Agile Text Analytics – Start with the Connections

By Ben Cortese *(Statistician)*

Finding the story within open end responses, social media posts, and other unstructured text data can be a challenge. While necessary to uncover the granular detail, coding exercises are often time and labor intensive. Word clouds provide attractive visualizations, but can lack the ability to identify themes within the text. An emerging solution to text mining, referred to as word networks, merges these two techniques and provides an agile solution while simultaneously removing the human element present in a coding exercise.

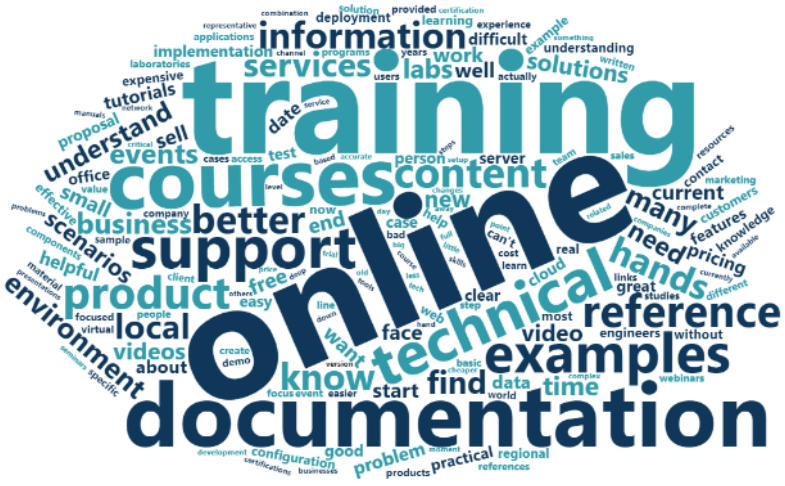
Word networks provide a multi-dimensional visual representation of common themes appearing in text. Clusters of words appear in a web that highlight relationships based on the entire body of text, whether that be open ended responses from a survey, social media posts, interview transcripts, etc. The web is driven by counting the number of times any two words appear close to one another. The stronger the connection, the thicker the link will appear between two words in the visualization. In addition to the connections in a word network, the number of total mentions of each individual word is accounted for via the size of the term in the network. This draws attention to both frequent terms and the most common themes in a body of text.

This technique fits into any agile research framework, developing high level themes from raw text quickly and accurately. Word networks identify the big picture and pinpoint areas worthy of more exploration. They can allow researchers to generate the topline view of a large qualitative project across multiple segments faster, allowing more time to focus on differences and deeper dives.

A case study comparing traditional text analytics to a word network is given below.

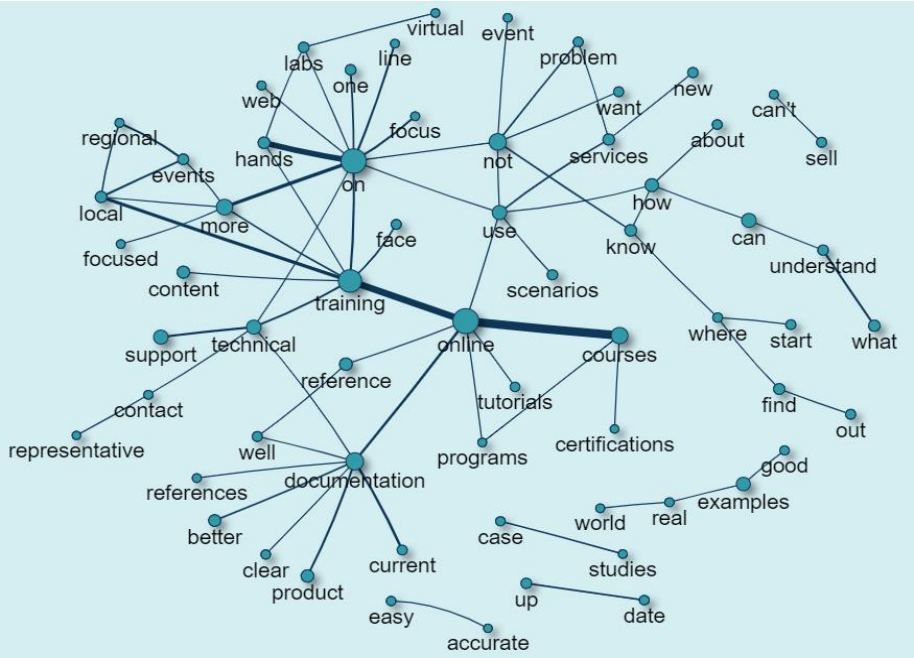
Consider an organization that would like to learn more about how best to familiarize customers with a new product offering. The traditional approach is to provide a series of codes and an associated word cloud. A snippet of such analysis is shown below.

Code	Frequency
Training	198
Technical Support	40
Documentation	32



The coding output is helpful, but requires an individual to read every comment, define codes based on those comments, and assign each comment one or more codes. While an important step in the analysis, this process is time consuming. The word cloud is representative of common terms within the comments, but does not draw any themes or connections from the individual words.

Alternatively, consider the word network visualization run on the same data below.
This output was generated in minutes and did not require reading of a single comment!



The visualization looks similar to a word cloud, but provides much greater detail. The terms are linked together based on their proximity to one another within each comment. The more times terms are used together, the thicker the connection between them.

It is clear from this network that "Online Courses" and "Online Training" are most frequent, along with "More Hands On", and "Local / Regional Events". A closer look indicates that "Real World Examples", "Clear and Current Documentation", and "Technical Support" are also common themes. A deeper exploration into the raw verbatim would complete the story using this top down approach to text mining.

Text analytics – whether it be through coding, word clouds and/or word networks – is a critical part in understanding the full story from the data. Word networks allow researchers to understand a high level version of the story quickly and confidently while more time is spent coding and interpreting word clouds.

About the Author

Ben Cortese has been with KS&R for nearly two years, working with a diverse client group on advanced analytics projects. As an expert R programmer, his focus is on exploratory analysis, Bayesian network modeling, key driver analysis, discrete choice and predictive modeling.

Education: PhD (Statistics), Syracuse University

Return To *Thought Leadership >*