Hearing Every Voice: SAS® Text Analytics for Federal Regulations Public Commentary

Emily McRae, Tom Sabo, Manuel Figallo, SAS Institute Inc

ABSTRACT

Regulations.gov was launched in 2003 to provide the public with access to federal regulatory content and the ability to submit comments on federal regulations. Public participation in federal rulemaking is encouraged as it supports the legitimacy of regulatory decisions, frames public acceptance or resistance to rules under development, and shapes how the public interest will be served. Manually reading thousands of comments is time-consuming and labor-intensive. It is also difficult for multiple reviewers to accurately and consistently assess content, themes, stakeholder identity, and sentiment. Given that individually proposed rules can exceed 10,000 comments, how can federal organizations quantitatively assess the data and incorporate feedback into the rulemaking process as required by law?

This paper shows how SAS® Text Analytics can be used to develop transparent and accurate text models, and how SAS® Visual Analytics can quantify, summarize and present the results of that analysis. This will significantly decrease time to value, leveraging capabilities that computers excel at while freeing up human intuition for the analysis of these results. Specifically, we will address public commentary submitted in response to new product regulations by the US Food and Drug Administration. Ultimately, the application of a transparent and consistent text model to analyze these documents will support federal rule-makers and improve the health and lives of American citizens.

INTRODUCTION

Electronic cigarettes, or "e-cigarettes/e-cigs" are small electronic devices, which are used to simulate some of the properties of smoking. They work by heating a liquid to expel an aerosol or vapor, which the user then inhales. The heated liquid and resulting vapor in an e-cig typically contains nicotine, and the devices are very often marketed as a smoking cessation aid. Proponents of e-cig devices claim that they are healthier than traditional cigarettes, as they don't produce tar or noxious gases. However, it is important to note that a) nicotine is still a highly addictive substance, b) e-cigs might contain contaminants or other unintended toxins and c) the long-term health effects of e-cig usage are unclear. [1]

Of particular concern is the growing usage of e-cigarettes in children and other non-smokers. The National Youth Tobacco Survey (NYTS) is conducted annually by the FDA and CDC to track rates of tobacco use in children (middle and high school students). The most recent data indicates that usage is trending upward as the proportion of high-school users who reported using e-cigarettes on 20 days or more in the last 30-day period, increased from 20% to 27.7% between 2017-2018. Overall, approximately 3.6 million children reported using e-cigarettes in 2018. [2]

If children and teenagers are not using e-cigarettes to quit smoking, what is driving this trend? One suggestion is the inclusion of flavors in e-cigarette liquid, which replace the traditional tobacco taste of cigarettes. Common flavors include mint/menthol, strawberry, chocolate, cinnamon, gummy-bear, and cotton-candy, among many others.[3] For context, conventional cigarette manufacturers are prohibited from adding flavors (other than menthol) to their products, a move the FDA implemented specifically to reduce youth tobacco usage.[4]

To further understand the disparate drivers of e-cigarette usage and to balance their potential benefit as smoking cessation tools versus their potential risk to under-age users, the FDA submitted a notice of proposed rulemaking to *regulations.gov* (Regulation of Flavors in Tobacco Products, Docket ID: FDA-2017-N-6565) and opened a public comment period. The notice specifically asks for '*comments, data, research results, or other information about, … how flavors attract youth to initiate tobacco product use and about whether and how certain flavors may help adult cigarette smokers reduce cigarette use and switch to potentially less harmful products.*'[5]

This notice attracted over 525,000 comments, of which ~23,000 were made publicly available. Using this data, this paper will showcase a repeatable solution framework for the analysis of *regulations.gov* comments, in addition to familiarizing readers with the capabilities of SAS® Visual Text Analytics and SAS® Visual Analytics.

SOLUTION DEVELOPMENT

SAS® Visual Text Analytics provides a comprehensive suite of text mining capabilities, including Natural Language Processing (NLP), contextual extraction of concepts, categorization by machine learning and sentiment analysis. This paper will show how these capabilities can be used specifically for the analysis of *regulations.gov* data. Furthermore, aspects of this analysis can be used across different federal agencies and regulations, and so form the backbone of a repeatable solution.

SAS® Visual Text Analytics capability	*Regulations.gov* analysis	Repeatable?
Concepts	Extract identity of referenced organizations	Y
Categories	Identify and quantify the document themes and form letters	N
Sentiment	Positive, Negative, or Neutral sentiment	Y

Table 1. SAS® Visual Text Analytics capabilities mapped to solution components

REGULATIONS.GOV API DATA EXTRACTION

Regulations.gov provides a public API for the extraction of documents (public comments) and dockets (the high-level organization folders, which contain the proposed regulations and any other background documents).[6] This paper uses code (courtesy of Manuel Figallo) to pull documents associated with Docket FDA-2017-N-6565.[7] The resulting data contains a unique ID for each comment, the comment text and other associated metadata, such as the posted date and title.

To prepare the data for analysis, we ran the following code:

```
/* Reformat Date variable and drop extraneous variables */
data work.reformat;
format Date Date9.;
set work._1_;
Date = input(postedDate, yymmdd10.);
drop agencyAcronym commentDueDate commentStartDate docketId docketTitle
postedDate;
run;
```

In addition, we ran sentence tokenizer code to break each individual comment into its constituent sentences and generate a sentence ID (sid).[8] This allows for more nuanced

categorization of comments, as each comment could potentially reference multiple themes. The resulting tokenized data set is now ready for analysis in SAS® Visual Text Analytics.

	documentId	sid	sentences
1	FDA-2017-N-6565-0126	1	Shouldnt we be focusing on the big picture here!!!
2	FDA-2017-N-6565-0126	2	Its about Tobacco ..
3	FDA-2017-N-6565-0126	3	E-liquid is a extract from the plant without all the chemicals in the cigarettes!!!
4	FDA-2017-N-6565-0126	4	People want to quit smoking and this is a healthier way of doing that !!!
5	FDA-2017-N-6565-0126	5	Electronic cigarettes devices are helping people with this problem of traditional cigarettes.
6	FDA-2017-N-6565-0126	6	Flavors, nicotine level and also you can have e-liquid with no nicotine!!!!
7	FDA-2017-N-6565-0126	7	I do use ESD products and it helped me quit cigarettes !!!

VIEWTABLE: Work.Comment_sentences

Figure 1. Example *regulations.gov* analysis data set, following sentence tokenization

CONCEPTS FOR CONTEXTUAL EXTRACTION

Given that we know the subject of e-cigarette usage is fairly controversial, we decided to identify the different interest groups or identities associated with each comment. For example, is the comment poster an e-cigarette user? A doctor or other health care professional? Do they represent or quote from a special interest group or professional organization? How often are specific e-cigarette manufacturers mentioned?

These questions can be answered by using the concepts node in SAS® Visual Text Analytics, which uses contextual extraction rules to identify and extract specific entities (defined as people, places of things of interest, used here interchangeably **with 'concepts'**). These contextual extraction rules use SAS proprietary LITI (Language Interpretation for Textual Information) syntax and use elements such as stemmed terms, parts-of-speech, and Boolean operators to define a specific context for a matched entity. This approach delivers high-precision results, especially in 'noisy' data.

IDENTIFY PROFESSIONAL ORGANIZATIONS

We initially noticed that many of these comments referred to research or positions stated by various professional organizations or lobby groups. When choosing to identify these organization entities, we first developed a series of simple rules, which define common organizational names. For example, the 'Academy of Such-and-Such' or the 'So-and-So Society', each one of these terms can be treated as a generic organizational 'flag.'

Model Studio - Build Models

FDA-2017-6565 > Concepts

Concepts   Edit a Concept

Predefined Concepts (0)

Custom Concepts (15)

COMMENT_IDENTITY

HELPER_CONS

    ORG_ID

    COMPANY_ID

```
1   CONCEPT:CENTER_ID
2   CLASSIFIER:College
3   CLASSIFIER:Council
4   CLASSIFIER:Convention
5   CLASSIFIER:Foundation
6   CLASSIFIER:Initiative
7   CLASSIFIER:Institute
8   CLASSIFIER:Network
9   CLASSIFIER:Organization
10  CLASSIFIER:Commission
11  CLASSIFIER:Society
12  CLASSIFIER:Trust
```

Figure 2. Concept LITI rules define initial organization entity 'flags'

However, as you can see from the matched blue text in Figure 3 below, this approach returns noisy results, as many of these terms can have other meanings.
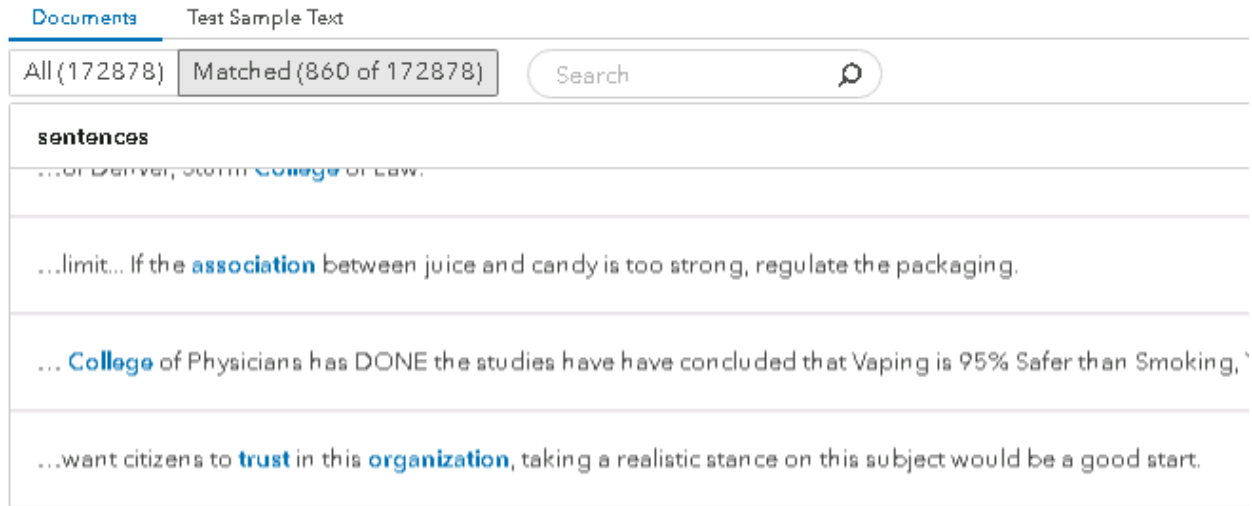

Figure 3. Matched text for concept *'ORG_ID'* including irrelevant 'noise'

It is necessary to further restrict matches on these organizational terms to a particular context. In this case, we chose to build on these initial definitions and specified that they must reference capitalized terms (denoted by the _cap token) and occur in a particular structure, as below.
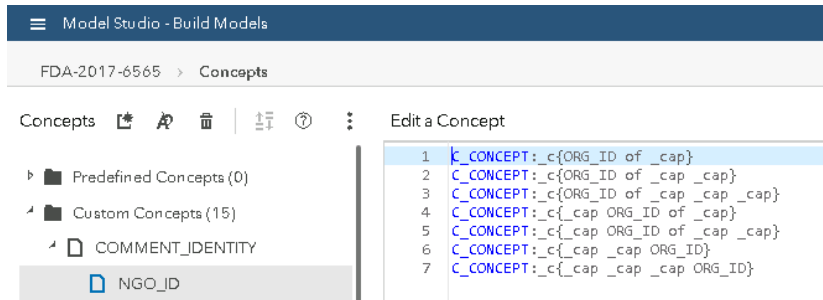
Figure 4. Concept LITI definitions are refined to include structure and context

Now the matched blue text indicated that we are predominantly matching the organizational entities we are interested in.
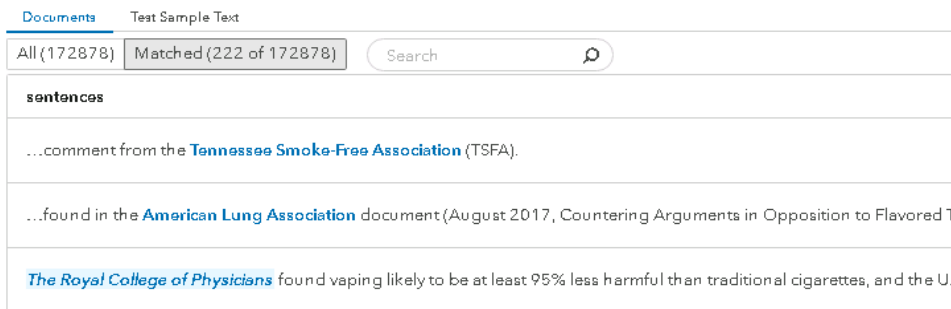
Figure 5. Matched text for concept *'NGO_ID'* showing improved accuracy

While we can generate accurate matches against organization entities, can we also improve the recall by adding in more definitions? And can we identify more organizational flags without reading through all the text?

SAS® Visual Text Analytics generates a 'similarity score', which is a measure of how likely it is for terms or phrases to occur in the same context. We used this feature to identify other organization flag terms. For example, using the 'Textual Elements' panel in the Concepts node, we searched for one of our flag terms – 'institute.'

Figure 6. Frequency and Parts-of-Speech for filtered term 'Institute'

Note the 'Role' column indicated that this term is present as both a Proper Noun (PN) and a Verb (V). We selected the PN role and clicked on the 'Similarity Score' icon (highlighted in red box). This approach identifies two additional organization definitions ('academy' and 'initiative') that we can add to our original ORG_ID concept definitions, thereby increasing the number of positive matches.

Figure 7. Green highlighted fields show terms related to 'institute', following term similarity scoring

DISAMBIGUATE MATCHES

The preceding sections indicated how it is possible to identify and restrict matches for organization entities to a particular context. But what if we need additional control over these matches? For example, the flag word 'center' is one of the definitions used to identify organizations, but also matches references to the Center for Tobacco Products (CTP), which is the FDA Center responsible for producing this *regulations.gov* docket. Can we remove the organizational matches for CTP?

SAS® Visual Text Analytics provides a special LITI rule type called REMOVE_ITEM, which allows users to disambiguate matches with a common, partial definition, which can mean more than one thing. We start by explicitly specifying the matches we want to remove in the 'FDA_ID' concept:
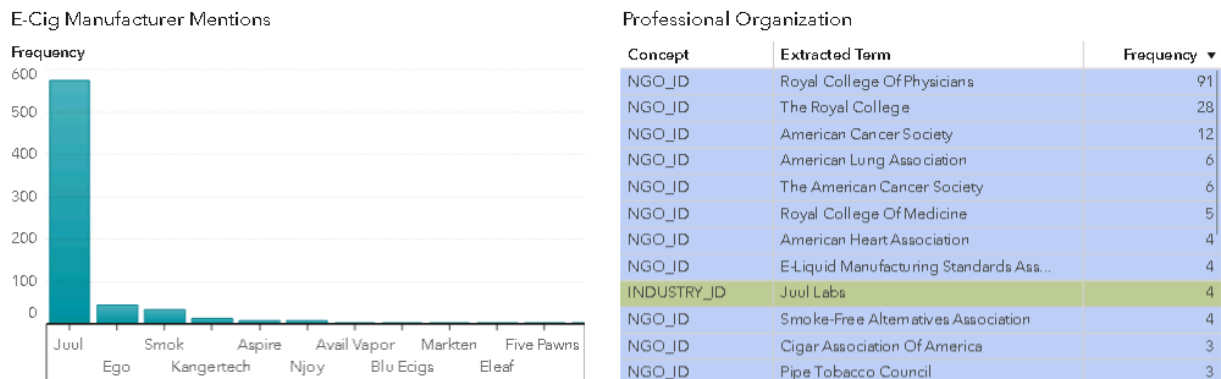

Figure 8. Concept LITI Classifier rules define unwanted matches

Then we use the 'ALIGNED' operator in a REMOVE_ITEM rule to return unrestricted matches on 'CENTER_ID' concepts, except where they also return 'FDA_ID' concept matches. We tested the results with some sample text, as seen below.

Figure 9. The LITI REMOVE_ITEM rule provides disambiguation by removing unwanted matches

Following the development of our concept rules, we used the Concepts score code produced by SAS® Visual Text Analytics to extract mentions of e-cigarette manufacturers and professional organizations. This newly structured data was then used as input for report-building in SAS® Visual Analytics. The combination of SAS® Visual Text Analytics and SAS® Visual Analytics allows users to quickly visualize and interact with the results of text analysis. In particular, this approach makes the results of text modeling by a select few analysts available to a much broader audience of stakeholders.

E-Cig Manufacturer Mentions

Professional Organization

Concept	Extracted Term	Frequency ▼
NGO_ID	Royal College Of Physicians	91
NGO_ID	The Royal College	28
NGO_ID	American Cancer Society	12
NGO_ID	American Lung Association	6
NGO_ID	The American Cancer Society	6
NGO_ID	Royal College Of Medicine	5
NGO_ID	American Heart Association	4
NGO_ID	E-Liquid Manufacturing Standards Ass...	4
INDUSTRY_ID	Juul Labs	4
NGO_ID	Smoke-Free Alternatives Association	4
NGO_ID	Cigar Association Of America	3
NGO_ID	Pipe Tobacco Council	3

Figure 10. Juul is by far the most referenced e-cigarette manufacturer. The Royal College of Physicians and the American Cancer Society top the list of referenced professional organizations.

CATEGORIZATION FOR THEME IDENTIFICATION

One of the main challenges associated with text analysis is simply that of data volume. The point at which it becomes very intensive to manually read through text is reached after a comparatively small number of documents. Think of a spreadsheet with a few hundred observations of structured data (dates, prices, quantities, and so on), which can be quickly interpreted with a representative graphic, versus the same number of free-form text documents, each of which might refer to any topic in any form! In this instance, with over 20,000 comments publicly available, how can we quickly and comprehensively assess the representative themes?

SAS® Visual Text Analytics addresses this issue by using a 2-step machine learning approach to categorize documents by the main ideas or themes across any volume of text data. Initially, the Topics node generates a series of lists, each of which contain related terms. These terms commonly occur together in a subset of the model document corpus. Therefore, it can be used to describe these documents, as well as differentiate them from the rest of the documents in the model corpus. In the second step, the Categories node combines the selected terms from the Topics model with Boolean operators to generate linguistic rules, which can be used to assess document inclusion for a category/theme.

SMOKING CESSATION

SAS® Visual Text Analytics categories automatically generated by the above machine-learning approach are initially labeled with a list of relevant terms. However, after reviewing the documents in each category, we chose to rename them in a more interpretable manner. It was very quickly apparent that the majority of opinions expressed in this data relate to

the ways in which e-cigarettes, or specifically e-cig flavors impact people's desire or ability to quit smoking.

Category	Example
TOBACCO_USAGE	I was a 2 **pack** a **day smoker** for over 10 **years** and the **day** I picked up a vape is the **day** I was able to quit cigarettes.
WHY_FLAVORS	Most of the people I know quit smoking (who quit smoking rather) stopped by **using flavors** OTHER than **Menthol** and Tobacco **flavors**;
RESUME_SMOKING	But lets just say if you took are flavors away from us I would more than likely **go back** to **smoking**
TRY_QUIT	I have **tried several times** to **quit** smoking **tried** every method none of them **work** until I started vaping
STOP_SMOKING	many people have benefited from these flavors and have **quit smoking**
SAVE_LIVES	**Vaping** can **save lives** if I can stop smoking cigarettes after 20 years of a pack and a half a day and now I'm down to 3 mg of nicotine
CHILD_USAGE	would you rather have **kids vaping** or smoking a cig I would rather hear a **kid vaping** than smoking to be honest.

Table 2. Categorization of *regulatons.gov* comments by themes and example matching text

Interestingly, though most opinions seemed to suggest general endorsement of the role e-cigs play in smoking cessation, the categorical analysis indicates that commenters chose to frame their arguments in several ways, ranging from documenting their previous smoking habits, as well as their previous attempts to quit smoking, to suggesting a return to traditional cigarettes if flavors were banned. Most of the commenters appear to be adult e-cigarette users and any reference to the abuse of e-cigarette flavors by under-18s appears to be largely ancillary.

We used the Category and Sentiment score code produced by SAS® Visual Text Analytics to assign these documents to categories and to generate sentiment scores. This data was then used as input to a bar-chart in SAS® Visual Analytics. By clicking on the bars in the graphic below, it is also possible to toggle through the associated documents for each category. The addition of SAS® Visual Analytics to this approach greatly improves the interpretability of this data for the purposes of sharing with a wider audience.

Figure 11. The most prevalent themes in these documents express support for e-cigarette flavorings or discuss their usage as smoking cessation aids.

FORM LETTER IDENTIFICATION

In addition to the volume and variety of documents submitted to *regulations.gov*, it is also apparent that some individuals participate in concerted campaigns centered around a common point of view. In many cases, these individuals chose to submit form letters, in which most of the document is identical and can contain only small customizations, if any. It would be very difficult, if not impossible, to manually identify and quantify these form letters. However, by using the output from the SAS® Visual Text Analytics Topics node, which produces topic scores for each document, followed by a clustering analysis, it is possible to group together documents, which can be very similar or identical.

We ran the Text Topics score code, followed by some additional post-processing code to produce clusters:

```
/* PROC Distance generates measures of distance/similarity that can then be
used as input for clustering  */
proc distance data=casuser._out_documents out=Dist method=Euclid;
var interval(_Col1_-- _Col17_ / std=Std);
ID documentId;
run;

/* PROC Cluster groups related documents */
proc cluster data=Dist method=Ward outtree=tree;
ID documentId;
run;

/* PROC Tree generates cluster membership */
proc tree data=tree out=clusters nclusters=1400;
ID documentId;
run;

/* Remove clusters of only 1 document */
data work.clusters2;
set work.clusters;
where documentId ^= clusname;
run;

/* Join documentId and clusters to original text */
proc sql;
create table clusters3 as
select C.*, S.sentences
from work.clusters2 C left join public.fda_forms_score S
on C.documentId = S.documentId;
quit;
run;
```

This process generates a new data set, in which form letters are grouped together under the same cluster. The following table shows an example form letter, in which the general outline is the same, but individual points are customized by each poster. This approach allows us to identify groups of individuals who are commenting as a bloc, while also quantifying their contribution to *regulations.gov*.

Cluster CL1467	
FDA-2017-N-6565-18305	**FDA-2017-N-6565-18360**
Declaration of [@advFirst][@advLast]	Declaration of [@advFirst][@advLast]
I, [@advFirst][@advLast], declare and state as follows:	I, [@advFirst][@advLast], declare and state as follows:
1. I am over the age of 18 years and have personal knowledge of the facts set forth below such that I would be competent to testify as a witness to the same if called.	1. I am over the age of 18 years and have personal knowledge of the facts set forth below such that I would be competent to testify as a witness to the same if called.
2. I am submitting this declaration in response to the FDAs request for public comments on the above-referenced docket regarding a proposed rule on the regulation of flavors in tobacco products, including electronic nicotine delivery system (ENDS), or vapor products.	2. I am submitting this declaration in response to the FDAs request for public comments on the above-referenced docket regarding a proposed rule on the regulation of flavors in tobacco products, including electronic nicotine delivery system (ENDS), or vapor products.
3. I am a resident of [@advCityState] and am 48 years of age.	3. I am a resident of [@advCityState] and am 53 years of age.
4. I started smoking cigarettes when I was 18 years old and smoked them for more than 25 years .	4. I started smoking cigarettes when I was 14 years old and smoked them for more than 35 years .
5. I have been using vapor products for 5 years and 11 months.	5. I have been using vapor products for 3 years and 5 months.
6. The categories of flavors of nicotine-containing e-liquid that I have used include Menthol/Mint, Fruit, Desserts. Of these, the flavor category that I use most often is Fruit.	6. The categories of flavors of nicotine-containing e-liquid that I have used include Tobacco, Menthol/Mint, Fruit, Desserts, Other Sweets, Other Flavors. Of these, the flavor category that I use most often is Desserts.
7. Before I started using flavored e-liquid products, I typically smoked one pack of cigarettes per day.	7. Before I started using flavored e-liquid products, I typically smoked one pack of cigarettes per day.
8. E-cigarettes have helped my smoking cessation more than any other product or medication on the market to date. Having tried nicotine patches, gum, as well as	8. Over my thirty plus years of smoking (Marlboro Red, Camel Light, Marlboro Medium/27, Dunhill, and Sobrane)
[…. abbreviated for length…]	[…. abbreviated for length…]
9. Since I began using flavored e-liquid products, I have been able to quit smoking cigarettes….	9. Since I began using flavored e-liquid products, I have been able to quit smoking cigarettes….

Table 3. Example form letters. Text highlighted in green is identical. Remaining text is customized by individual poster.

CONCLUSION

The objective of this paper was to map SAS® Visual Text Analytics capabilities to the analysis of *regulations.gov* data, such that this approach can be replicated by readers for their own federal agencies or interests. Specifically, we showed how LITI contextual

extraction can be used in SAS® Visual Text Analytics Concept rules to identify referenced organizations, manufacturers or comment poster identities. In addition, we showed how SAS® Visual Text Analytics Topic and Category modeling can be used to identify document themes and quantify form letters. This process is summarized in Figure 12, below.
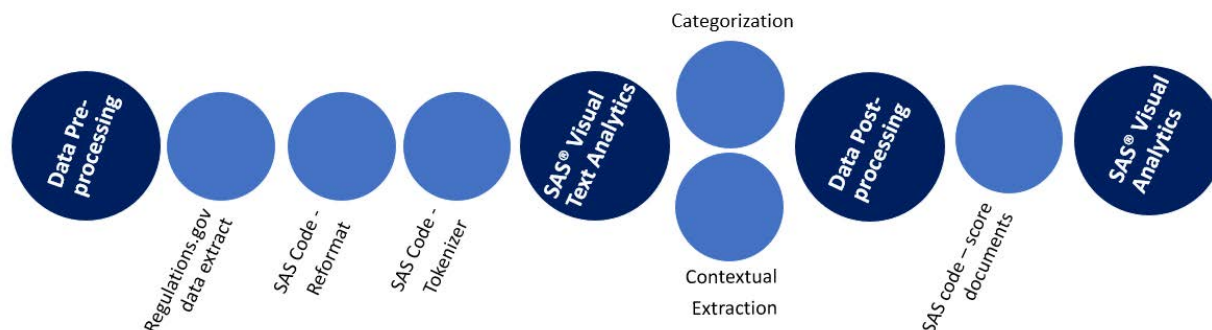
Figure 12. Overview of regulations.gov solution process

Regulations.gov is a very valuable source of text data and a key bridge between members of the public and the federal agencies, which serve their interests. Federal agencies are required by law to include the public in their rule-making process. According to section 553 of the Administrative Procedure Act, "*the agency shall give interested persons an opportunity to participate in the rule making through submission of written data, views, or arguments with or without opportunity for oral presentation.*" Traditionally, this has been accomplished by a manual review of public comment data, and in the case of controversial regulations, can involve reading thousands of documents to identify themes, sentiment and form letters.

SAS® Visual Text Analytics offers three key benefits for the analysis of *regulations.gov* data:

1. Consistency: human reviewers often struggle with ambiguity or nuanced definitions when assigning document themes. An algorithmic approach avoids these inconsistencies by generating a statistical representation of categories from term and document frequencies

2. Transparency: SAS® Visual Text Analytics models are not black-box, and results can be easily queried and interpreted. This supports the legitimacy of federal analysis for these comments and the overall rule-making process

3. Scalability: SAS® Visual Text Analytics models can be scaled to accommodate large volumes of data and category definitions. This is especially valuable as manual review of text is time-consuming and labor-intensive

This methodology can be applied to any public comments, as well as a wide variety of other text sources. The development and use of a text solution for *regulations.gov* analysis has the potential to greatly improve the accuracy and efficiency of these review efforts, particularly by reducing time to value. For example, if a docket receives 10,000 comments, and each comment takes 5 mins to manually review:

> 10,000 comments   x   5 minutes   =   833 hours   /   20 FTE weeks

In contrast, the process documented in this paper took approximately a week to develop. This represents a significant time-saving improvement, with the opportunity for continued improvement as the reuseable solution is applied to more text.

REFERENCES

[1] E-cigarettes: Good news, bad news Available at:
https://www.health.harvard.edu/blog/electronic-cigarettes-good-news-bad-news-2016072510010

[2] 2018 National Youth Tobacco Survey Available at:
https://www.fda.gov/TobaccoProducts/PublicHealthEducation/ProtectingKidsfromTobacco/ucm625887.htm

[3] Flavored e-cigarette use: Characterizing youth, young adult, and adult users Available at:
https://www.sciencedirect.com/science/article/pii/S2211335516301346

[4] Tobacco Control Act, section 907(a)(1)(A) Available at:

https://www.fda.gov/tobaccoproducts/guidancecomplianceregulatoryinformation/ucm263053.htm

[5] Docket FDA-2017-N-6565 Available at:
https://www.regulations.gov/document?D=FDA-2017-N-6565-0001

[6] Regulations.gov Developers Info Available at:
https://regulationsgov.github.io/developers/

[7] Regulations.gov API extraction code Available at:
https://github.com/sasgovernment

[8] Sabo, T., Pilz, A. 2018 "Using SAS® Text Analytics to Assess International Human Trafficking Patterns"

ACKNOWLEDGMENTS

RECOMMENDED READING

- *SAS® Visual Text Analytics 8.3: User's Guide*

- *SAS Text Analytics for Business Applications: Concept Rules for Information Extraction Models:* Jade, Teresa. Wilsey, Biljana Belamaric. Wallis, Michael. *Forthcoming 2019*

CONTACT INFORMATION

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

> Emily McRae
> SAS Institute
> Emily.mcrae@sas.com