Paper RIV-10

Uncovering Patterns in Textual Data with SAS® Visual Analytics and SAS® Text Analytics

Daniel Zaratsian, SAS Institute Inc.; Mary Osborne, SAS Institute Inc.; Justin Plumley, SAS Institute Inc.

Presented by Meera Venkataramani

ABSTRACT

SAS® Visual Analytics is a powerful tool for exploring data to uncover patterns and hidden opportunities. The challenge with the growing amount of data is that the majority is unstructured, often in the form of customer feedback, survey responses, social media conversation, blogs, news articles, and other unstructured text. By integrating SAS® Visual Analytics with SAS® Text Analytics, you can uncover patterns in both structured and unstructured data, while enriching and visualizing your data with customer sentiment and categorical flags, and uncovering root causes that primarily exist within unstructured data.

This paper highlights a case study that provides greater insight into unstructured data and demonstrates advanced visualization, while enhancing time to value by leveraging SAS® Visual Analytics high-performance, in-memory technology and SAS® advanced text analytics capabilities.

INTRODUCTION

If you take a look at all of your customer touch-points, all of the opportunities that your business has to interact with customers, what percentage are you leveraging today? A standard response that we hear from clients is somewhere between 1% and 10%. Not only is data collection a problem, but targeting and filtering this data for relevant information is one of the core challenges facing many organizations today. It is essential for organizations to have the tools and technology that enable them to extract pertinent information and visualize it for business decisions and insight.

This paper is relevant to a business audience and has three sections:

Section 1 – The first section highlights “the why” and “the how” of visualizing unstructured data. If your organization is inundated with unstructured data and you are looking to gain a competitive advantage, text analytics plays a primary role in this strategy. Having the ability to visualize and report on your results requires a process that is scalable and automated, leading to better business decisions.

Section 2 – The second section discusses additional use cases and applications of using SAS® Visual Analytics to help explore unstructured data. The goal of this section is to illustrate past success with this technology and to offer ideas around new and innovative ways of analyzing and visualizing your unstructured data. Applications for this technology are not limited to social media, but also include call centers, surveys, document collections, e-mails, adjustor notes, maintenance notes, trouble tickets, and more.

Section 3 – The third and final section demonstrates a case study of the technology and process used to analyze and visualize Twitter content from the 2013 Super Bowl. This case study is relevant in many ways because of the large volume of both structured and unstructured text, and it also mimics unexpected events and changes in “customer behavior” around an event. A good business parallel for this would be a new product launch, media publications, or other viral events that reflect a company’s reputation in a negative or positive light.

WHY IS THIS IMPORTANT?

Unstructured data continues to grow in volume, variety, velocity, and the overall value that it provides to organizations. This trend offers both opportunities and challenges. There has also been a shift in the approach that analysts take to derive value from unstructured data. Depending on your experience, background, and business requirements, you might start with a simple word cloud, progress to text clustering, natural language processing, taxonomies, or even starting to establish complex relationships using ontologies and network graphs. No matter where you are on this spectrum, the end goal is to extract meaningful and actionable information from the raw text. The majority of customers achieve this through some form of visualization.

This topic is also important because the two technologies discussed in this paper—text analytics processes and the subsequent visualization—are naturally symbiotic. On one hand, the best analysis of unstructured data might not be of much value unless the results can be visualized and presented in a way that is easy to understand, regardless of whether the audience will be colleagues, executives, customers, or investors. On the other hand, the data that drives

a good visualization needs to be accurate, correctly formatted, properly constructed, and also relevant in a way that aligns with the business requirements.

HOW TO START VISUALIZING YOUR UNSTRUCTURED DATA

Although the process, technology, and analytical approach used to analyze unstructured data are driven by the business requirements, as a best practice many organizations follow a similar pattern. This pattern is illustrated below in **Figure 1**.

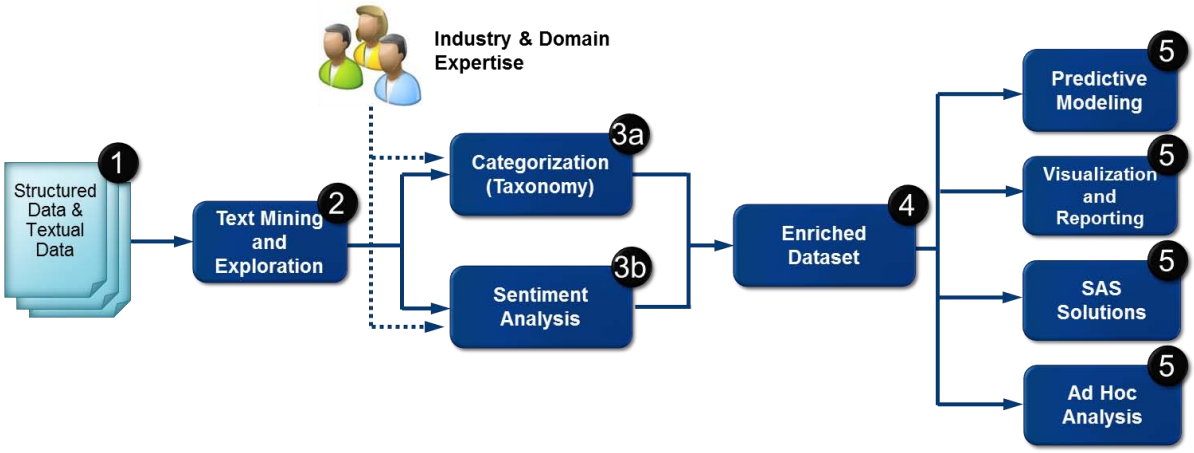

Figure 1. SAS Text Analytics Process Flow Diagram

Aside from defining the business requirement(s), the first step in the analytics process is information retrieval. In many cases, organizations have both internal and external data which they wish to aggregate and analyze. In addition, this data typically contains both structured and unstructured data. A good example of this is in the financial services industry, where a banking customer might have call center notes from financial advisors, transactional data, demographic data, and might also be looking to integrate online social data. SAS has the ability to collect and aggregate this information from disparate sources using a combination of SAS/ACCESS® engines and SAS® Crawler, that provide web crawlers, RSS feed crawlers, file crawlers, and pluggable crawlers for Twitter, Facebook, Google, Bing, YouTube, and others.

The challenge with unstructured data is that it is free-form, opening the door for misspellings, abbreviations, emoticons, misusing terms (for example, writing *their* instead of *they're*), and many other challenges. In addition, the analyst typically does not know what is contained within the raw text. Therefore, an essential part of this process is to use text mining to explore and extract key elements in the data, relationships, topics, and clusters, all of which greatly benefit from part-of-speech analysis and standardization such as the ability to stem terms or automatically identify misspellings.

Whether your organization is looking to enhance your predictive models with unstructured data, organize and categorize content, assess sentiment, or a combination of these, it is often necessary to take a dual approach to model development. This involves leveraging statistics to uncover "what you don't know" and also a rule-based approach, enabling you to accurately integrate the knowledge of your domain experts and that of your industry experience.

The next step, depending on your business requirements, will be to enrich the original data set by creating new variables that identify sentiment and categorize the unstructured data. This step might be the most time-consuming, but it is one of the most valuable pieces of the process because the technology will help to extract and tag the unstructured data, thereby turning the free-form text into structured data.

The enriched structured data can then be consumed in existing predictive models, used for segmentation or forecasting, and fed into SAS Visual Analytics. At this stage, it's important to know whether the goal of the visualization is analytical exploration or operational reporting. With data exploration, it might be beneficial to have variables to allow for general pattern discovery, where operational reporting might already have a focused scope in mind.

CUSTOMER APPLICATIONS

Before you can generate the most benefit from visualizing text, the data should go through a text analysis process. Text analysis uses powerful tools for unstructured data that have traditionally been used for structured data. What types of text analysis lend themselves to visualization? More than you might think.

As depicted in **Figure 1**, text mining takes place in the early stages of the text analytics process. This part of the workflow enables users to perform interactive discovery on vast amounts of text. It helps analysts to pinpoint key themes that exist in the data, so they can more effectively narrow their focus and filter out the noise.

EXPLORATION

One of the most user-friendly and powerful representations of text is the idea of Concept Linking using SAS® Text Miner. Concept linking is elegant in its simplicity. It quickly shows associations or relationships between terms. The more highly associated terms are, the thicker the line between the terms:

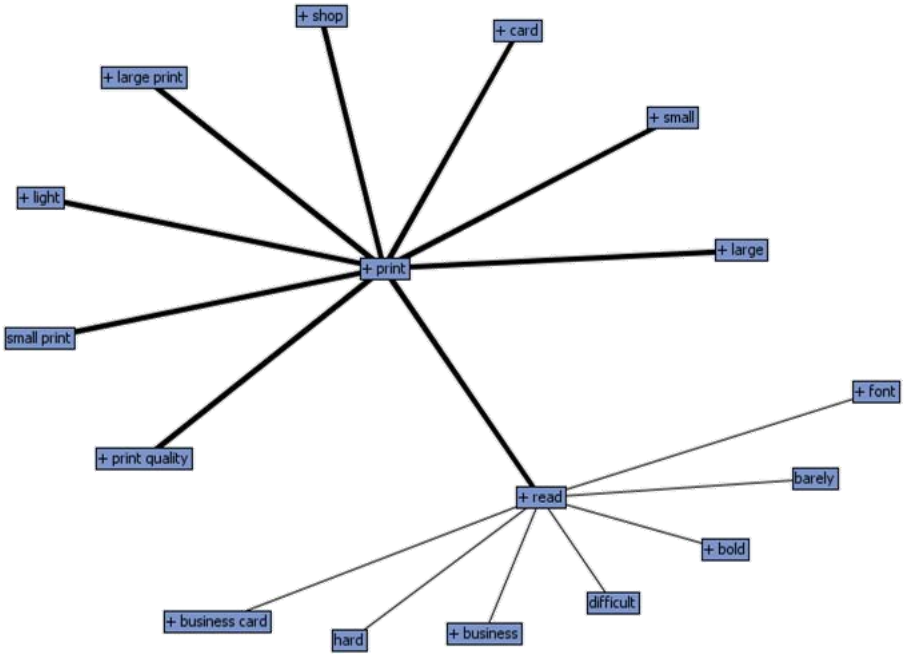

Figure 2. Concept Link Illustrating Term Association for the Word "Print"

CATEGORIZATION

These relationships can aid in the creation of a taxonomy by providing some insight into how terms are used together in a document collection. For example, in the context of document publishing, if the term "Print" is associated with the terms "large" and "small", it might make sense to create a parent node in a taxonomy called **Print** with a child node called **Small**. A taxonomy is a set of categories (and often subcategories) that are typically related in a hierarchy. A sample taxonomy for what we have described can look like **Figure 3**.

Taking this idea a step further: If you are assessing sentiment and are seeing a high number of negative comments related to "Print", the words associated with "Print" could be used to build a root cause analysis. Contextual extraction could be used to identify places where "print" (as a noun) or "font" is used in a sentence with the word "small" (including synonyms) or "large" (including synonyms), and so forth. The end result might look like this:

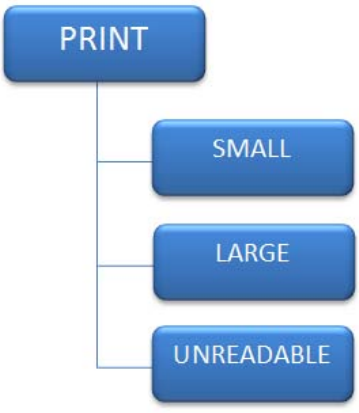

Figure 3. Example of a Taxonomy

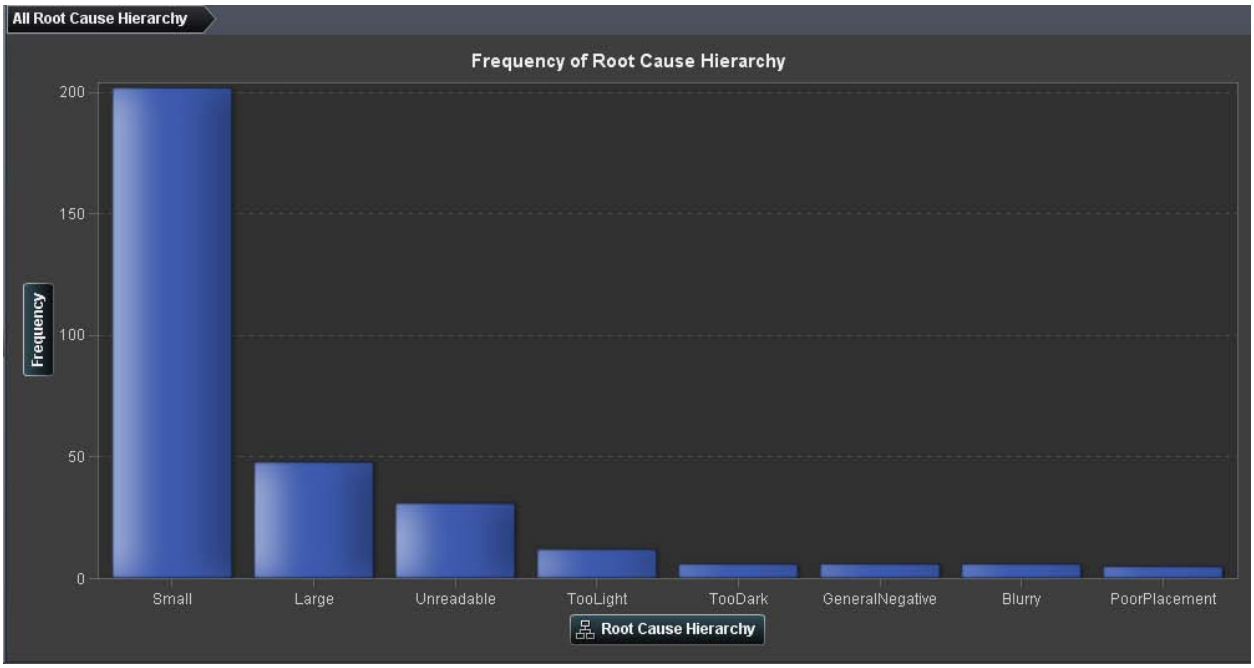

Figure 4a. Root Cause Hierarchy Showing the Distribution of Topics Found within Unstructured Text

This image tells us, at a glance, that “print” (as a noun) and “font” occur in sentences with the term “small” and its synonyms far more often than any of the other levels.

Drilling into the **Small** bar yields a relationship between (a) terms that mean or indicate “print” and (b) words that *describe* “print”.

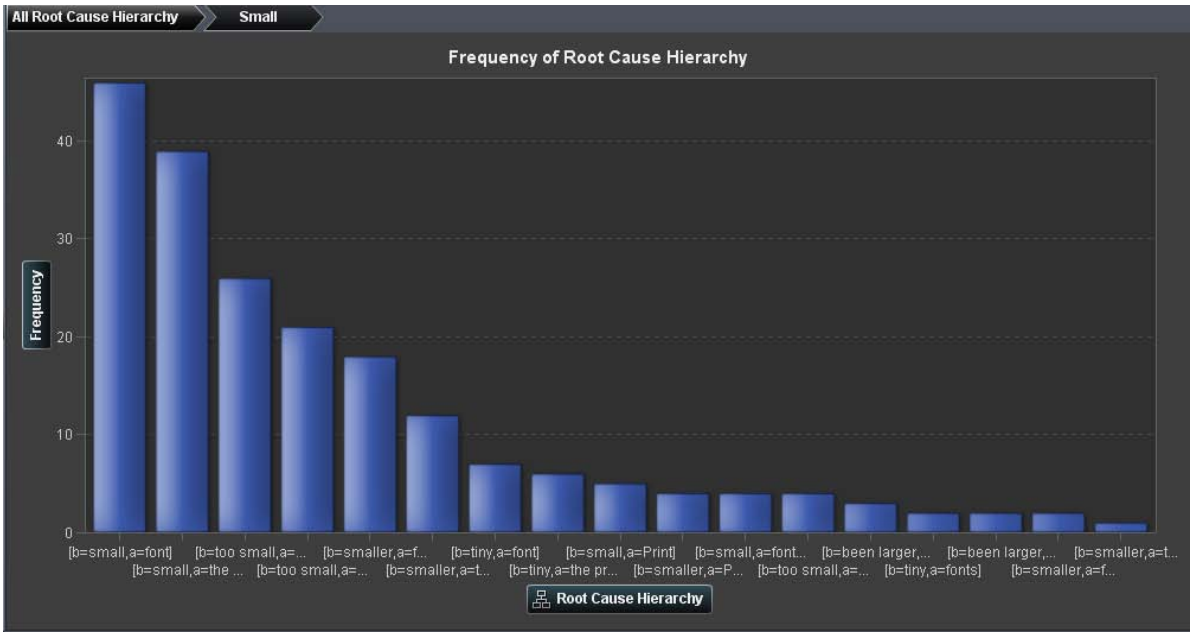

Figure 4b. Hierarchy Drilldown into the Term “Small” to Visualize How Customers Are Using the Word “Small”

Drilling one final time reveals sentence snippets that provide the actual context in which the terms were used.

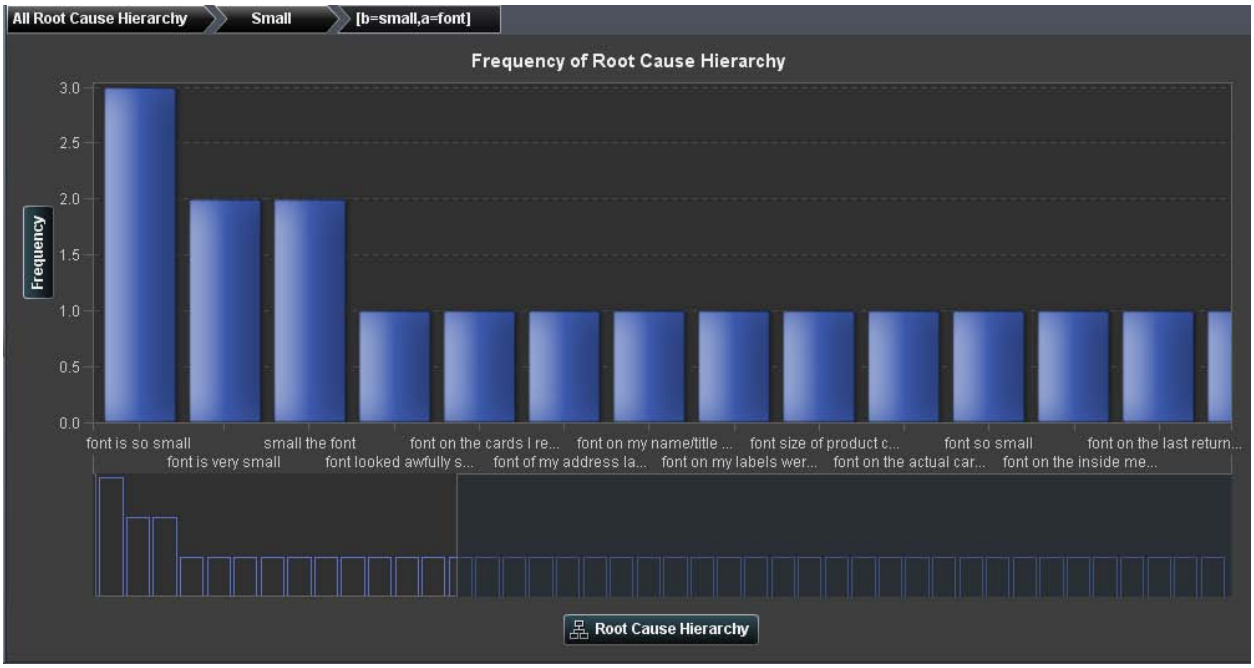

Figure 4c. Hierarchy Drilldown into the Terms "Small" and "Font" to Discover Root Cause

With a hierarchy such as this defined, we can quickly drill from high-level concepts into actual verbatim of customers.

SENTIMENT ANALYSIS

Sentiment analysis provides the backbone to voice of the customer projects. The ability to derive a positive or negative tone from surveys, call center notes, and social media is critical to understanding customer reaction to products or services. Common visuals used for sentiment analysis are red/green bar charts with red indicating negative sentiment and green indicating positive. In the chart below, we can see that **Print** sentiment is more highly negative than positive, and the table below highlights the negative responses with regard to **Print**.

Figure 5. Sentiment Analysis of "Print" Related Customer Reviews

ENTITY EXTRACTION

Categorization and sentiment analysis processes create additional structured information that describes each distinct unstructured text data after taking into consideration the entire text field. For example, the process could tell us that the record belongs to a category like **Print** or that the comment displayed negative sentiment about **Quality**.

Sometimes, however, we want to find information like “What people are named in this record?”, “What product did the comment reference?”, and so forth. For these cases, we can use an entity-extraction process that can enable us to identify instances where an important person, event, or location has been identified. For example, if we want to identify key locations discussed in a corpus of documents and then find the documents associated with a *specific* location, we might use a drill path found in **Figure 6**.

Figure 6. Visualizing and Filtering Extracted Entities

TAXONOMY VALIDATION AND NOISE FILTERING

Beyond exploration of text, root cause analysis, and positive/negative displays of sentiment, there are other areas where visualizing unstructured data can be beneficial.

Taxonomy creation, as a part of content management, can be a time consuming process. It requires an understanding of the document collection so that the documents can be properly categorized. Once key themes are identified using text mining and leveraging domain expertise, the actual creation of the taxonomy can begin. This becomes an iterative process by which the taxonomist builds categories and associated business rules, which then need to be validated. Once the taxonomist is comfortable with the structures in the taxonomy, that taxonomy can be applied to the full document collection through an automated process.

Visualization can aid the taxonomist in developing and validating the taxonomy. Hierarchies that represent the category/subcategory (or parent/child) relationships in the data can be built on the fly, enabling quick drills from one level to the next.

Consider the following taxonomy layout based on scoring financial news documents using the SAS® Industry Taxonomy Rules for Media and Publishing out-of-the-box taxonomy:

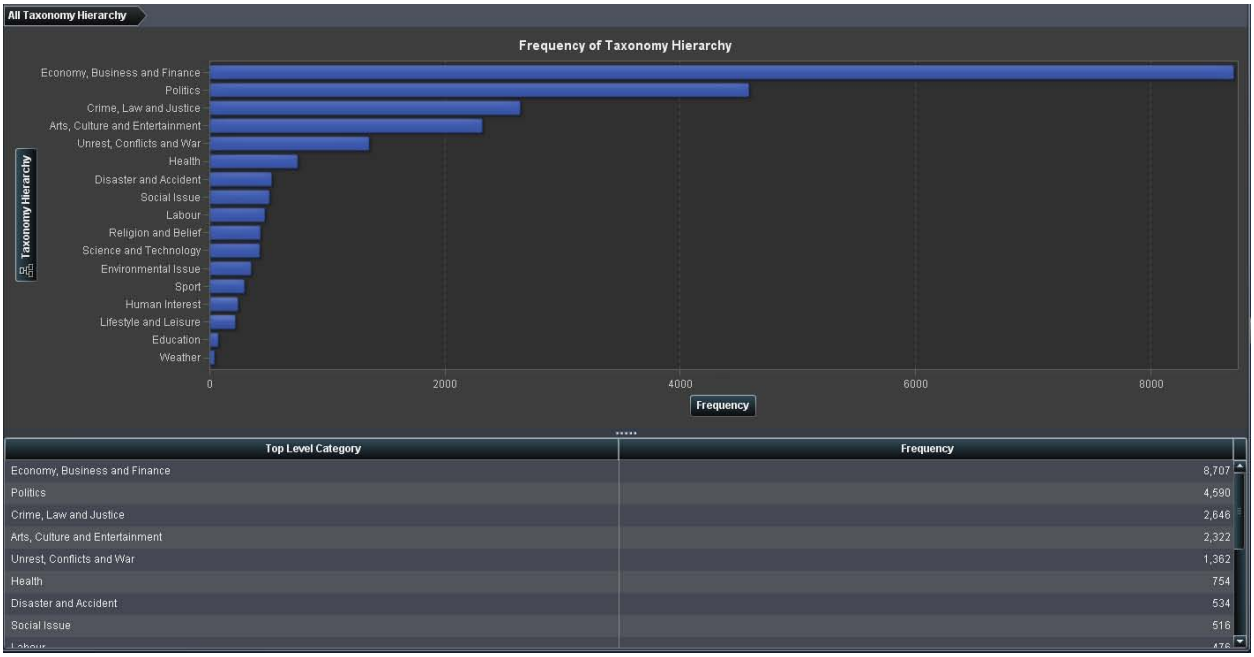

Figure 7. Taxonomy Visualization of the SAS Media and Publishing Hierarchy

There are some documents that are being categorized under **Sport**. As the taxonomist, perhaps we know that we should not have any documents triggering for Sport. Upon further review, there is a rule under Sport for the word "sai". A *sai* is a martial arts weapon. SAI is also the acronym for the Supreme Audit Institution in India. In this case, by scoring the documents against the taxonomy and visualizing the output, we were able to discover an ambiguity in our taxonomy rules that lead to a misclassification.

Consider social data. Social data grows exponentially. Many organizations are capturing social data and storing it in cheap storage environments like Hadoop. Many are doing so without first examining whether the data they are capturing is valuable. Social data is big data. In order to realize value from it, it becomes necessary to separate potentially valuable or relevant content from extraneous data. Fortunately, another use of a taxonomy can help with this task: to filter the relevant from the irrelevant content. Drillable visualizations help business users validate the results and determine that what was classified as "noise" truly was information that could be removed.

CASE STUDY: SUPER BOWL BLACKOUT

To illustrate SAS technology around unstructured data and visualization, we targeted Twitter data pertaining to the 2013 Super Bowl, which was held in New Orleans on February 3, 2013. Super Bowl XLVII was a matchup between the Baltimore Ravens and the San Francisco 49ers.

As you would expect, Twitter users had a wide variety of things to say about the Super Bowl. As part of this case study, SAS collected approximately 4.85 million tweets from January 11, 2013, through February 6, 2013. Because of this broad timespan, our tweets covered not only Super Bowl discussions, but also Super Bowl predictions throughout the playoffs, as well as non-football-related content such as the Beyoncé half-time performance, Super Bowl party chats, and a wide variety of speculation around the impromptu Super Bowl blackout.

The wide variety of conversations, and the nature of unstructured data, required extensive data cleansing and filtering to do the following:

- Remove the irrelevant tweets.
- Identify misspellings and abbreviations. For example, synonyms for the San Francisco 49ers can include SF49, niners, 49ers, and 49er.
- Focus the analysis in order to visualize and report on key insights.

Identifying and grouping misspellings, synonyms, and abbreviations enriches the clustering results, eliminates noise, and aids in the discovery of emerging topics. This process is iterative and might require input from subject matter experts, but the end result is an industry- and company-specific mapping of key terms and phrases that reduces noise to expose insightful topics buried within your data.

We followed a similar process, as illustrated in **Figure 1**, by starting with text mining and data exploration. Our first goal was to uncover misspellings and abbreviations within the data, ultimately helping to identify emerging topics through the Super Bowl data set. One area that we chose to focus on as part of this analysis is the Super Bowl blackout. This occurred a few minutes into the second half of the Super Bowl. At approximately 8:37 PM EST, the lights on one-half of Mercedes-Benz Superdome suddenly went out. This case study will investigate and visualize the viral tweets and categories discussed throughout this blackout.

First, using SAS Text Miner, topics and term/phrase associations were extracted to help understand the common themes across the Twitter subset of blackout chats. The key topics were:

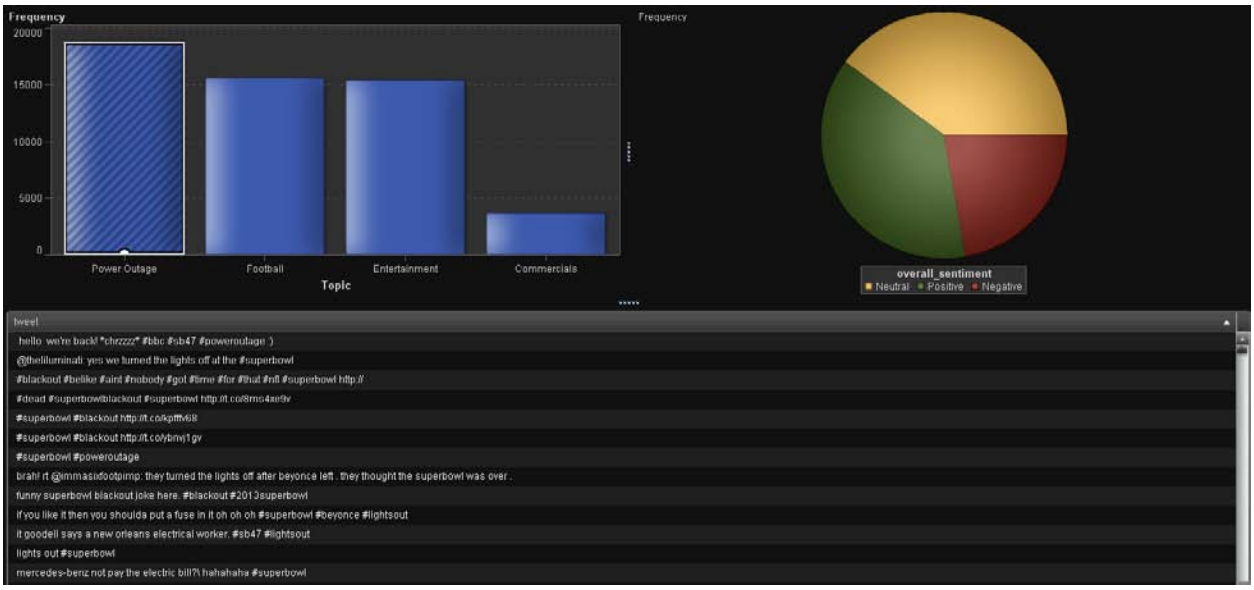

Figure 8. High-Level Categories Found within the Tweets Pertaining to Super Bowl Blackout

To build on these key topics, we also wanted to use network analysis to visualize the relationship between topics and the most influential Twitter users (in this case, “Influential Twitter users” are based on the number of followers they have, the frequency of Tweets and ReTweets, and the overall frequency of Super Bowl-related Tweets). There are several ways to generate network graphs using SAS, these include SAS/GRAPH®, SAS® Customer Link Analytics, and SAS® Social Network Analysis solutions.

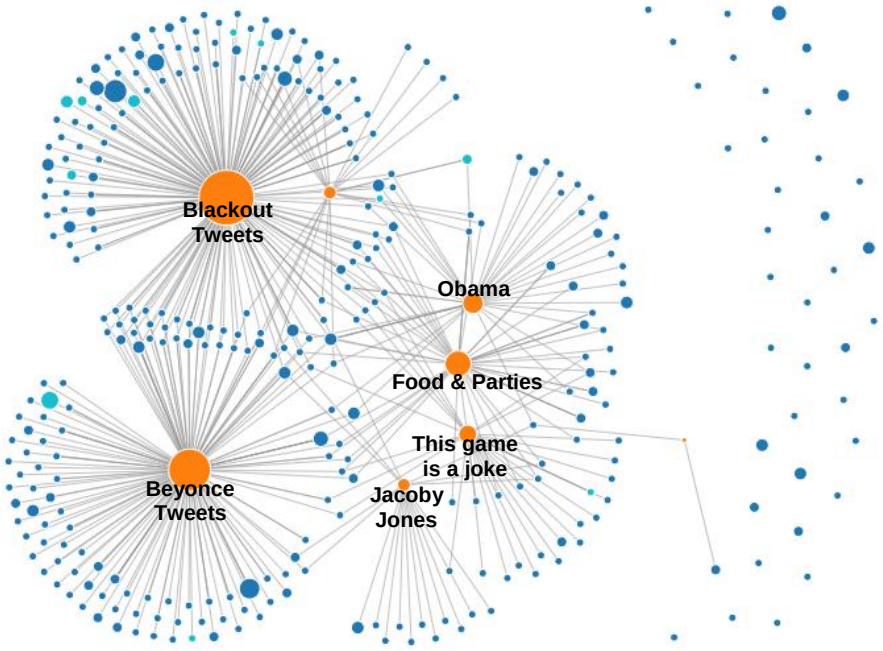

Figure 9. Network Diagram of Influential Twitter Users/Authors (Blue Dots) and Topics (Orange Dots)

The analysis and exploration of the blackout tweets exposed several significant behaviors of twitter users. The second half of the Super Bowl started at approximately 8:30 PM EST. As many of you might remember, the 49ers kickoff was returned by Jacoby Jones for a record-breaking 108-yard kickoff return and a Baltimore Ravens touchdown. On Twitter, it was also interesting to discover emerging topics where people were saying “this game is boring,” “game over!”, and even one tweet came across Twitter at 8:33 PM EST saying “Turn off the Lights #Superbowl #NFL.” This is an interesting coincidence, because nearly 5 minutes later, half of the lights went dark!

The blackout lasted nearly 35 minutes. During this time, a wide variety of conversational topics took place on Twitter. These topics, the authors who promoted the topics, and the associations across topics are all visualized in the network graph shown in **Figure 9**. Two very nimble and intelligent marketing departments took advantage of the blackout by creating an on-the-spot social media marketing tactic. These two marketing tactics came from Oreo and Walgreens. Oreo took advantage of this opportunity and, within 11 minutes, they were able to launch a very clever marketing campaign as shown in **Figure 10**.

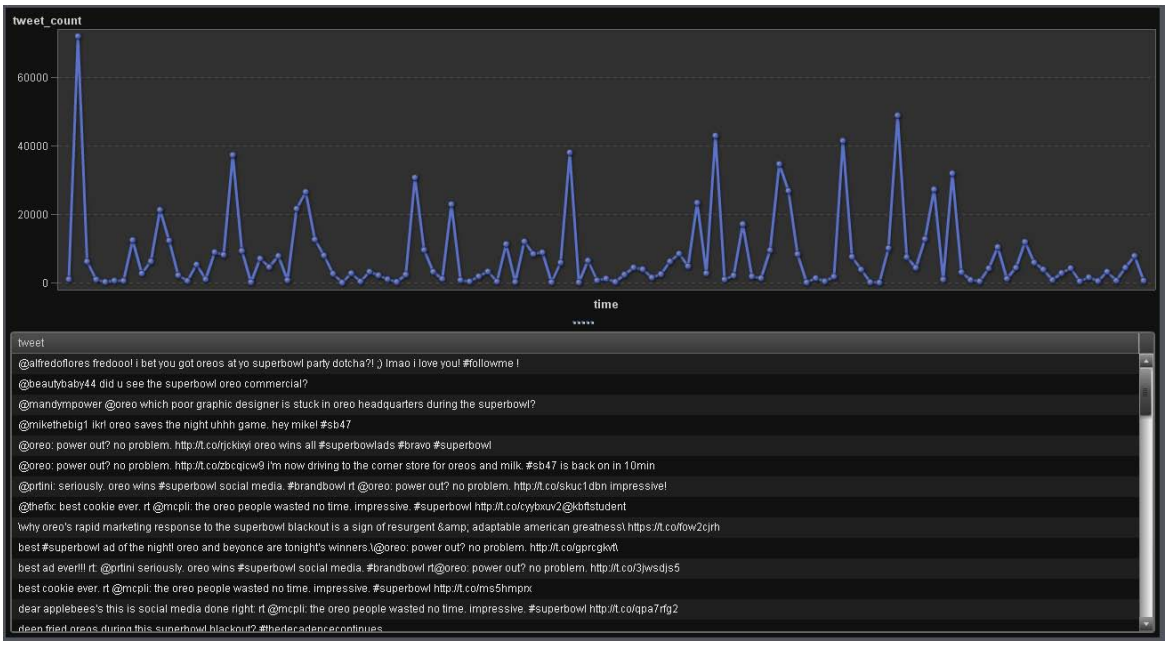

Figure 10. Strategic Marketing Tweets from Oreos During the Super Bowl Blackout

Although Walgreens did not come up with a visual, they took only 7 minutes to respond and tweet "We do carry candles" as shown in **Figure 11**. This tactic generated several retweets and also drove several post-Super Bowl articles discussing their marketing efforts, especially beneficial for Walgreens who didn't have to pay 3.8 million for a Super Bowl advertisement and still received good media coverage.

Figure 11. Strategic Marketing Tweets from Oreo and Walgreens During the Super Bowl Blackout

CONCLUSION

As a growing number of organizations are realizing the value of unstructured data, visualization technology makes text analysis more accessible to a larger audience than ever before. Unstructured data analysis can be a daunting task. Having the technology to visualize the results expands the reach and insights of the analysis. When thinking about visualization, go beyond the traditional structured data and consider the value that graphs, charts, and other visuals bring to your call center notes, trouble tickets, surveys, and social data. Unstructured data and its value are growing exponentially; companies that leverage this technology will have the advantage.

REFERENCES

- SAS Visual Analytics: http://www.sas.com/software/visual-analytics/data-visualization.html
- SAS Text Analytics: http://www.sas.com/text-analytics/
- SAS Text Miner Knowledge Base: http://support.sas.com/software/products/txtminer/index.html
- SAS Sentiment Analysis Knowledge Base: http://support.sas.com/software/products/san/index.html
- SAS Content Categorization Knowledge Base: http://support.sas.com/software/products/ccs/index.html

RECOMMENDED READING

White paper: “Sifting Through the Noise of Social Media”. Available at http://www.sas.com/reg/wp/corp/47188

CONTACT INFORMATION

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

Dan Zaratsian
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
Work Phone: (919) 531-8862
E-mail: dan.zaratsian@sas.com

Mary Osborne
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
Work Phone: (919) 531-2765
E-mail: mary.osborne@sas.com

Justin Plumley
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
Work Phone: (919) 531-8573
E-mail: justin.plumley@sas.com