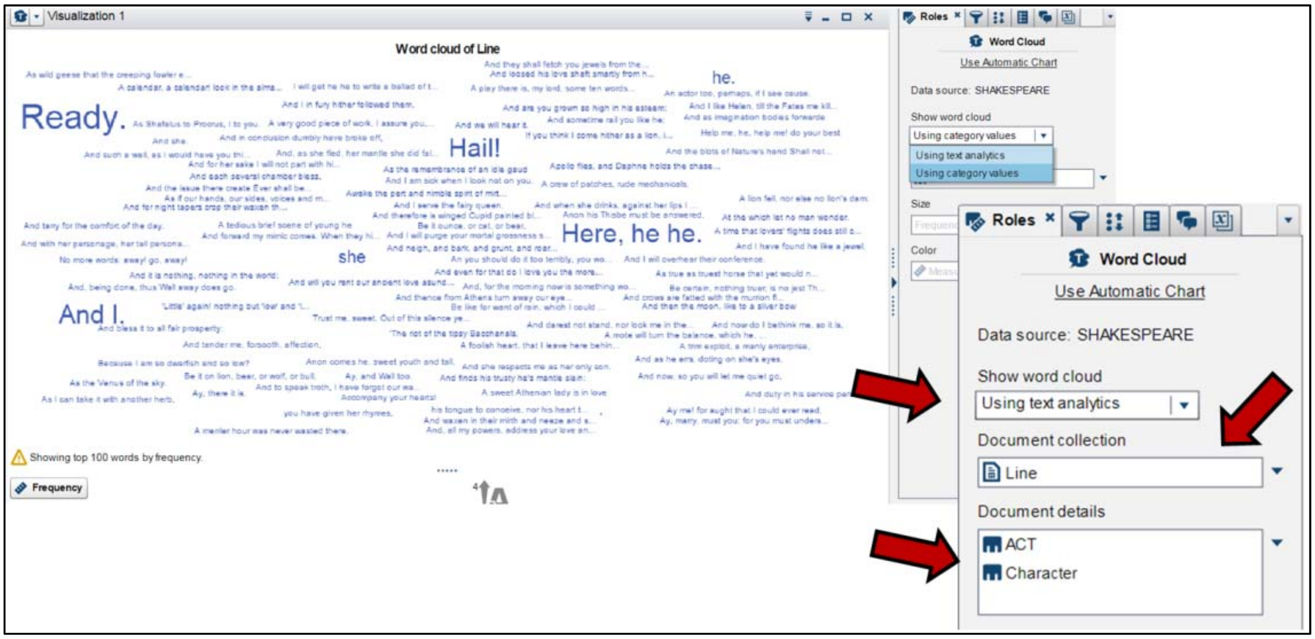
VA needs to know what field contains the unstructured text to be utilized in the text analytics word cloud. Without this designation, your Word Cloud object will assume you are using Category values which is a different type of word cloud that analyzes phrases for prevalence rather than Text Analytics with allowing for Sentiment Analysis.

6. Right click on the variable Line under the Data section located to the left of your workspace and select “Document Selection”.

Switching to the pane to the right of our workspace provides tabs that allow us to customize and control the content, look, analytic factors and much more as desired for the end results.

7. Under the Rolls tab, If not already selected, select “Using text analytics from the Show word cloud drop down list.

8. Under the Documentation Collection category, either drag and drop the variable from the far left pane “Line” into the workspace or select it from the list of variables available when clicking on the down arrow.
9. If you have any remaining variables you want to capture for additional reporting and investigation, include those under Document Details. We will review their application later in this paper.



Data source: SHAKESPEARE

Show word cloud: Using text analytics

Document collection: Line

Document details: ACT, Character

Caption 6 – Selecting Word Cloud Rolls & Properties

The “Filters” tab Filters × provides many opportunities to control the content of the data included in your Word Cloud. While we are not going to apply any during this discussion, some application of this tool would usually be helpful. If your data contains rows with content you do not want to include, you can filter out those entire rows from being included. For our illustration, suppose I don’t want to include any of the conversation in A Midsummer Night’s Dream provided by the Fairies. In that case, I would want to filter out any rows for the Characters Peaseblossom, Cobweb, Mote, and Mustardseed.

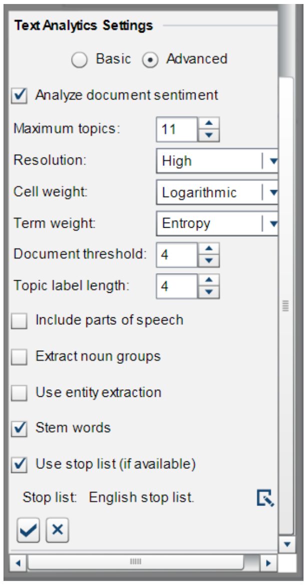
To continue with our analysis of this famous play, we select the “Properties” tab. Here we have the opportunity to make many selections that affect our output.

10. From the Properties tab, create a name for your Visualization.   If you would like to create your own graph title, uncheck the “Generate graph title” and provide your own text.
11. In the section labeled Text Analytics Settings, located on the Properties tab, check the box labeled “Analyze Document Sentiment” and then accept the change.
12. In the same section, change the selection from Basic to Advanced. Additional options become available. Also select “Stem words” and “Use stop list”.

For the purpose of this paper, we will make modifications to the following selections: Stem Words, Use Stop List and Stop List: English stop list. Definitions for all of the selection options on this tab are available from the SAS online documentation (Institute). Stem words considers all words with the same base word (stem) as a single word. Example: love, loved, loving, loves.

The use of the English stop list is immensely helpful. The list can be downloaded from the SAS Support website and is called engstopl.sas7bdat. Depending on the permissions controlled by the SAS administration supporting your Visual Analytics environment, you can either load the stop list or have the administrators load it for all to use.

The stop list contains a list of English words we don't usually want to be considered for analytical purposes. Words such as "the", "and", "at" are eliminated from analysis. Because the list is in a SAS dataset format, additional words may be added (or removed) based on your unique needs.



Text Analytics Settings

○ Basic ● Advanced

☑ Analyze document sentiment

Maximum topics: 11

Resolution: High

Cell weight: Logarithmic

Term weight: Entropy

Document threshold: 4

Topic label length: 4

☐ Include parts of speech

☐ Extract noun groups

☐ Use entity extraction

☑ Stem words

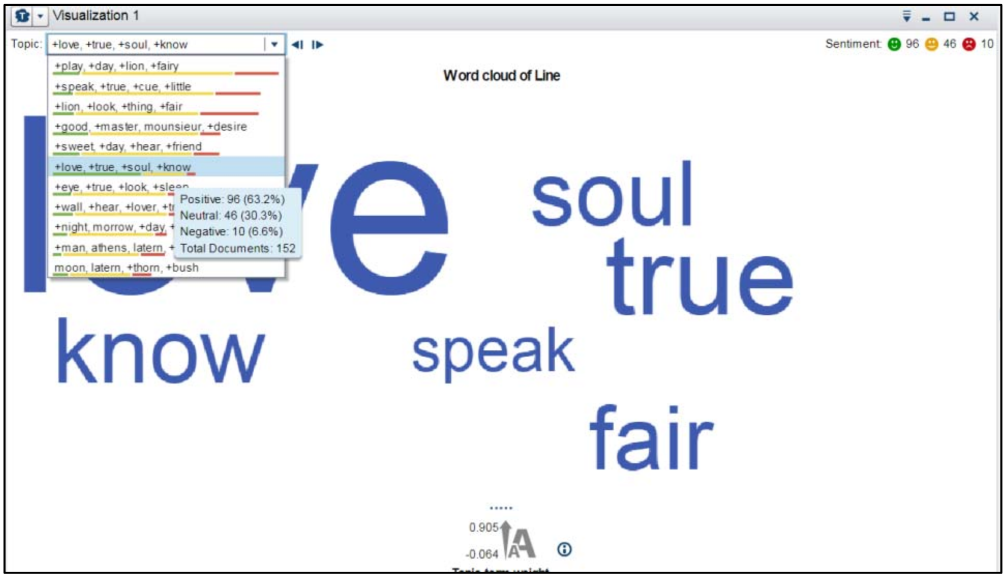
☑ Use stop list (if available)

Stop list: English stop list.

Caption 7 – Properties tab

INTERPETING THE WORD CLOUD

At first glance, the words that appear in the newly defined word cloud may seem as if they don't represent the text you provided as input entirely. That is because the Text Analytics feature has taken our unstructured data, applied statistical methods, and generated meaningful word groupings. These groupings are referred to as Topics. A list of the most relevant topics are listed in the drop down list in the top left of the Object workspace pane. The words visually represented in this example are specifically associated with the specific topic selected.



Visualization 1

Topic: +love, +true, +soul, +know

+play, +day, +lion, +fairy
+speak, +true, +cue, +little
+lion, +look, +thing, +fair
+good, +master, mounsieur, +desire
+sweet, +day, +hear, +friend
+love, +true, +soul, +know
+eye, +true, +look, +sleep
+wall, +hear, +lover, +t
+night, morrow, +day, +
+man, athens, latern, +
moon, latern, +thorn, +bush

Sentiment: 96 46 10

Positive: 96 (63.2%)
Neutral: 46 (30.3%)
Negative: 10 (6.6%)
Total Documents: 152

Caption 8 – First glance of Text Analytics Word Cloud

The value of statistically derived topics comes from topics being identified for which current human interpretations may not have considered relevant. Knowing A Midsummer Night's Dream, Love is certainly a topic associated with many of its events.

Additional analytic interpretation is provided in Caption 8 for our +love +true+soul+know topic. As selected in step 11 above, the sentiment of the Line is derived. Note the Green, Yellow and Red sentiment faces in the top right of Caption 8. They represent the entirety of this top with Green being positive, Yellow neutral and Red negative.   The distribution of sentiment is visually included in the bar directly underneath each topic in the drop down list. If I were to be more interested in the most positive of topics, this visualization captures my attention. There is additional value for such representation of sentiment. What if it showed there were more negative sentiments surrounding the +love +true+soul+know topic than expected? Attention would be drawn and I'd be able to focus on the specific comments associated with the results.

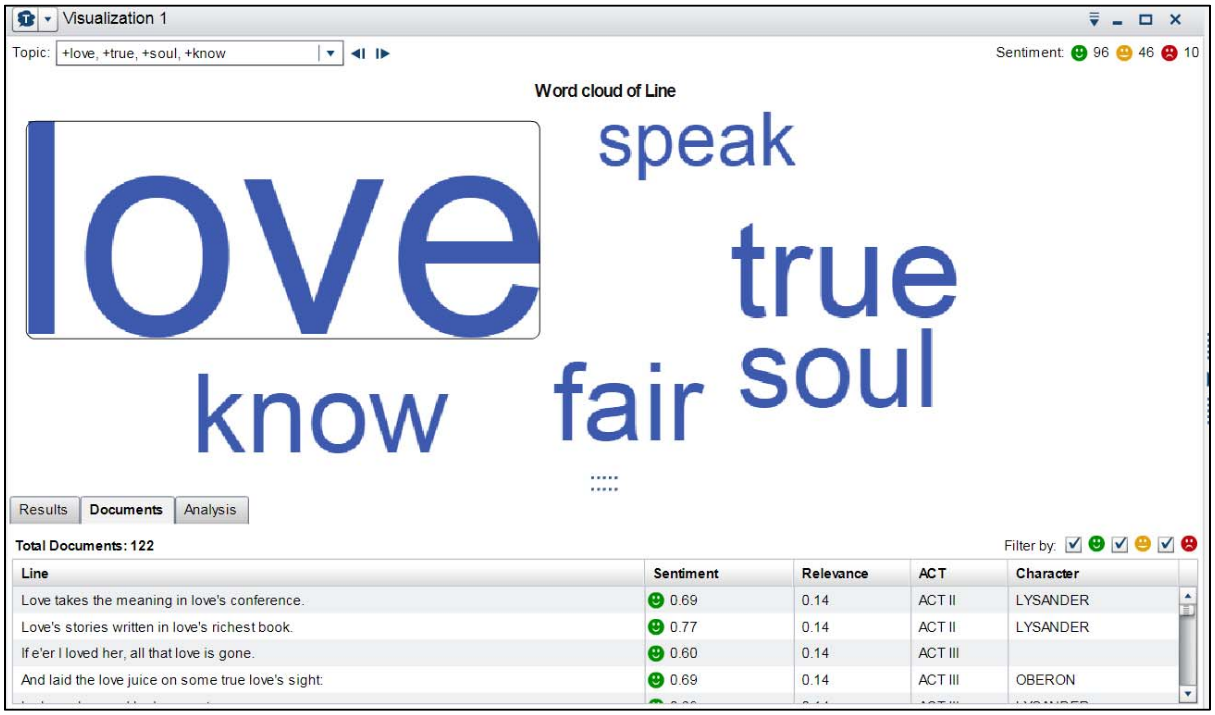
DOCUMENT DETAILS

Now that a topic of interest based on sound statistical methods has been identified, we want to explore the results further. Some items one would want to consider when analyzing the current scenario.

- Was there a character in the play that was focused on this topic more than others?
- Was there an Act in the play that focused on this topic more predominately than the others?
- Was there an Act or Character which more strongly represented positive sentiment than the others?

INTERPETING THE WORD CLOUD

To create data in a format to answer these questions, select a single word in the Word Cloud for the +love +true+soul+know topic. I've highlighted the word "love". Magically, the lower portion of our page is populated with the specific details for our selection. Not only do we have the lines spoken in the play associated with this topic/word, but the sentiment and topic relevance are provided. Not the columns for the variables Act and Character are included in the report. These appear because we identified them as data items of interest in step 9.

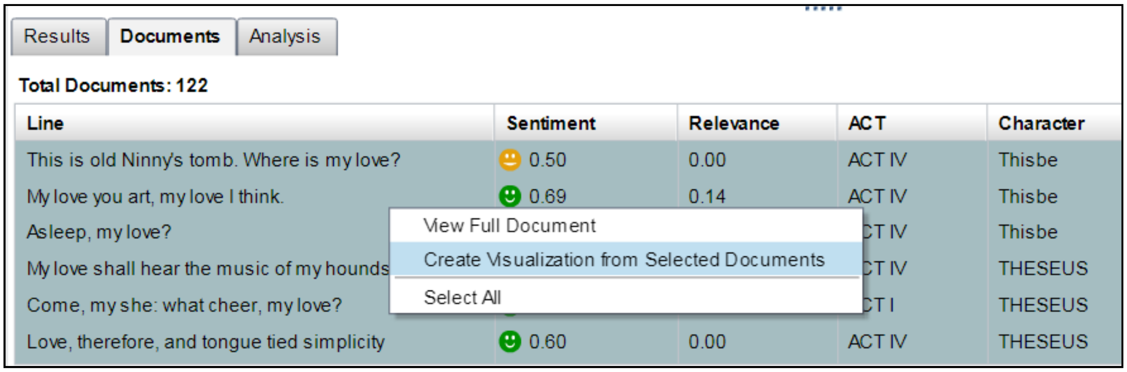


Total Documents: 122

Line	Sentiment	Relevance	ACT	Character
Love takes the meaning in love's conference.	0.69	0.14	ACT II	LYSANDER
Love's stories written in love's richest book.	0.77	0.14	ACT II	LYSANDER
If e'er I loved her, all that love is gone.	0.60	0.14	ACT III	
And laid the love juice on some true love's sight:	0.69	0.14	ACT III	OBERON

Caption 9 – Document Details for "Love"

Note in Caption 9 within the Documents detail to the right are check boxes for the 3 sentiment levels. The number of total unstructured data contributions is provided. One also has the ability to sort the columns, rearrange and resize the report. This information may also be exported for distribution or used to create additional reporting objects within VA. As seen in caption 10, right click on the data and select Create Visualization from Selected Documents.

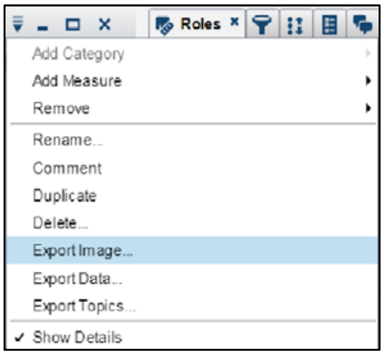


Total Documents: 122

Line	Sentiment	Relevance	ACT	Character
This is old Ninny's tomb. Where is my love?	0.50	0.00	ACT IV	Thisbe
My love you art, my love I think.	0.69	0.14	ACT IV	Thisbe
Asleep, my love?			CT IV	Thisbe
My love shall hear the music of my hounds			CT IV	THESEUS
Come, my she: what cheer, my love?			CT I	THESEUS
Love, therefore, and tongue tied simplicity	0.60	0.00	ACT IV	THESEUS

Caption 10 – Create additional reporting

Not only is the data detail easily distributed, the Word Cloud itself is can be exported. Simply select the chevron drop down list located at the top right of the pane containing the word cloud. The resulting image will be saved as a .png and available for distribution. Additional controls and actions are also available in this drop down that should be explored.



Caption 10 – Export Image & other actions

CONCLUSION

While Shakespeare provided us with unstructured text to utilize in the demonstration of SAS Visual Analytic Word Clouds using Text Analytics, what are some practical business applications of this tool? Most companies provide products and services. Capturing customer opinions in regards to the quality of products and services is of upmost importance to the companies continued success. We can provide a number of stars to express our opinion on the performance of a play, but a textual response for the motivation of the score is far more valuable. Some responses are short and to the point, others are long and reflect on a number of different points with a mix of positive and negative. It would be nearly impossible to provide any deeply meaningful analytics around such surveys using just score ranges and human discernment of the comments.

Let's say I attended a rendition of A Midsummer Night's Dream. When surveyed, I provided the highest score of 5 for my love of the play. The survey also allowed me to comment for which I provided "I loved the play but I couldn't see well from the balcony". Many companies don't have the resources to explore comments further than the negative scored responses. It's likely the producers of the play hadn't even considered the view from the balcony. Using traditional methods of analyzing surveys, this point could have easily been lost.

Text Analytics and the visual capabilities of Word Clouds allow businesses to not only evenly determine topics of interest but identify specific points within topics. Its further capabilities provide visual results to socialize and analytic data to be used for further exploration. The ability to derive considerably more quality information from unstructured data gives any company an edge on the competition.

REFERENCES

SAS® Online Documentation. “Working with Word Clouds.” Accessed February, 2018. Available at: http://support.sas.com/documentation/cdl/en/vaug/68648/HTML/default/viewer.htm#n1oo7kmcwcn1rsn1ll0bo0wuzgdn.htm

ACKNOWLEDGMENTS

I would like to thank Erin Davis of *SAS*® and Tom Keefer for the help and support of this paper.

RECOMMENDED READING

- *SAS® Documentation Online http://support.sas.com/documentation*
- *SAS Communities https://communities.sas.com/*

CONTACT INFORMATION

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

Jenine Milum
Atlanta, GA, USA
jeninemi@yahoo.com