Paper SAS061-2014

Uncovering Trends in Research Using Text Analytics with Examples from Nanotechnology and Aerospace Engineering

Tom Sabo, SAS Federal LLC

ABSTRACT

Understanding previous research in key domains can help R&D organizations focus new research in non-duplicative areas and ensure that future endeavors do not repeat the mistakes of the past. However, manual analysis of previous research efforts can prove insufficient to meet these ends. This paper highlights how a combination of SAS® Text Analytics and SAS® Visual Analytics can deliver the capability to understand key topics and patterns in previous research and how this combination applies to a current research endeavor. We will explore these capabilities in two use cases. The first will be in uncovering trends in publicly visible, government-funded research (Small Business Innovation Research or SBIR) and how these trends apply to future research in nanotechnology. The second will be visualizing past research trends in publicly available NASA publications, and how these might impact the development of next-generation spacecraft.

INTRODUCTION

Organizations have an interest in understanding past research trends for a variety of reasons. They may be interested in staffing their own organization, government or otherwise, to have in-house subject matter expertise to deal with an influx of patent or other requests associated with a particular emerging technology. Or they may be interested in ensuring that future research investments in a given technology area, such as nanotechnology, are non-duplicative. They may be interested in leveraging previous research performed in a given area, and would be interested in the means to quickly assess a number of related questions regarding those past research trends.

However, these organizations often resort to manual analysis of research, due to a lack of text analytics technology or the knowledge to apply it to research-oriented tasks. Manual analysis has its limitations. First, it's best for producing qualitative results, which can't help you put actual data points in the proper context. Second, manual analysis takes an extraordinary amount of time and resources, which the organization may or may not have[1].

This is where SAS can step in, enabling a combination of SAS Text Analytics and SAS Visual Analytics to deliver the capability to understand key topics and patterns in previous research and how they apply to a current research endeavor. In this paper, we will explore a repeatable framework for research analytics including the technology involved in this solution as well as a five-step process for generating and using this framework. We will explore this through two examples. The first will be in uncovering trends in publicly visible, government-funded research (SBIR) and how these trends apply to future research in nanotechnology. The second will be visualizing past research trends in publicly available NASA publications, and how these might impact the development of next-generation spacecraft.

The five-step process for generating and using the framework for research analytics is as follows:

1. **Data acquisition and preparation for text analytics**: Data is acquired for each of our examples through web interfaces and is converted into a SAS data set using SAS® Enterprise Guide®.
2. **Text analytics**: For each of our two examples, we will take different approaches to text analytics. For our SBIR example, we will apply text topics and promotion of applicable topics to categories using SAS® Contextual Analysis. For NASA, we use text clustering and text topics algorithms from SAS® Enterprise Miner® and SAS® Text Miner.
3. **Data preparation for visual analysis**: Both text analytics approaches generate data sets that are prepared for ingestion into SAS Visual Analytics using SAS Enterprise Guide.
4. **Ad hoc exploration**; This is accomplished with SAS Visual Analytics.
5. **Interactive report generation and use**: This is also accomplished with SAS Visual Analytics.

In the following sections, we will take a detailed look into the business value that is generated by each of our examples, as well as explore the implementation of the five-step process outlined above.

UNCOVERING NANOTECHNOLOGY TRENDS IN SBIR/STTR RESEARCH

As discussed in the previous section, organizations have an interest in understanding past research for a variety of reasons, and would benefit from the means to quickly assess a number of related questions regarding those past research trends. All these requirements would benefit from a framework for research analytics. For this example, we chose the SBIR/STTR program to obtain research data from, because it is highly applicable to government agencies that seek to leverage research funded by other government agencies. The Small Business Innovation Research (SBIR) program is a highly competitive program that encourages domestic small businesses to engage in Federal Research/Research and Development (R/R&D) that has the potential for commercialization. Through a competitive awards-based program, SBIR enables small businesses to explore their technological potential and provides the incentive to profit from its commercialization[2].

In particular, we will explore leveraging this framework for the following questions:

- What are the applications of research in a given domain, such as nanotechnology? Can we visualize this over time and locations? Where has research been increasing or declining?
- Within nanotechnology, how are terms and concepts interrelated?
- Once we determine these current trends in nanotechnology, can we extract these reliably from up-to-date data?
- Can we explore this research in an ad hoc manner, by category, program, and geography?

DATA ACQUISITION

Obtaining the data for the SBIR/STTR conversion was a straightforward process from www.sbir.gov. On the **AWARDS** tab, SBIR/STTR provides an easy-to-use interface to obtain all research in the past 20 years related to nanotechnology as a CSV file. We limited our search to abstracts and titles that included the term “nano” as a sub-segment of any searchable word. This gave us a data set of 2416 rows and 26 columns of data. The columnar data was mostly structured, but also included a title and abstract section. A quick conversion in SAS Enterprise Guide from the CSV file that we obtained from the SBIR site gave us the SAS data set ready for text analytics within SAS Contextual Analysis.

TEXT ANALYTICS – APPLYING SAS CONTEXTUAL ANALYSIS

We registered the SAS data set using SAS® Management Console, then were able to select and load the data set within the SAS Contextual Analysis interface. This enabled us to select the desired column (abstract) for text analytics and extract terms and topics from the data set.

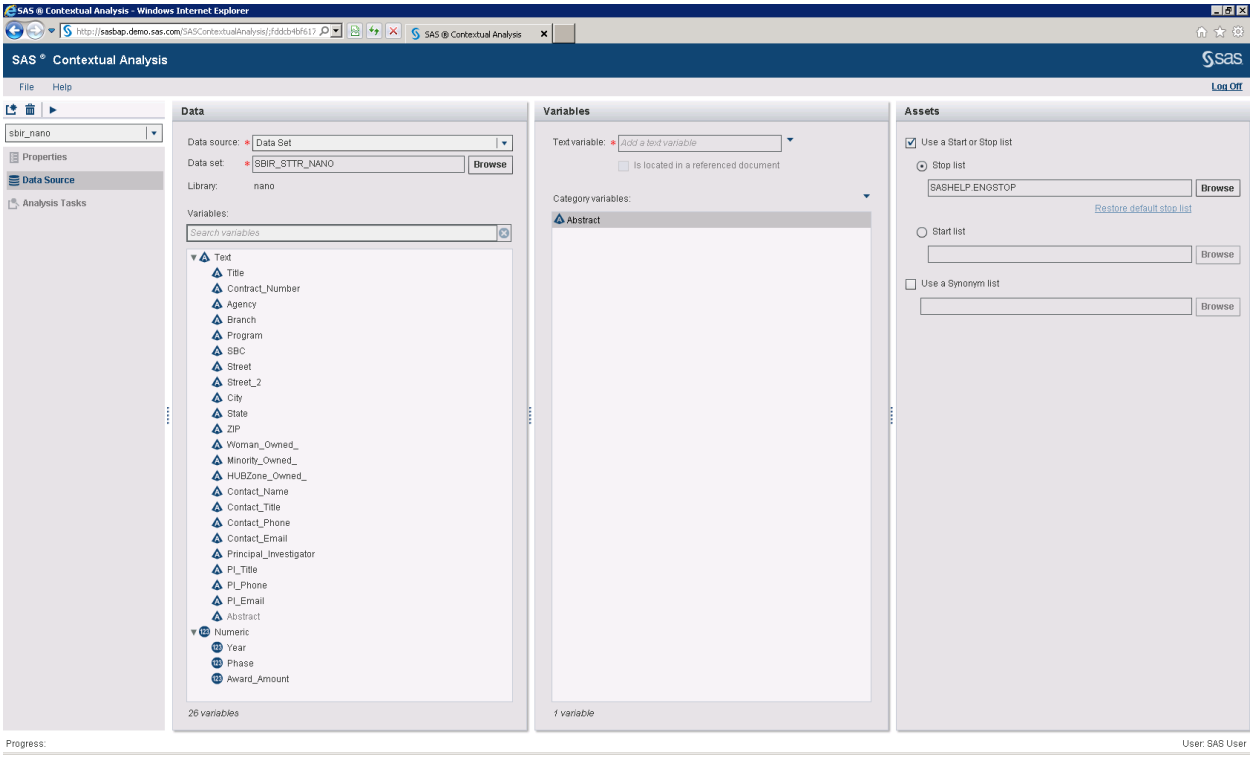

Figure 1. Selecting a Data Set for Term and Topic Extraction Using SAS Contextual Analysis

Topic extraction generated a variety of relevant topics related to nanotechnology by highlighting the top five terms for each topic, then also enabling drilldown via the provided word cloud into other terms relevant to the topic. In Figure 2, the terms and relevant documents associated to a cancer treatment and diagnosis topic are highlighted in the right-most and bottom-most panel. Also note the other topics that came out of this text analysis, related to energy storage, composite fibers, microelectronics, protective coatings, fuel and propulsion, and chemical/biological sensing technology. In each case, the top five terms give a good indication of the research covered by that topic.

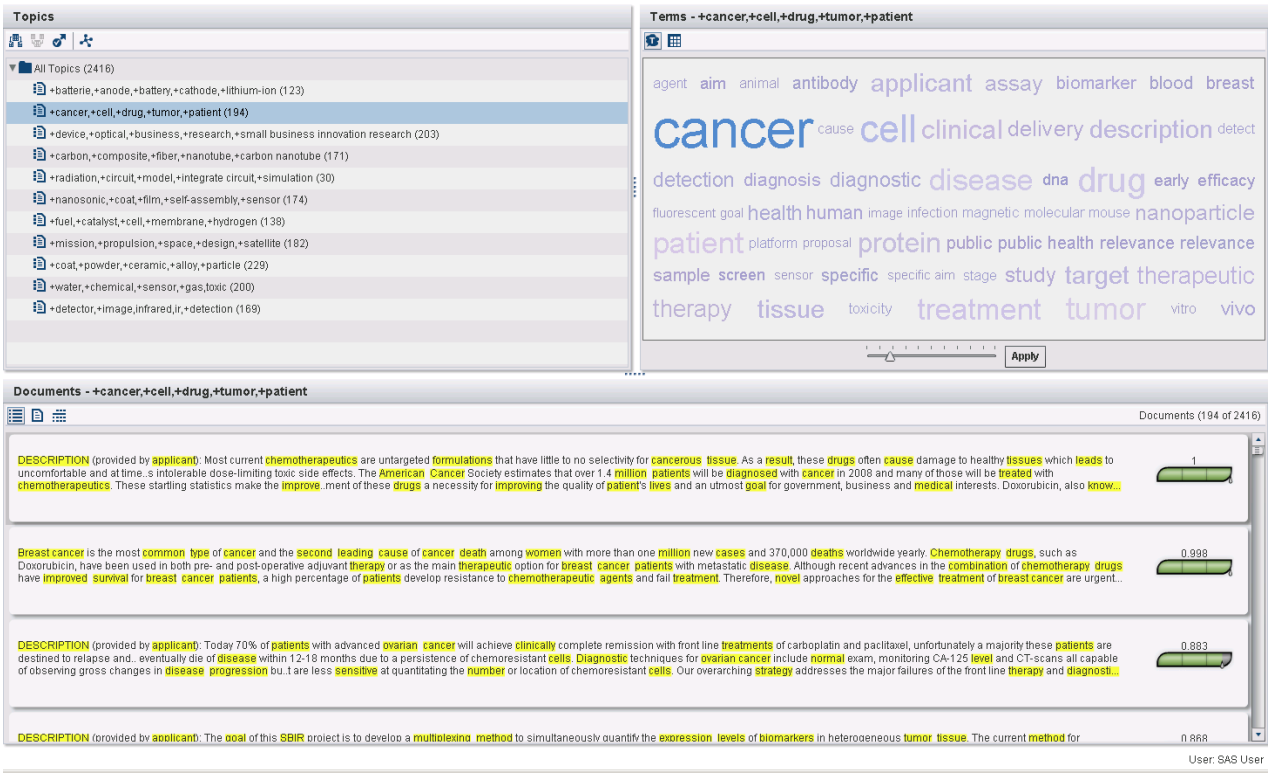

Figure 2. Topic Extraction Results, Highlighting Cancer Treatment and Diagnosis

It is valuable from the standpoint of understanding a topic to visualize how the terms within the topic are interrelated. To further explore a topic, SAS Contextual Analysis builds a term map of the topic. This enables us to explore, as shown in Figure 3, how key terms related to cancer and therapies are applied and co-occur in the topics. So, for instance, the presence of the word “tumor” alone indicates a topic match, or the presence of the terms “cancer” and not the term “manufacture.”

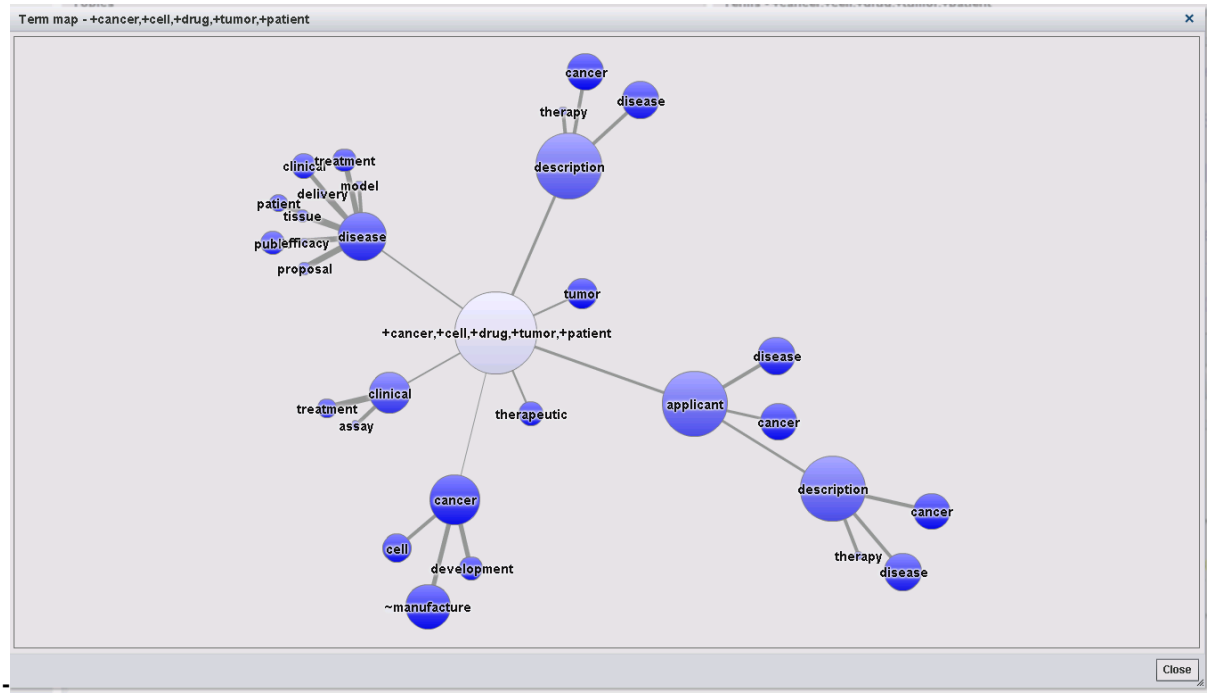

Figure 3. Term Map Related to the Topic Surrounding Cancer Treatment

The next step in the text analytics process is to promote select topics from the topic extraction into categories. Category promotion assists our approach in three ways.

1. It automatically provides us textual membership rules that denote category membership. This is different from topics because there is now a concrete method to determine whether new documents fit a given category.
2. These automatically generated rules can be modified to include subject matter expertise, customized and split to differentiate key topics (such as general therapies from cancer therapies), or to include stand-alone custom rules independent of those generated from the topics.
3. The categories can be used to score new incoming data sets.

As part of this exercise, we promoted 10 relevant topics to categories, including one topic related to spacecraft, which was split into topics related to satellite and vehicle propulsion and spacecraft system design. We could then score all of our documents that originated from topic extraction to determine how well the rules performed, as noted in Figure 4, which showcases all the generated categories and focuses in on the rule performance for spacecraft system design. These auto-generated rules can also give us an indication of how broad our topic is, and whether it may need to be split into multiple topics to differentiate.

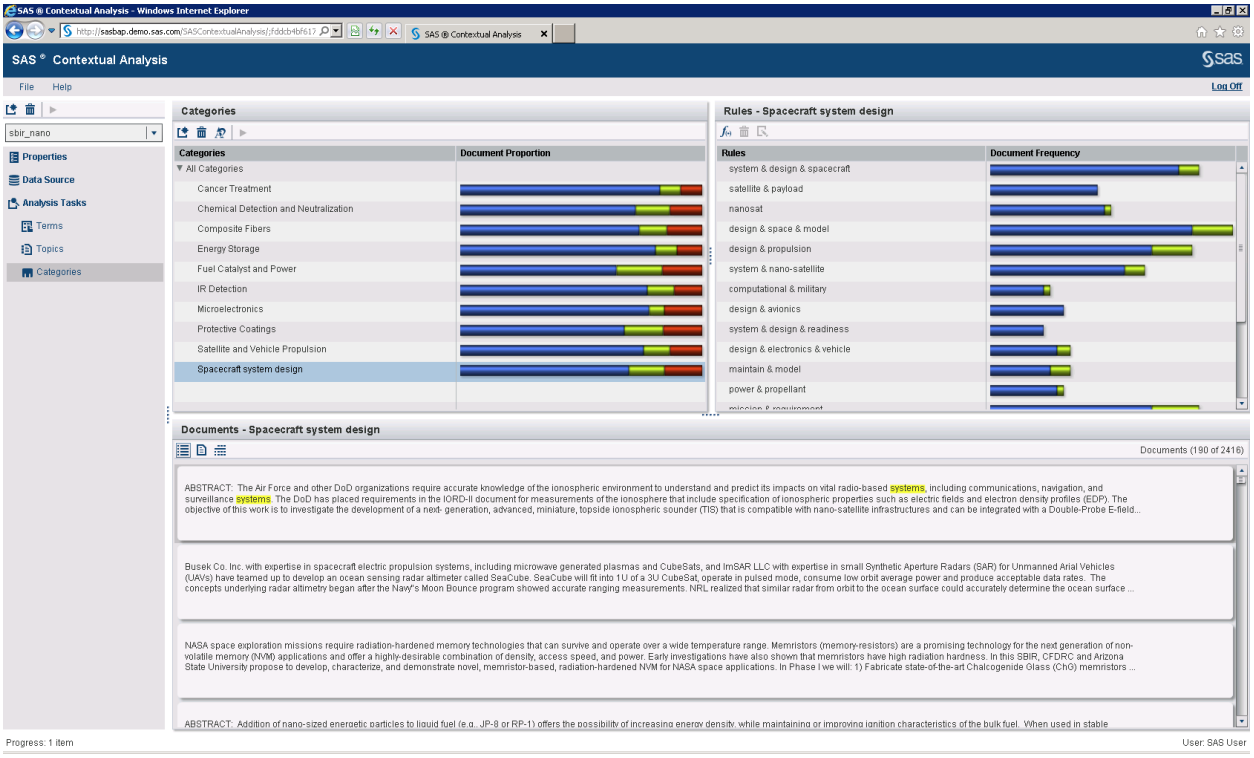

Figure 4: Rule Performance of Categories Generated from Topics

DATA PREPARATION FOR VISUAL ANALYTICS

SAS Contextual Analysis produces a number of SAS data sets in the file system for use in downstream analytics or visualizations—or, in this case, visual analytics. For the purposes of this example, we take one such data set (doc_category_rule_ds), which includes our original data set of 26 variables plus all the instances where a category or individual category rule fired upon a given document abstract. We utilized SAS Enterprise Guide just to include binary variables that denoted category membership, and also relabeled the categories appropriately. (See Figure 5.) This produces the final SAS data set, which will be used in SAS Visual Analytics.

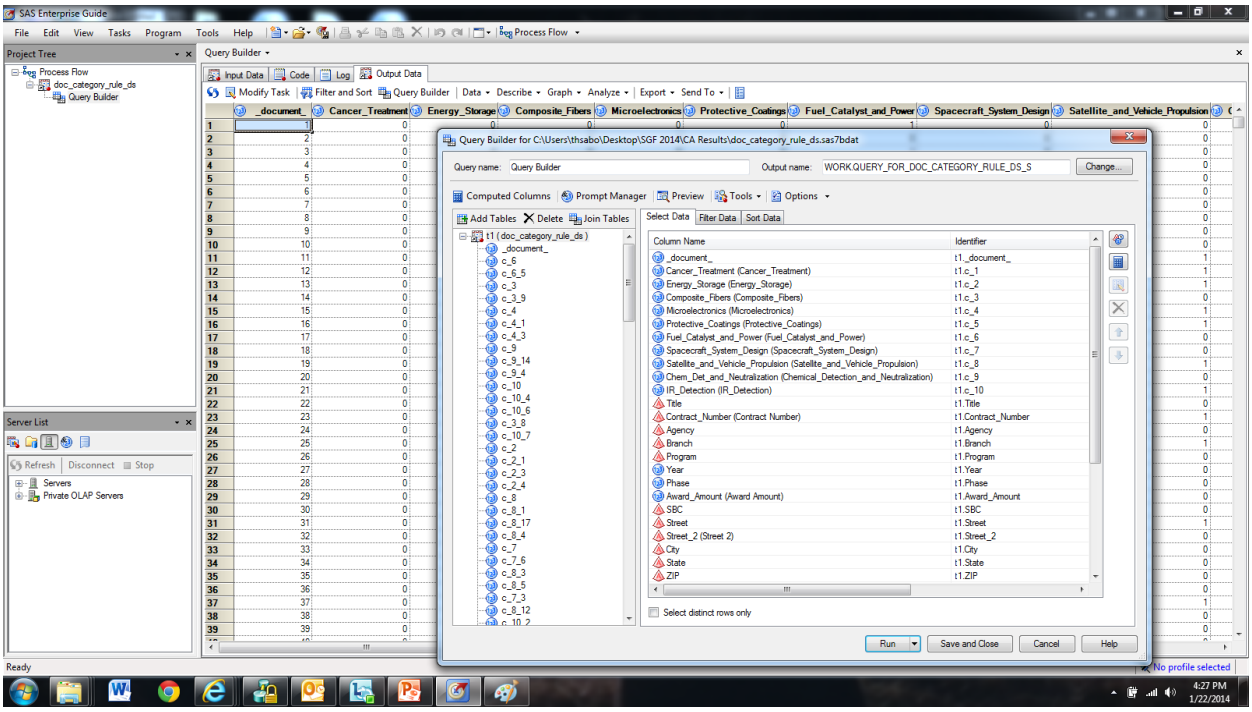

Figure 5: Using SAS Enterprise Guide for Data Preparation Leading into SAS Visual Analytics

AD HOC EXPLORATION

Given the new flat-file, we can use SAS Visual Analytics for exploration of the data. In particular, we may be interested in visualizing how these newly extracted categories of nanotechnology research span time and geography. Figure 6 depicts five of these categories over time. In particular, note how we can interactively filter out years that had a minimal amount of related data, and note how certain categories, such as Spacecraft System Design are trending up over time, while others such as Protective Coatings are trending downwards. Others remain somewhat flat after an earlier spike, such as Chemical Detection and Neutralization. This trend analysis serves to indicate where to dig deeper to investigate rising trends in nanotechnology.

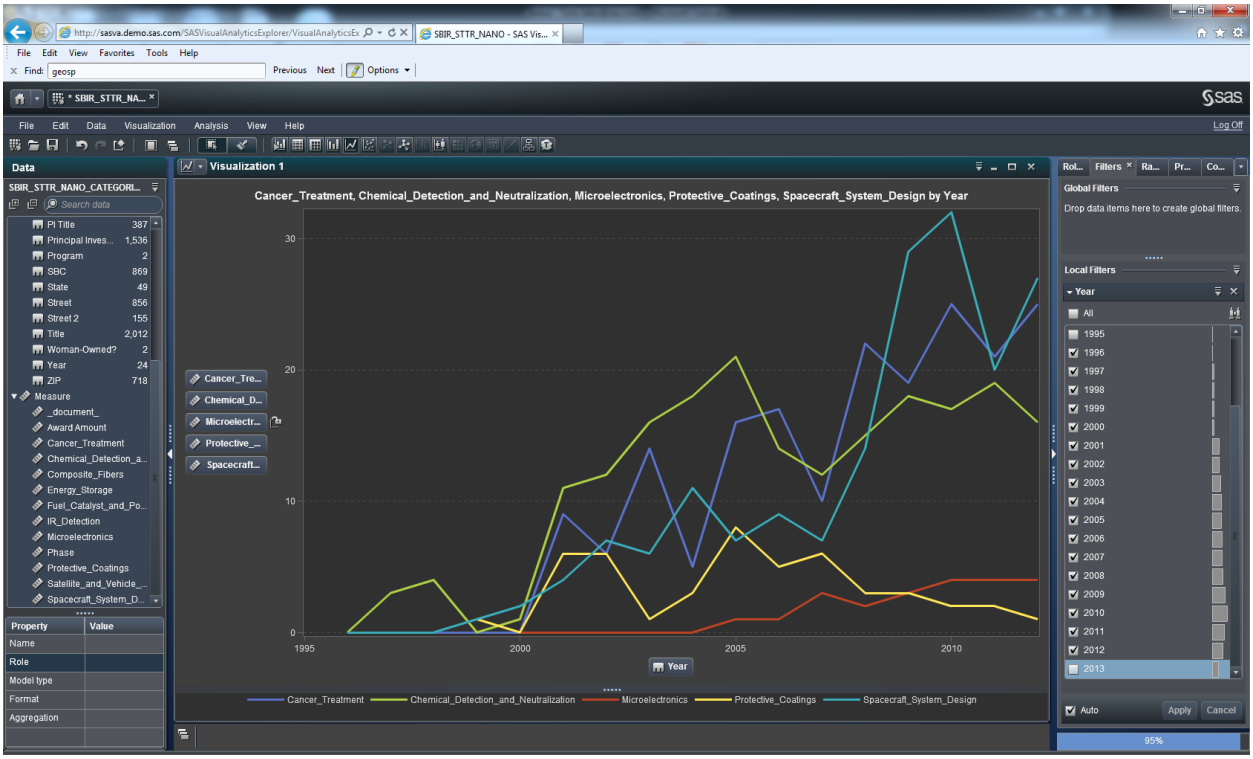

Figure 6: Nanotechnology Trends Depicted over Time

We could also explore how the nanotechnology categories span geospatially. We generated a hierarchy that enables dynamic exploration and drilling from the state to ZIP code level, so that analysts could explore the distribution of research at a higher level, and drill down into any state of interest. Figure 7 depicts the research at the state level. Figure 8 enables us to drill down into the state of Virginia to explore an anomalous amount of the SBIR-STTR funded research that occurs at one ZIP code. This ZIP code accounts for approximately 10% of the SBIR/STTR funded nanotechnology research in the past 20 years! Finally, we can dynamically filter by the ZIP code 24136 in Figure 9 to explore the research trends that occur at one facility, primarily involving protective coatings, but also related to spacecraft system design, satellite propulsion methods, fuel, and composite fibers. Similarly, we could generate research profiles of particular organizations just as easily.

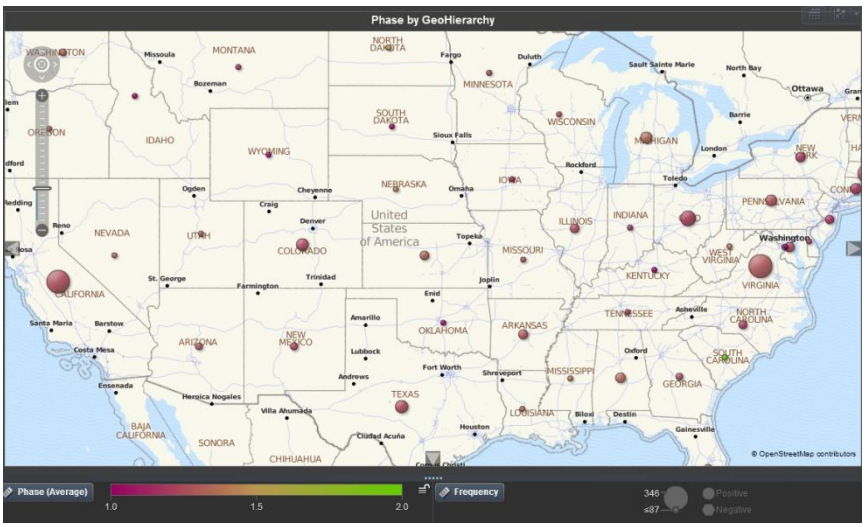

Figure 7: SBIR-STTR Nanotechnology Research Distributed Geospatially

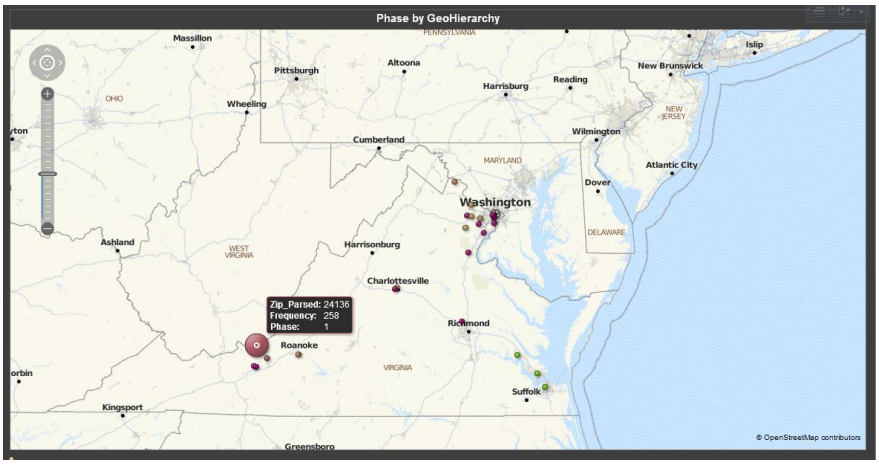

Figure 8: Drilldown into State of Virginia to See Anomalous Amount of Research at a Particular ZIP Code

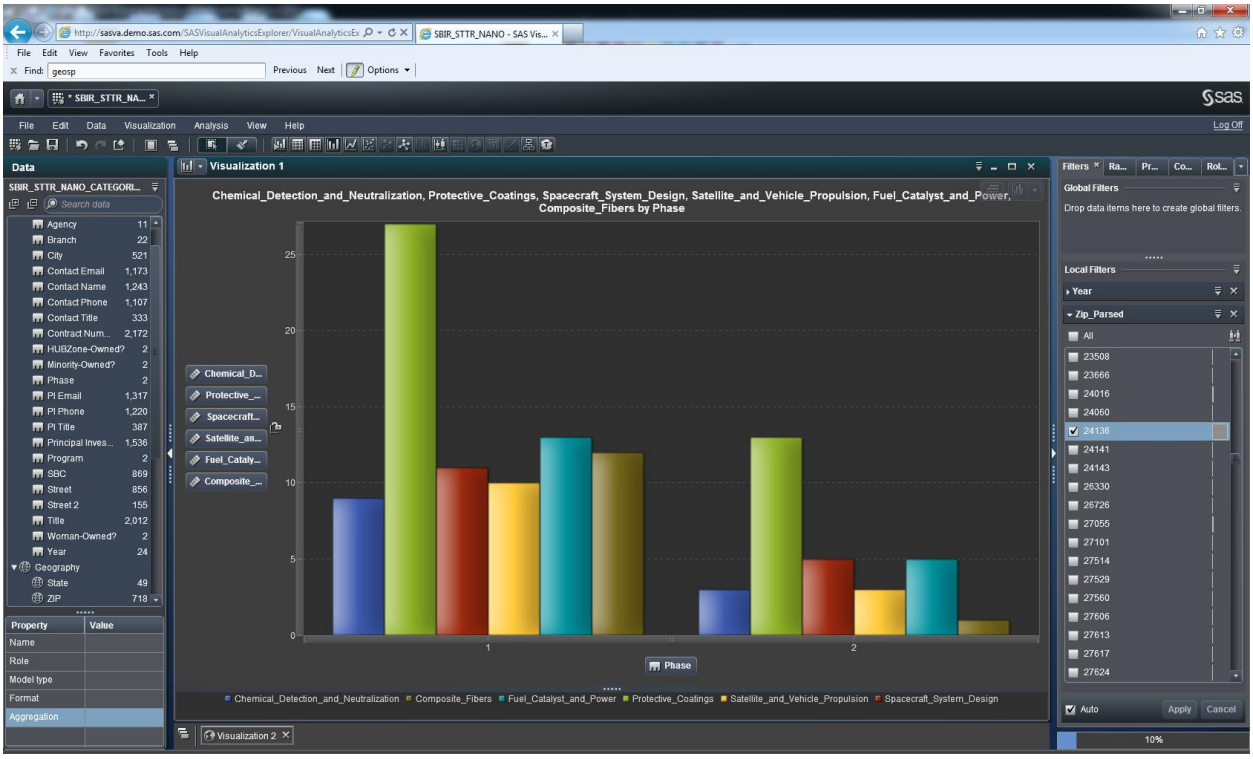

Figure 9: Research Profile of Virginia ZIP Code

INTERACTIVE REPORT GENERATION AND USE

One question remains—can we report on this research in an ad hoc manner, by category, program, and geography? We will undertake this question through the use of the SAS Visual Analytics interactive reporting capabilities.

In particular, we want to be able to explore the data in the report by interactive filtering. We could filter first by the program (SBIR/STTR), then by funding agency, including DOD, HHS, and NASA. We could explore the data geographically and then depict the trends in the data over time, enabling us to explore data in a particular time period. These interactive exploration objects and filters are very straightforward to set up in SAS Visual Analytics. For instance, in Figure 10, we have set up a dashboard to explore all the relevant research to nanotechnology in the SBIR program funded by the DOD and occurring in California. In Figure 11, we use this same dashboard to visualize all the research funded by NASA. Finally, in Figure

12, we showcase all the data funded by DOD and NASA occurring in Virginia, particularly focused on the one ZIP code listed earlier. As a DOD agency researcher or NASA researcher, I may be interested in leveraging research funded by the other organizations outside my own for the purposes of funding further non-duplicative research, or building upon the results of previous research efforts and funding new initiatives appropriately with this foresight.

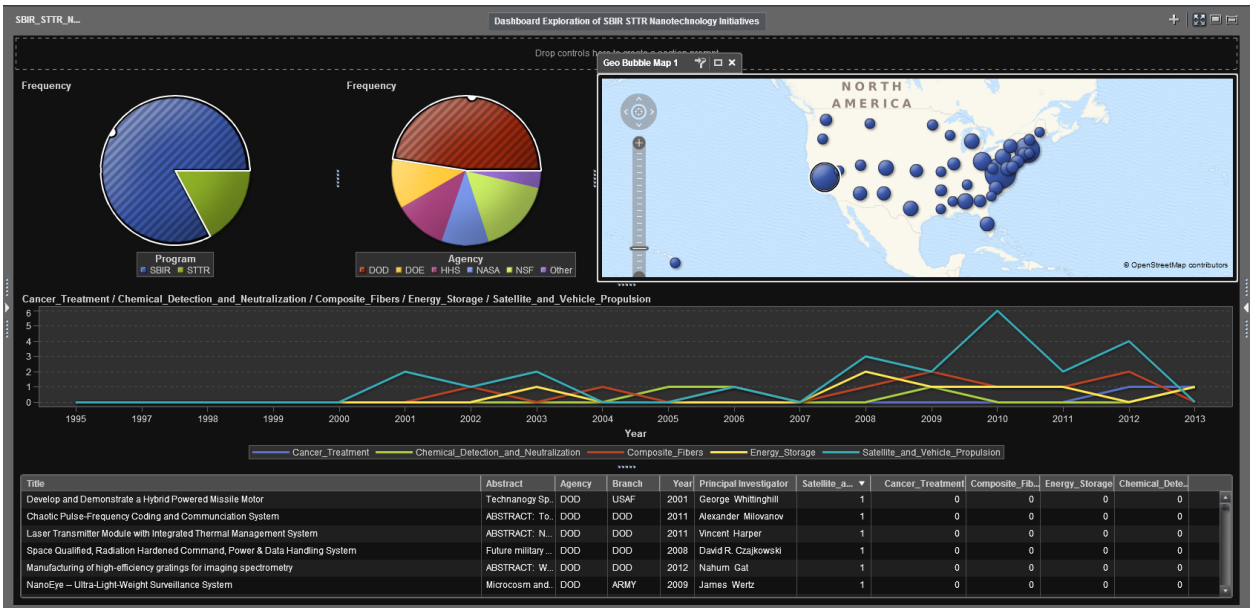

Figure 10: DOD Funded SBIR/STTR Nanotechnology Research Occurring in California

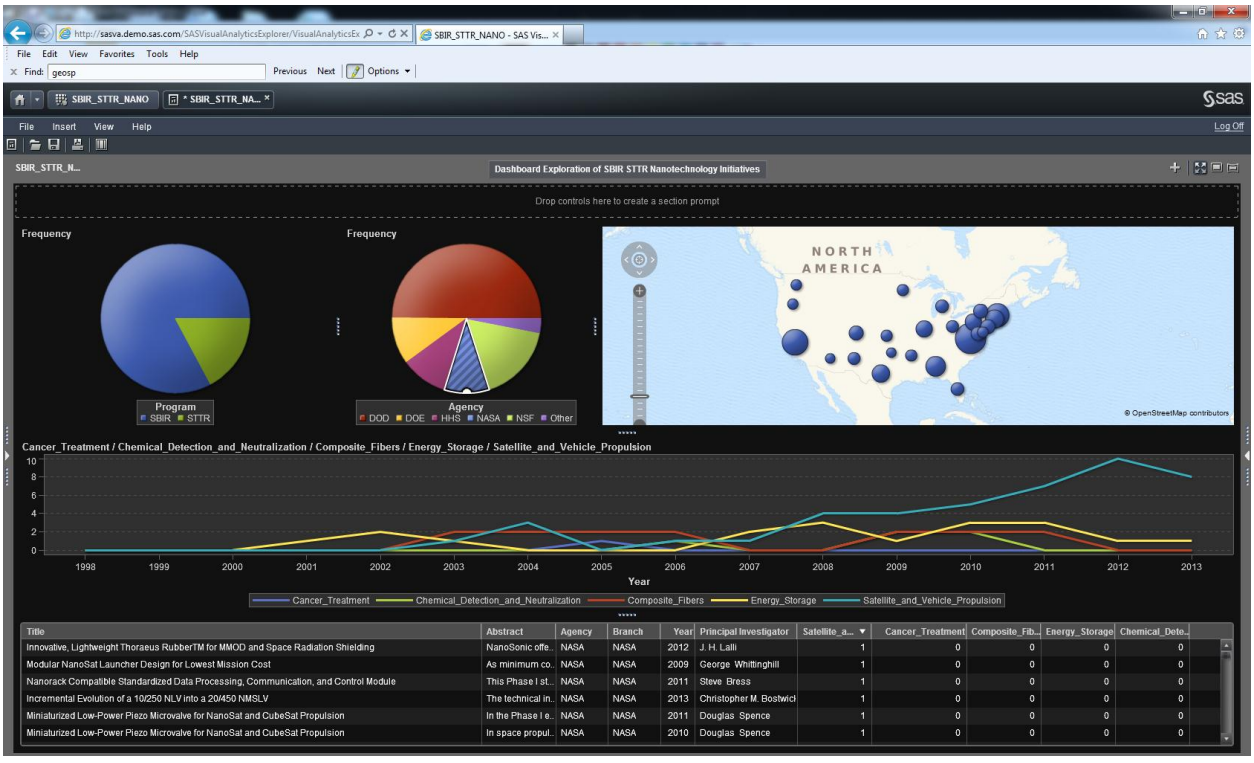

Figure 11: NASA Funded SBIR/STTR Nanotechnology Research

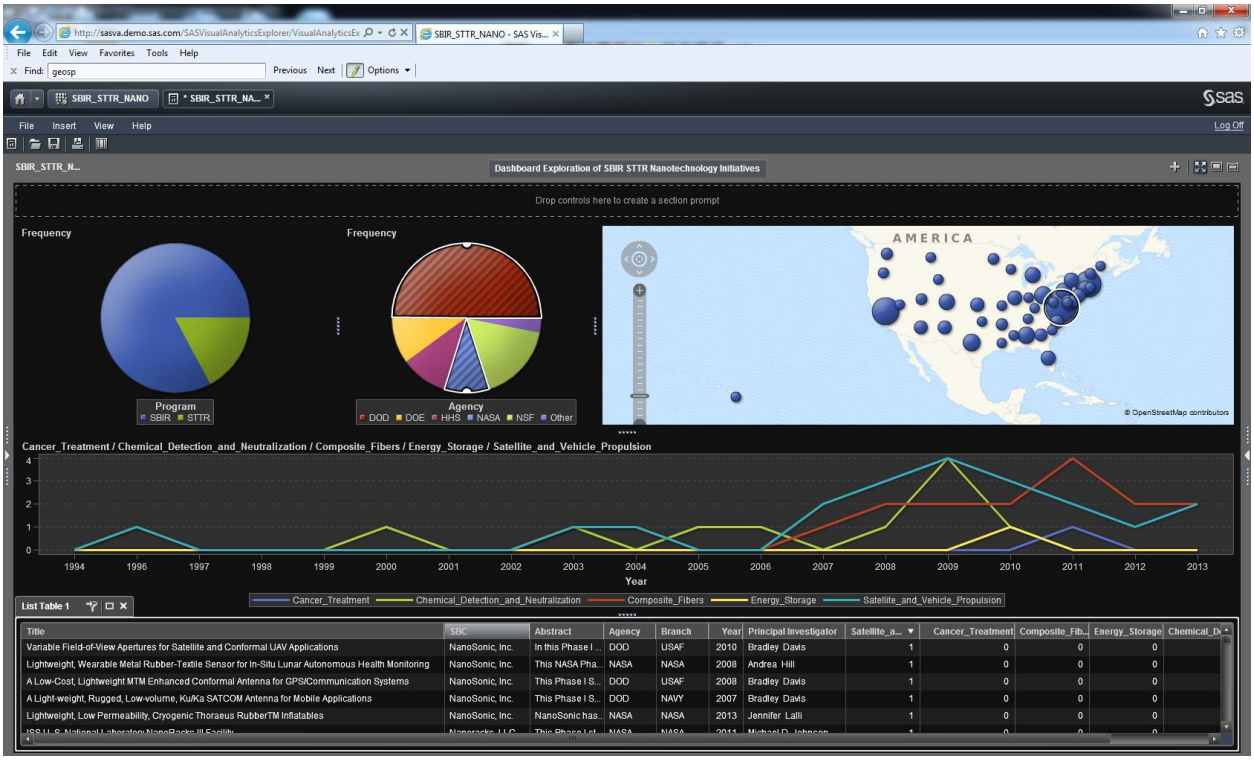

Figure 12: DOD and NASA Funded Nanotechnology Research Occurring in Virginia

DISCUSSION

We have demonstrated how to apply our text analytics capabilities and visual explorations to the variety of questions set forth at the beginning of this example. Through topic extraction and category promotion, we discovered the various applications of nanotechnology research, from cancer treatment to spacecraft system design. Through visual analytics explorations, we can depict this over time and determine which trends are increasing, such as spacecraft system design, and which are decreasing, such as protective coatings. Capabilities within SAS Contextual Analysis enable us to determine how terms and concepts are interrelated within a nanotechnology topic, and also enable us to pull in new external data sets for categorical scoring. Finally, the SAS Visual Analytics reporting capabilities enable us to explore the data set by category, program, and geography. All these capabilities form a framework for research analytics that enables a researcher to make more informed decisions on funding future research efforts, and ensures that new research initiatives are non-duplicative. Anticipating future research trends can also assist government agencies to staff accordingly to handle the influx of requests related to these emerging technologies.

It is worth noting that the generation of these visualizations, both exploratory and ad hoc dashboard, took no more than a couple of hours. The associated text analytics and data preparation also took no more than a couple of hours, highlighting the ability of SAS Contextual Analysis to streamline the topic extraction and subsequent category promotion text analytics tasks. The visualizations and explorations depicted here just scratch the surface of what is possible using SAS Text Analytics and SAS Visual Analytics, and we will explore additional capabilities in the next example.

ESTABLISH A FRAMEWORK FOR RESEARCH ANALYTICS IN AEROSPACE ENGINEERING

To support research endeavors, analysts often rely on supporting research documents to prepare for and guide their new initiatives. For example, if they are working to develop a next generation spacecraft, they need to quickly analyze relevant research in key areas so that past challenges are taken into consideration and mitigated. If an agency is building a new type of multistage rocket, it would want to

avoid previous issues with stage separation, which occur when smaller rockets are attached and used to launch the space vehicle. As the rockets run out of propellant, they must be released to reduce the mass of the remaining rocket (without affecting the flight trajectory) and enable the remaining propellant to accelerate the rocket to its final speed and altitude.

For analysts responsible for ensuring that a new design avoids problems in past designs, they need to ask—and answer—questions such as the following:

- What are the relevant topics of research that I need to consider?
- What are the trends in this research over time?
- Where are the research centers that have the expertise in these research areas?
- Who are the experts on this topic?[3]

In this section, we will explore a framework for research analytics that deals specifically with this issue of stage separation and follows the five steps identified earlier. We will identify the cross-cutting areas of research using SAS Text Analytics, and depict ad hoc explorations and interactive reporting using SAS Visual Analytics.

DATA ACQUISITION

For this study, we obtained data from the publicly accessible NASA Technical Reports Server (NTRS)[4]. NTRS provides access to aerospace-related citations, full-text online documents, and images and videos. The types of information include conference papers, journal articles, meeting papers, patents, research reports, images, movies, and technical videos—scientific and technical information (STI) created or funded by NASA. By searching for the term “State Separation,” we were able to identify and download a data set of 582 documents with 16 columns of data, including a title and abstract section. Again, SAS Enterprise Guide was used to convert the CSV file into a SAS data set for text analytics.

TEXT ANALYTICS – APPLYING SAS ENTERPRISE MINER AND SAS TEXT MINER

SAS Enterprise Miner and SAS Text Miner provide a graphical user interface for extracting topics and clusters. Topics are differentiated from clusters not just by the algorithms that define them, but by the fact that one document can have multiple topics associated with it, while each document fits into exactly one cluster. As a result, the underlying data sets that represent these results are different. Topic results are a number of binary variables, while cluster results are stored in a single nominal variable. This impacts what visualization techniques apply to these text analytics results.

Figure 13 highlights a graphical flow in SAS Enterprise Miner and SAS Text Miner used to generate text topic and text cluster results.

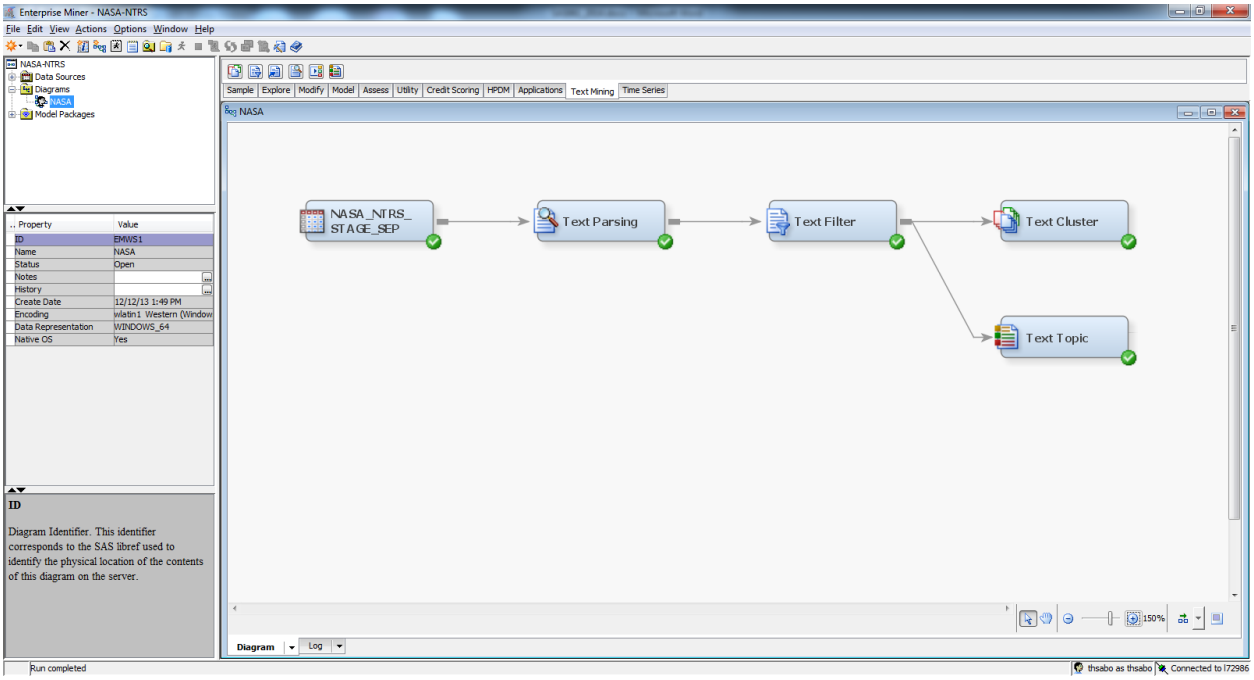

Figure 13: SAS Enterprise Miner and SAS Text Miner Graphical Flow for Text Mining NTRS Documents

The flow is configured to point to a data set—namely, the SAS data set that we generated in the previous step using stage separation abstracts pulled from the NTRS server. It is configured to designate the abstract field as the text to be parsed. A text parsing and text filter node is applied with default settings. Finally, a topic node and a cluster node are used to obtain text topic and cluster results.

While default settings are used for this example, it is worth noting that the Properties panel on the left side of the graphical user interface provides a wider range of configuration options than what is available in SAS Contextual Analysis. Further differences between SAS Contextual Analysis and SAS Enterprise Miner and SAS Text Miner will be reviewed at the end of this paper.

Figure 14 highlights the topics extracted using SAS Enterprise Miner and SAS Text Miner. These topics are similar to those extracted using SAS Contextual Analysis in structure because they use the same underlying technology. An interactive browser enables limited ad hoc exploration of these topics, which will be enhanced using SAS Visual Analytics in a later step.

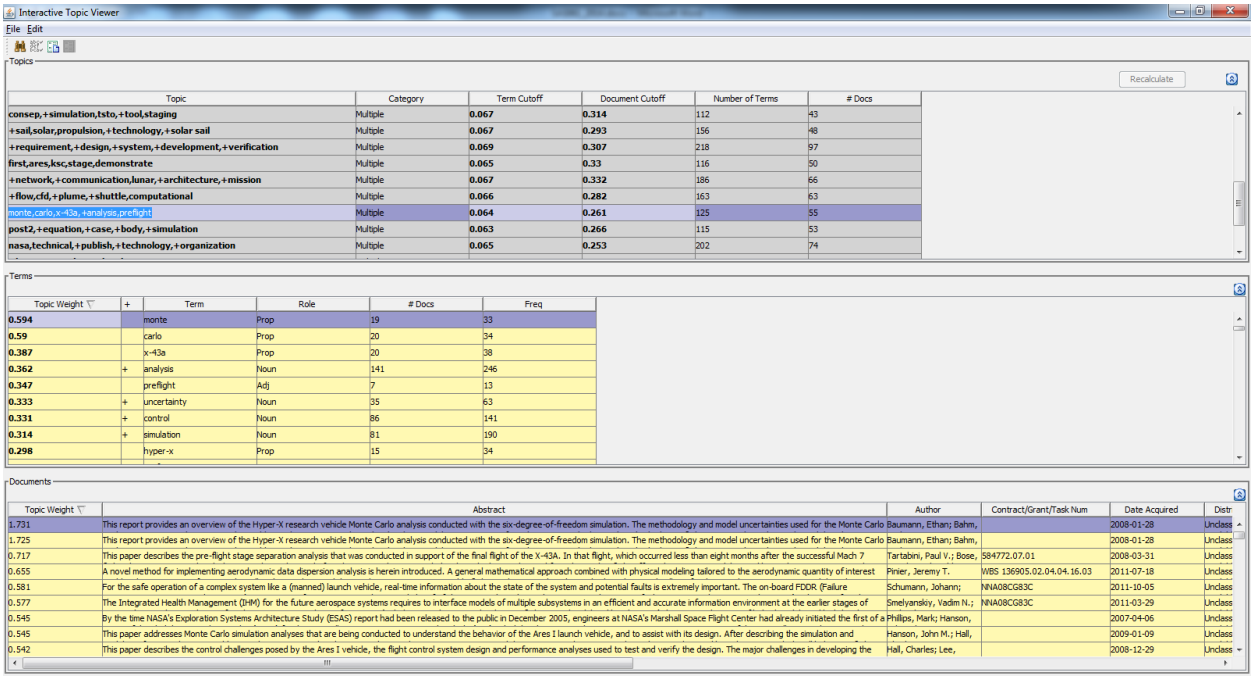

Figure 14: SAS Enterprise Miner and SAS Text Miner Generated Topics with Monte Carlo Analysis Highlighted

Figure 15 highlights the clusters extracted using SAS Enterprise Miner and SAS Text Miner. In this case, we defined the cluster properties to match exactly 25 clusters with 12 descriptive terms, and we wind up with similar descriptive terms as the text topics. Again, the key difference between topics and clusters is that a single document can have multiple associated topics, such as wind trajectories and Monte Carlo analysis, while clustering associates one document with exactly one cluster.

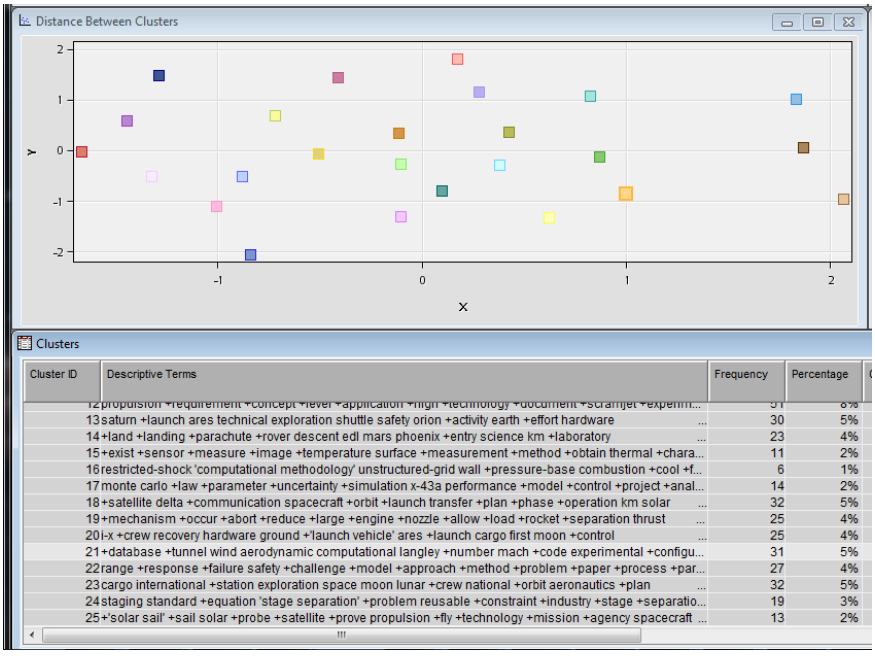

Figure 15: A Portion of the Text Cluster Results Window Highlighting 25 Clusters

DATA PREPARATION FOR VISUAL ANALYTICS

Similar to SAS Contextual Analysis, SAS Enterprise Miner and SAS Text Miner produce a number of SAS data sets on the file system for downstream use. We merge three of these together using SAS Enterprise Guide. These files include textcluster_train for the text cluster results (this file also includes all the original structured data), textcluster_clusters for cluster lookup values, and text_topic train for text topic results. As part of this process, we also relabel the text topics and clusters with easy to understand names. So, for instance, the topic “monte, carlo, x-43a, +analysis, preflight” becomes “monte carlo analysis.” We accomplished this for clusters by copying the textcluster_clusters into an Excel file, relabeling this lookup table there, and then importing the results back into a SAS data file prior to merge. We rename the topics as part of the query_builder that joins the three tables and selects the data that we will use in SAS Visual Analytics.

AD HOC EXPLORATION

While similar explorations as given in the example case for SBIR-STTR nanotechnology analysis are possible, here we will focus on additional visualizations, as well as one that is more appropriate for the single variable that results from cluster analysis.

Clusters can be visualized as a tile chart as shown in Figure 16. In fact, in this case, we generated a hierarchy (on-the-fly) that begins with the text clusters. So, for instance, users can explore the various categories generated from the analysis, including Solar Sail technology, Hypersolic Propulsion, and Wind Aerodynamics, each of which is a cross-cutting area of research that is related to stage separation. Users can then drill down to see the NASA Centers where work in these topics took place, and drill further down to see the publications resulting from these NASA Centers. Additionally, a color gradient can be applied to the clusters to determine how various topics cross-cut the clusters. This could reveal, for instance, a correlation between research around flight trajectories and, specifically, the Ares_1-X spacecraft.

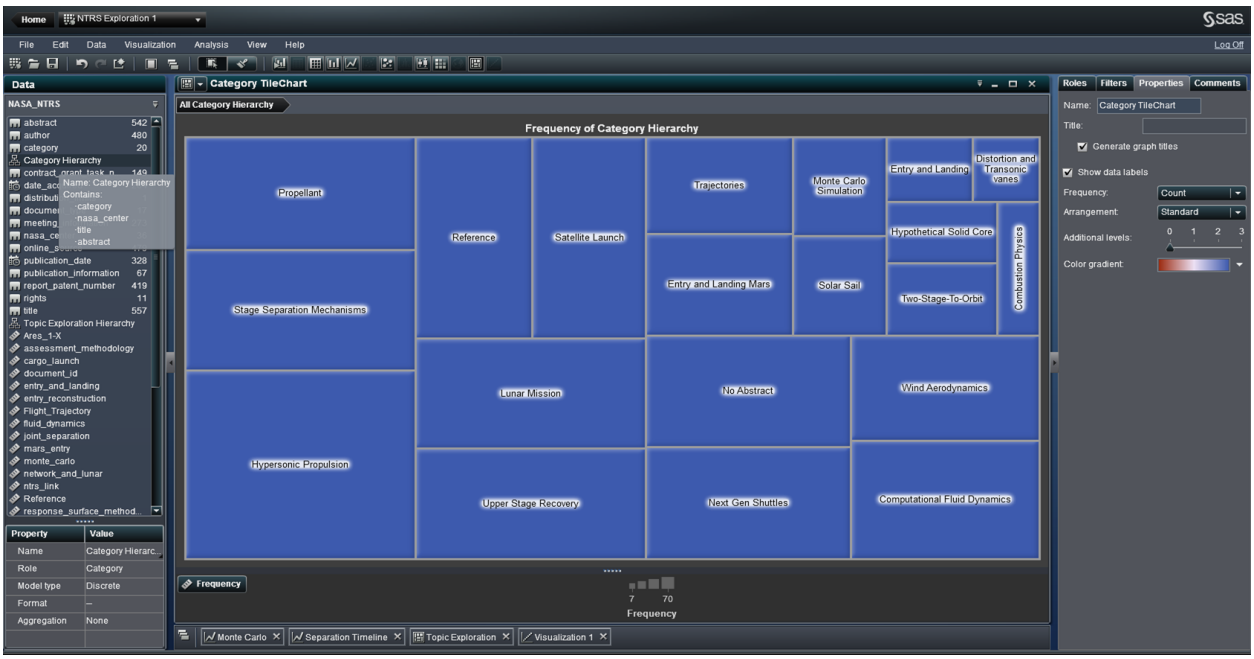

Figure 16: Tile Chart in SAS Visual Analytics Depicting a Drillable Hierarchy Beginning with Research Clusters

Building further upon this, a correlation analysis against any of the variables used in the analysis can be pulled up as an ad hoc exploration. Figure 17 showcases a correlation analysis of selected measures, each one of which is a topic that we imported. Using this, for example, we can visualize how the topics related to the Ares 1-X are correlated with topics related to cargo launch, flight trajectory, and vibration analyses. Each of these are related to stage separation, and research in these areas can be leveraged for next-generation spacecraft.

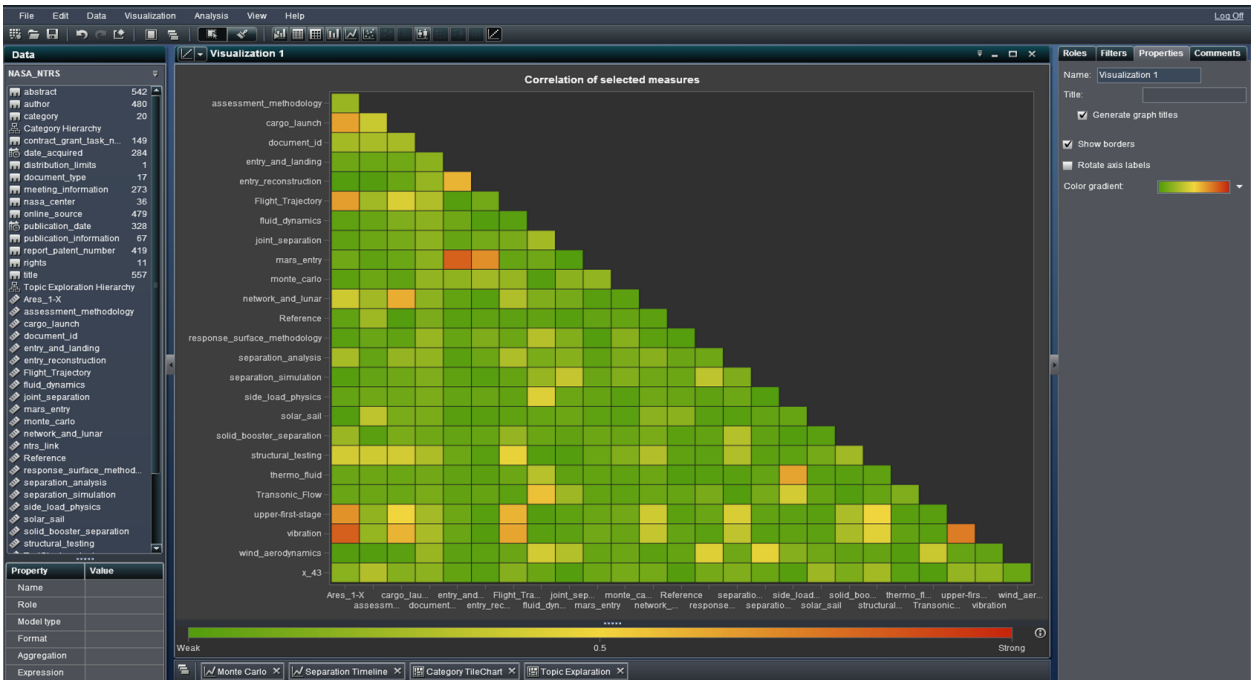

Figure 17: Correlation Analysis of Generated Topic Variables Related to Stage Separation

INTERACTIVE REPORT GENERATION AND USE

Figure 18 depicts a snapshot of a very similar dashboard visualization as used in the previous example for SBIR-STTR nanotechnology. However, the notable difference is that there is a button bar across the top that enables the user to begin explorations by clusters. Thus, a user may be interested in exploring two state-to-orbit studies, filter by only the conference papers generated for those studies, and look specifically at research coming out of Langley Research Center. Users are provided with the relevant authors, publication dates, and details of the research coming out from those studies.

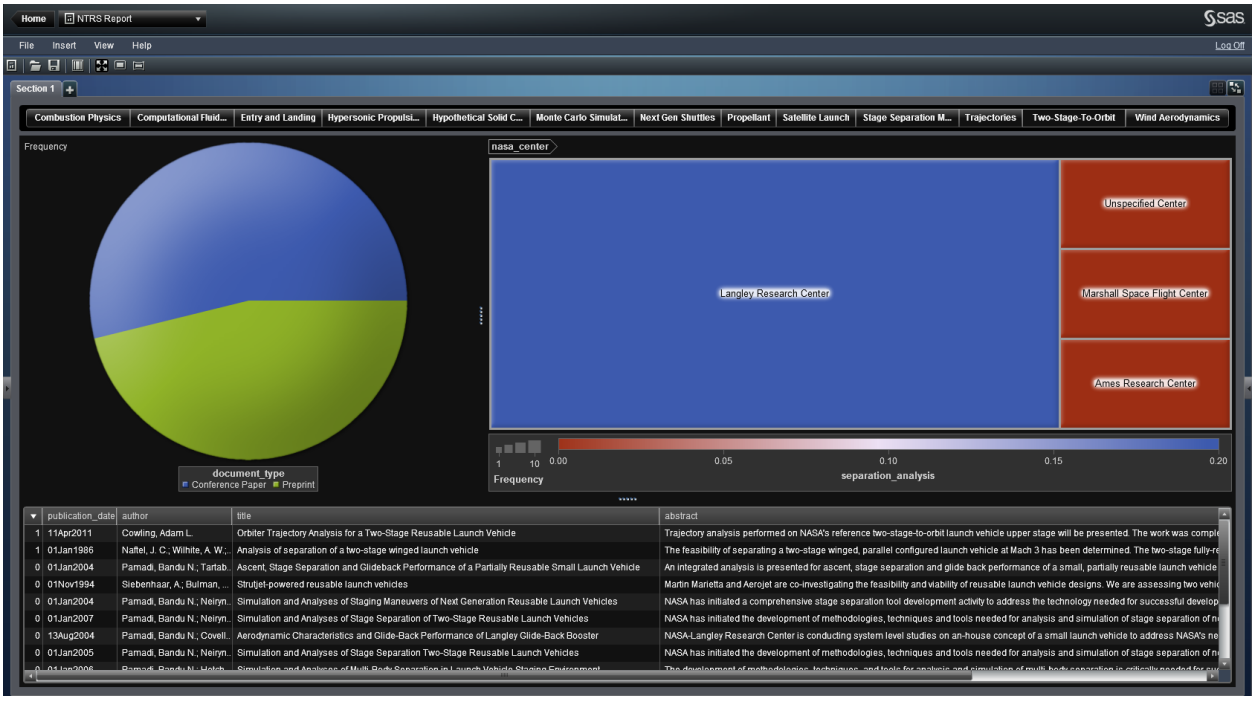

Figure 18: NASA NTRS Reporting Dashboard for Cross-cutting Research Topics to Stage Separation

DISCUSSION

In summary, using text analytics to identify relevant cross-cutting areas of research can help researchers visualize how these content areas trend over time, point them to the contributing research centers, and help them access expert contributors faster. To return to our example, this puts researchers in a much better position to ensure that future stage separation problems within next-generation spacecraft do not occur.

CONCLUSION

This paper illustrated a framework for research analytics enabled through SAS Text Analytics and SAS Visual Analytics technology. Specifically, in two examples, we obtained and prepared data for text analytics, performed text analytics, prepared data for exploration and reporting, and used SAS Visual Analytics for this ad hoc exploration and interactive reporting. Organizations can leverage this framework, or elements thereof, in their own research efforts.

We explored in this paper text analytics enabled via two forms of SAS technology, SAS Text Miner and SAS Contextual Analysis. The two products use similar technology in different ways. For instance, both products use technology to designate text rules given a target variable. SAS Enterprise Miner and SAS Text Miner use this technology to assist with predictive modeling, while SAS Contextual Analysis uses this technology to promote topics to categories. While you could accomplish topic-to-category promotion using SAS Text Miner along with another SAS product, SAS® Enterprise Content Categorization, SAS Contextual Analysis streamlines this capability by incorporating functionality from both products specifically for the purposes of topic-to-category promotion, and additional scoring against new data sets. However, SAS Contextual Analysis enables only the generation of text topics, versus topics and clusters, which are available in SAS Enterprise Miner and SAS Text Miner. Having reviewed the visualizations for each example, the choice of technology may be partially dependent on the desired end result, for the ability to explore and interact with the data.

Technology continues to further enable text analytics and visual exploration capabilities against text analytics results. As this process continues to unfold, new solutions that use this combined technology set will emerge and existing ones will evolve. This intersection of technology and evolving solutions will remain an ongoing field of study and real-world application for the foreseeable future.

REFERENCES

1. Sabo, Tom. 2014. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data,” p. 1. http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.
2. SBIR/STTR. http://www.sbir.gov/about/about-sbir.
3. Sabo, Tom. 2014. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data,” p. 4. http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.
4. NASA Technical Reports Server. http://www.sti.nasa.gov/find-sti/#ntrs.

CONTACT INFORMATION

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

Name: Tom Sabo, Sr. Solutions Architect
Organization: SAS Federal LLC
Address: 1530 Wilson Blvd.
City, State ZIP: Arlington, VA 22209
Work Phone: +1 (571) 227-7000 x51717
Email: tom.sabo@sas.com