SAND2019-11202C

Using Text Explorer to Inform and Enhance Risk and Issue Application Development & Functionality

By: Scarlett Marklin & Yvonne Petrova
Sandia National Laboratories

Introduction to Text Analytics with JMP

Text analytics within big data is gaining traction not just because the method displays "hidden" findings within unstructured data, but text analytics is a great way to merge the quantitative and qualitative aspects of data analysis to provide a greater understanding of phenomena. While free programs such as R and Python certainly do a wealth of analyses for textual data, they lack the visualization and the interactivity that makes JMP a powerhouse in the text analytics field. Researchers are no longer "flying blind" when it comes to pre-processing text and analyzing the results derived. JMP allows a fluid interaction between the analysts and the data, much like a painter when moving across a canvas to paint a picture. JMP allows greater flexibility in painting a picture and getting a solid jump start on diving into what story the data tell within the unstructured text without the need for analysts to have an in-depth background in coding. JMP allows users to learn coding as they go, in real time, with hands-on fluidity. The text explorer in JMP offers a solid platform for pre-processing and becoming familiar with the nuances of the data.

Visualization and interactivity make JMP a powerhouse in text analytics.

The added functionality of JMP Text Explorer includes:

- default values for baseline understanding of data,
- the option to show text which provides greater clarity behind how terms and phrases are mentioned,
- alphabetical and count ordering of terms and phrases to provide greater insight into areas of further investigation,
- summary statistics regarding overall words analyzed, terms found, and non-missing cases,
- the option to add and remove stop words interactively while pre-processing,
- managing stop words, phrases, recodes and stem exceptions without re-running analyses or modifying code to address changes, and
- using regular expressions to automate steps of pre-processing, increasing usability for analysts and provides friendly user functionality not found in R or Python.

JMP offers a unique hands-on experience with text data that is not found in any other platform, analysts can deep dive into the data without getting bogged down by the need to be an expert in JSL or have extensive coding skills. Analysts can focus more on answering the research questions posed instead of debugging code or figuring syntax required to arrive at a result. This cuts down on the time it takes to pre-process data. With JMP's flexible interactive design, analysts can complete an analysis in multiple ways, bringing the "fun" back to functional data analysis.

JMP brings the "fun" back to analysis.

Text Analytics at Sandia National Laboratories

Specifically, text analytics via the text explorer in JMP has been essential in the evaluation and customization of applications at Sandia National Labs for improved functionality and usefulness. Recently, Sandia management requested an analysis of data from the Assurance Information System (AIS). AIS was created in 2012 to fulfill a NNSA mandate to provide a "reliable system that consistently and accurately reflects assurance activities and guides the user through data entry to enable informed, data-driven decision making" (https://sharepoint.sandia.gov/sites/AIS/SitePages/AIS%20home.aspx). During the transition from the old, outdated AIS to a new tool called Sage, there was an established need to learn and improve upon the issues and consistent problems that plagued AIS. To improve consistency and accuracy, and to streamline data capture, details were needed from text captured in AIS.

Three core research questions were asked by management.

1. What are the kinds of issues being noted?
2. What are the causes?
3. What are the actions?

To answer these questions, text fields of interest were analyzed.

- Issue Statement
- Cause Description
- Action Description
- Results Description

Five separate analyses were conducted looking at the text to address the core research questions.

	Issue Statement	Action Description	Cause Description
With Causal	Table 1a Terms Analyzed=1,969	Table 3a Terms Analyzed= 2,253	Table 2 Terms Analyzed= 2,501
Without Causal	Table 1b Terms Analyzed = 319	Table 3b Terms Analyzed = 3,812	N/A due to no cases to analyze

In completing analyses, these text fields of interest were analyzed for those observations that were noted as causal and those without causal.

For risk and issues management, text fields are key in describing and explaining what the risks and issues are. Yet, these fields are unstructured and often difficult to navigate while pre-processing data. Often, text pre-processing is long and tedious, being described as an artform in many venues. There are ways to automate pre-processing based on the nature of the data. One way to automate within JMP is through use of regular expressions which can be reused when needed.

There are ways to automate pre-processing.

The Sandia text fields analyzed had a range of data input from empty fields whiched offered no insight to fields which contained excessive text, many fields included references to other documents with no additional information. As text analytics had not previously been used on risk and issues management data, this meant that there was no baseline from which to start. Prior to this effort, only counts of risks and issues and percent of open/closed were analyzed. Considering the nature of the data and the method employed, very close collaboration with the subject matter experts (SMEs) within risk and issues management was required to establish reasonability checks along the way. Due to the existing unstructured nature of the data, as part of the deliverable from this effort, suggestions were provided for categories regarding issues, causes, actions, and results which needed to be incorporated within the new tool as drop-down selections. If suggested categories could be added to the new tool during development, this would increase usability and create greater robust and quality data capture for future analyses. Furthermore, based on the text provided for results, sentiment analysis performed indicated whether actions taken had effectively worked to improve or correct the issue or risk, did not work to improve the issue or risk or could not be determined by the results provided.

	Result Description	
	With Causal 5,868 cases 165,730 tokens 1,857 terms	**Without Causal** 8,001 cases 281,056 tokens 3,151 terms
Results in which actions did not have desired outcomes or were inadequate in some way	6 topics	3 topics
Results did not specify any action outcomes (could not be determined), but offered improvements that could/should be made	3 topics	5 topics
Results in which actions noted had desired outcomes	1 topic	2 topics

The methodologies employed for risk and issue analysis included:

- Perform initial clean-up of data (correcting spellings, stemming, punctuation, lower cases, removal of words that have larger frequencies, but don't offer any discriminate analysis, etc.).

- Review terms, stems and phrase lists JMP found (based on frequencies), determining if any information required separate indicator columns within JMP for additional questions.
- Use additional sources including the JMP community (https://community.jmp.com/), JMP books found within JMP itself (Help>>>Books), regular expression resources and tutorials (https://regular-expressions.info; https://regexcrossword.com/), GitHub (https://github.com/) and semantic resources (https://tone-analyzer-demo.ng.bluemix.net/)

The analysis selected was Latent Semantic Analysis, Singular Vector Decomposition (SVD) to analyze results (categories/topics). Term frequency-inverse document frequency (TF IDF) takes the log of number of times the term is used vs. the number of documents (cases) used. It then devalues words used frequently and gives weight to more important words used in each case. Weighting used was centered and scaled to mirror principle components output. Topic analysis –rotated SVD provides a certain group of terms which aids in trying to find buckets or categories.

Outcomes

In analyzing clusters for issues, causes, actions and results, JMP allowed for recommendations and improvements during development of Sage (the latest risk and issues tool) which not only addressed the research questions, but also provided additional answers to questions not yet considered, including but not limited to ways of mapping areas of risk (e.g. location) via drop-down configuration, having radial buttons to note systemic issues and risks (those repeated), proposing an alternative measurement to risk likelihood and consequence matrices, etc. Additionally, recommendations were made for approaches to capturing more robust data and decreasing the amount of time it took users to enter necessary information into the tool.

For example, from the text results, it was noted that users, while capturing issues and risks in AIS, were often doing the minimum and many optional fields of interest were left blank. Some issues noted stemmed from disagreement with management or colleagues and/or personality clashes, which should not have been captured in the tool. This result highlighted users lack of understanding for what should be captured as a risk or issue in AIS and what should not. It was evident that AIS fell short because of tool design which was difficult to use and did not provide clear steps for entering data to help guide users. The analysis highlighted systemic issues across the labs and further highlighted that the same actions were being taken with the same results – these actions were inadequate. These findings supported SME opinion that issues were not being appropriately addressed and follow-ups were not adequate to limit re-occurrence of issues.

The most impactful results from this effort were:

- Stronger results than R
    - Pre-processing time was reduced by half!
    - JMP alerted to indicators not readily apparent in R!
    - More efficient due to focusing on analysis instead of debugging code!
- Gaps and inconsistencies in the historical data were highlighted
- Hidden information within AIS was clarified (the good, the bad and the downright ugly)
- A clear path forward for enhancements/design of new tool was provided
    - Based on analysts desires to get more granular findings
    - Greater accuracy with data capture
- SMEs were provided with text derived categories

Demonstrated Methodology

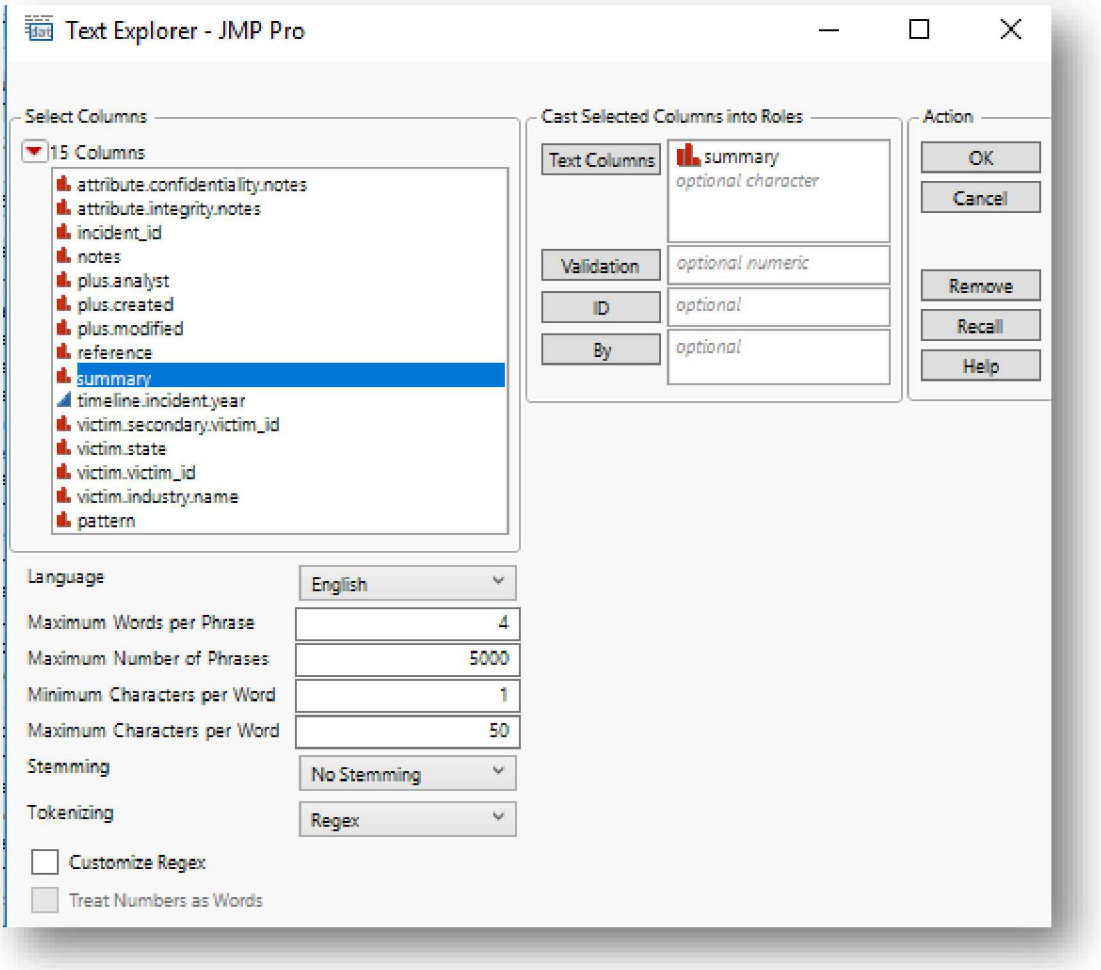
Due to the sensitive nature of the actual Sandia data used for the Sandia effort, this paper uses an unclassified public dataset on data breaches in the United States to demonstrate the pre-processing methodology employed to highlight how JMP was efficient and aided in enhancements for risk and issue application development and functionality.

Default for Baseline

JMP Text Explorer (Analyze>>>Text Explorer) offers users the ability to

- Select from multiple languages (e.g. English, Chinese, French, etc.)
- Set maximum words per phrase (e.g. 4 would mean having phrases up to 4 words)
- Set maximum number of phrases (default is 5,000)
- Set minimum characters per word (e.g. wanting to limit how words based on the amount of characters), and
- Set maximum characters per word that allows limiting how long a word can be.

Users can also select to stem (default is no stemming) or tokenize (regex) prior to looking at the data. This initial step offered by the text explorer allows users a “baseline” view of the data to gain greater understanding behind the written word and to begin addressing some pre-processing steps required to clean the data to a better degree for analysis.



Capitalizing on JMP's Interactive Display Options

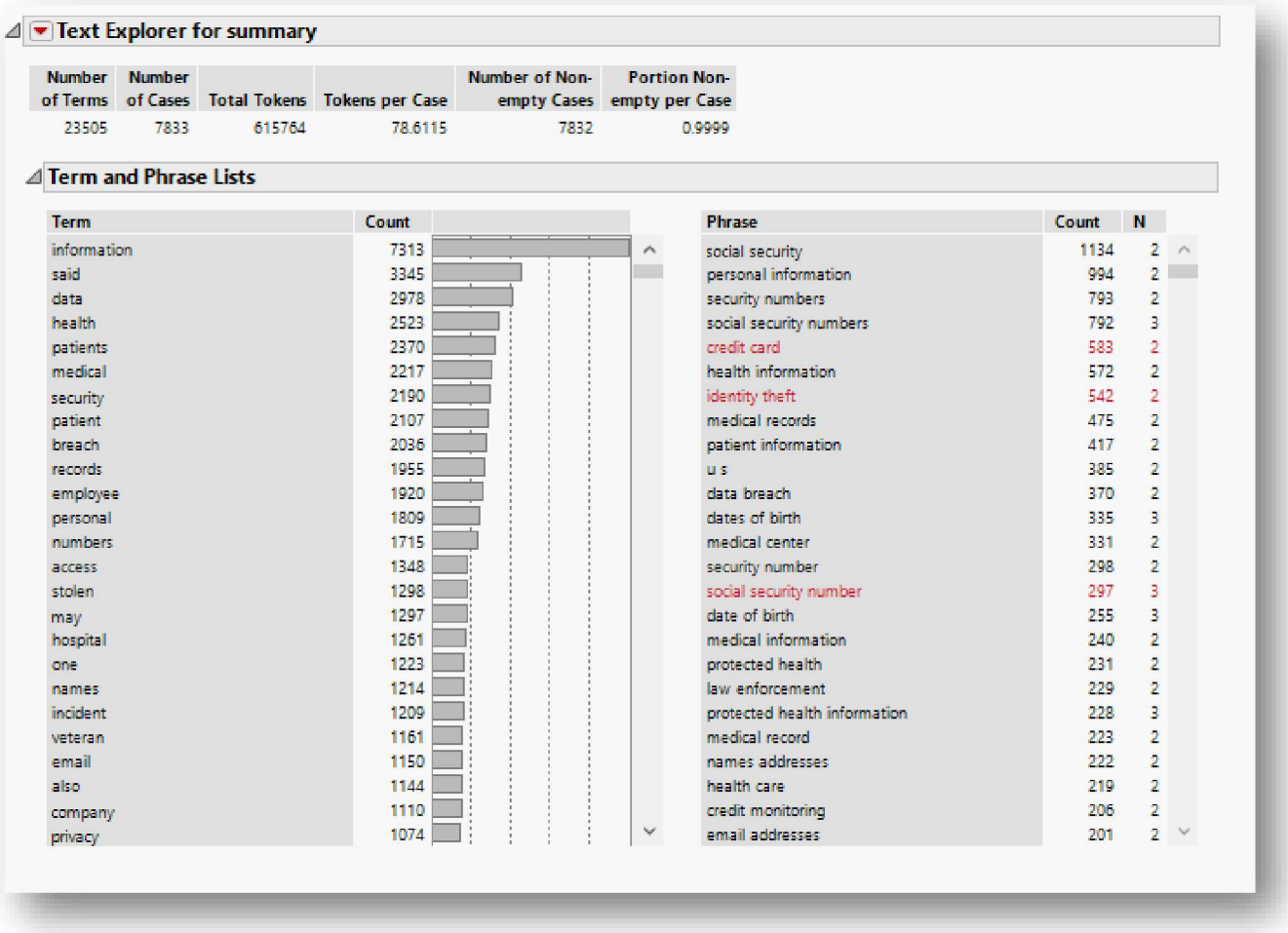
Visual Display of Term and Phrase Frequencies and Summary Statistics

Text Explorer offers users the ability to look at terms and phrase lists, based on either default or specified values provided by the user to show the frequency of the words or terms mentioned most often. JMP provides users with a summary of:

- Number of terms found
- Number of cases analyzed
- Overall word count (tokens)
- Average tokens (words) per case (observation), and
- Missing data, noting the number of non-empty cases (observations/rows) and the proportion.

These summary statistics are extremely important when analyzing data as having a low response rate will limit the generalizability of the results and should be considered when performing analysis to

inform questions regarding participation, engagement, sensitivity of the field, etc. For our example, using the default setting for data breaches within the United States from 2013 to year ending 2018, it is easy to see the top list of terms and phrases first using the default (frequency). This provides an initial overview of information found within the dataset.



Text Explorer for summary

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
23505	7833	615764	78.6115	7832	0.9999

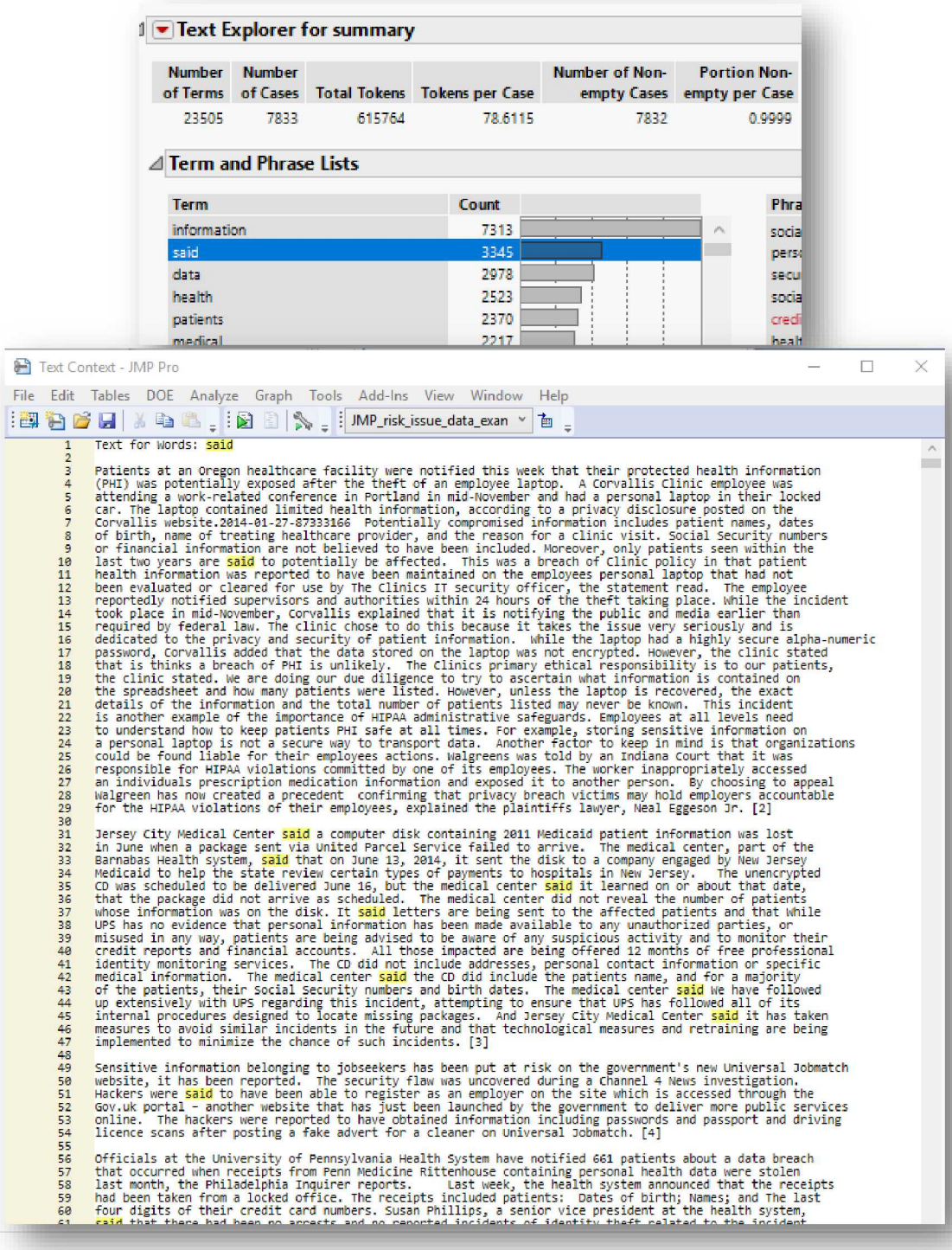
Term and Phrase Lists

Term	Count
information	7313
said	3345
data	2978
health	2523
patients	2370
medical	2217
security	2190
patient	2107
breach	2036
records	1955
employee	1920
personal	1809
numbers	1715
access	1348
stolen	1298
may	1297
hospital	1261
one	1223
names	1214
incident	1209
veteran	1161
email	1150
also	1144
company	1110
privacy	1074

Phrase	Count	N
social security	1134	2
personal information	994	2
security numbers	793	2
social security numbers	792	3
credit card	583	2
health information	572	2
identity theft	542	2
medical records	475	2
patient information	417	2
u s	385	2
data breach	370	2
dates of birth	335	3
medical center	331	2
security number	298	2
social security number	297	3
date of birth	255	3
medical information	240	2
protected health	231	2
law enforcement	229	2
protected health information	228	3
medical record	223	2
names addresses	222	2
health care	219	2
credit monitoring	206	2
email addresses	201	2

Show Text

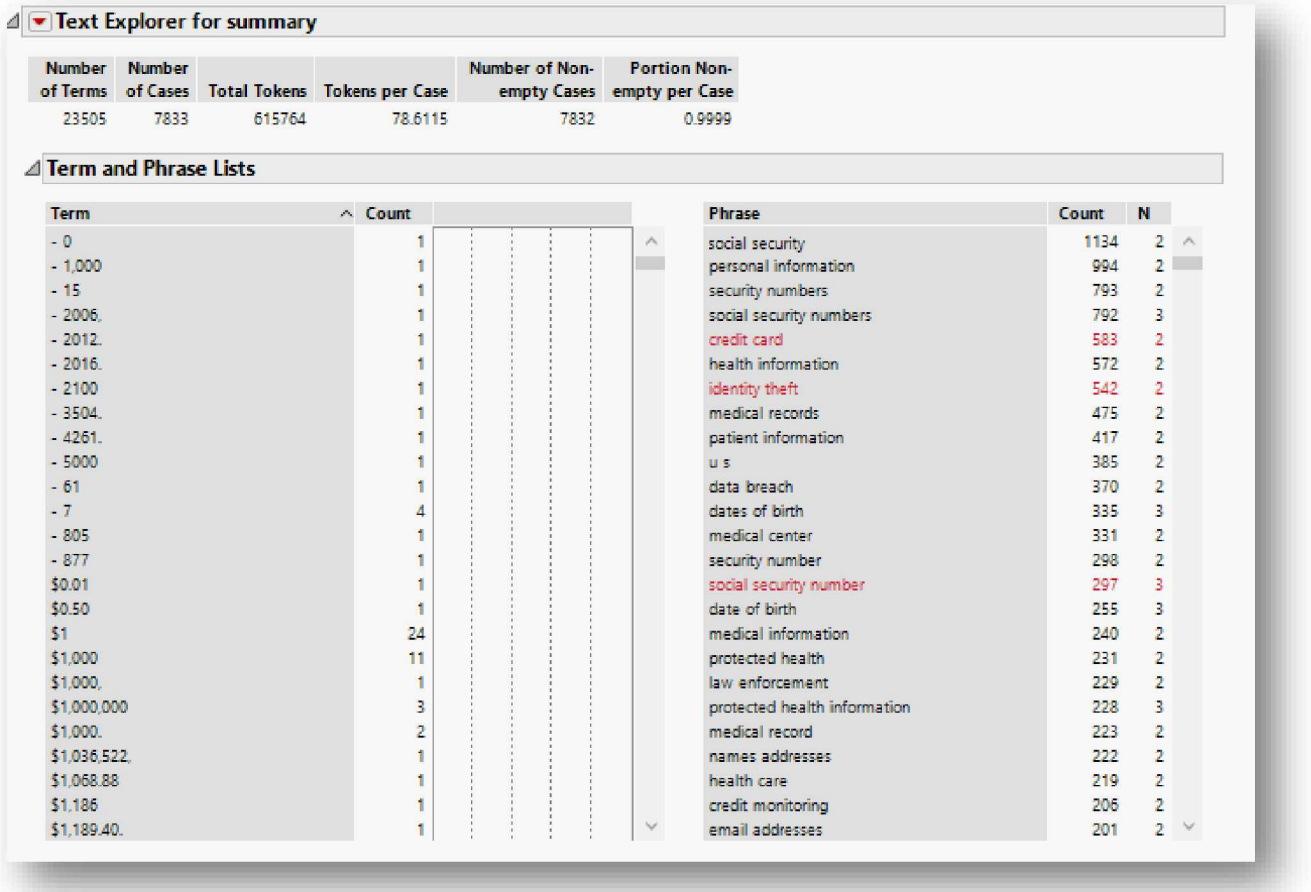
The highly useful “Show Text” option found within JMP’s Text Explorer (Right click on term>>> Show Text) allows users to look at how the word is used within each observation/row it is found within the corpus (dataset). For example, prior to any major pre-processing the word “said” is the second most frequently found word in the dataset, mentioned 3,345 times. Too see how this word is used in each instance found, the “Show Text” option displayed below provides greater insight into its relative importance to the data and the research question being asked.



If the analysis is answering the research question: “What were the risks and issues found from the data breaches between 2013 and 2018 within the United States?”, the word “said” can now be eliminated (treated as a stop word) and not used for analysis in looking at topics or clusters that will be used to answer the research question.

Red text shows key phrases of interest that are built into JMP.

Additionally, the option to select terms alphabetically (Right Click within Term Window>>>Alphabetical Order) can show greater areas of focus for pre-processing. This feature allows users an early insight into steps that should be taken in preprocessing regarding the focus for the analysis. For example, in using the alphabetical ordering feature in JMP, users can see numbers that do not look important. The “Show Text” feature can be used to gain greater insight into these values. For example, some numbers are associated with dollar values ($), so further exploration is needed to determine if these values are important. Notice that the red phrases listed in the example are already key phrases of interest built into JMP. Users can add more phrases of interest, but credit card, identity theft and social security number are already key phrases in JMP.



Text Explorer for summary

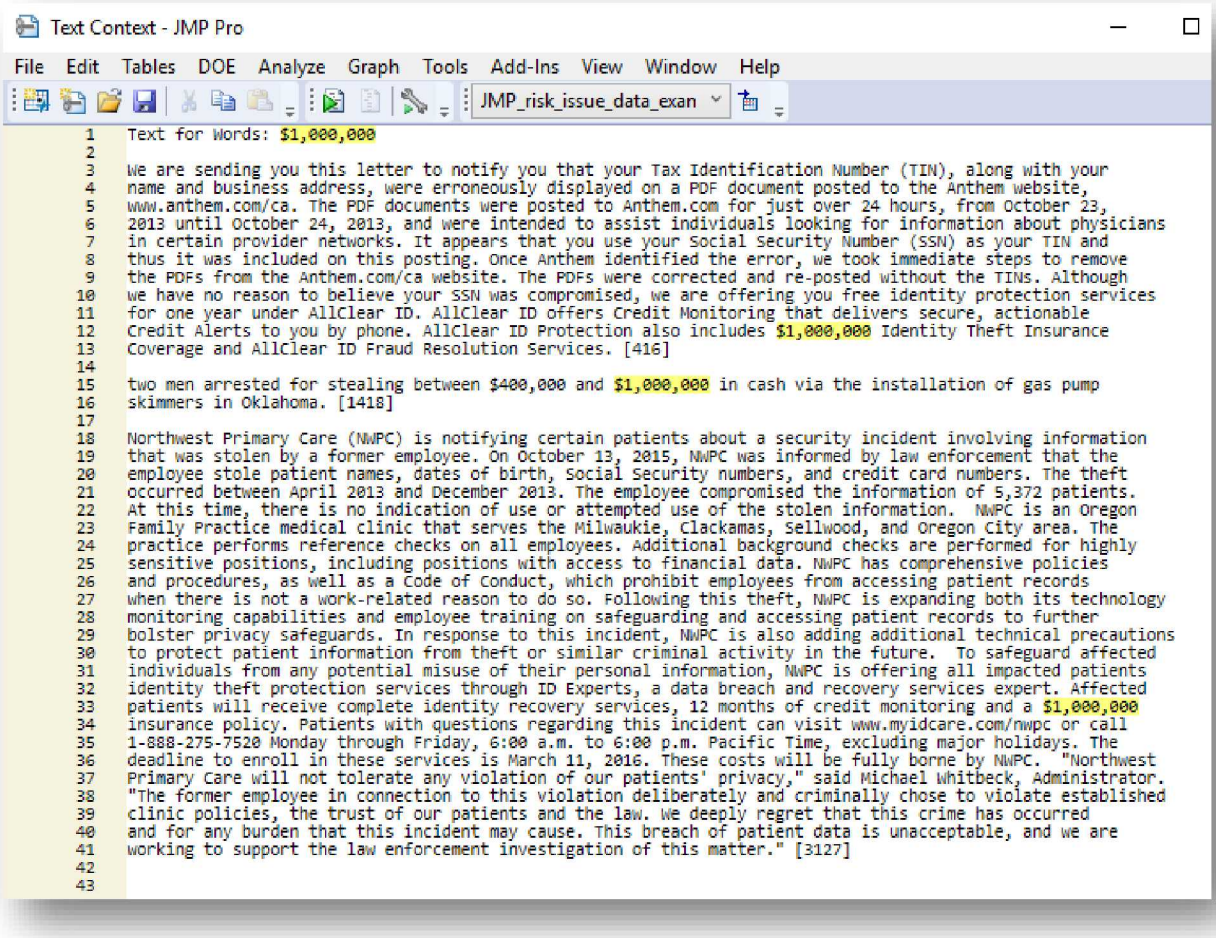
Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
23505	7833	615764	78.6115	7832	0.9999

Term and Phrase Lists

Term	Count
- 0	1
- 1,000	1
- 15	1
- 2006.	1
- 2012.	1
- 2016.	1
- 2100	1
- 3504.	1
- 4261.	1
- 5000	1
- 61	1
- 7	4
- 805	1
- 877	1
$0.01	1
$0.50	1
$1	24
$1,000	11
$1,000,	1
$1,000,000	3
$1,000.	2
$1,036,522.	1
$1,068.88	1
$1,186	1
$1,189.40.	1

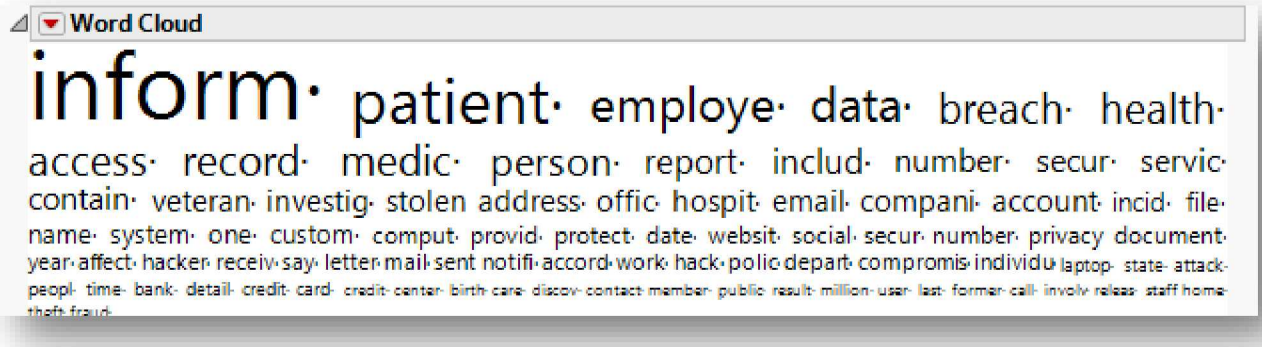
Phrase	Count	N
social security	1134	2
personal information	994	2
security numbers	793	2
social security numbers	792	3
credit card	583	2
health information	572	2
identity theft	542	2
medical records	475	2
patient information	417	2
u s	385	2
data breach	370	2
dates of birth	335	3
medical center	331	2
security number	298	2
social security number	297	3
date of birth	255	3
medical information	240	2
protected health	231	2
law enforcement	229	2
protected health information	228	3
medical record	223	2
names addresses	222	2
health care	219	2
credit monitoring	206	2
email addresses	201	2

For our example, let's select the values of $1,000,000 to review what context this term is used. From the example, this value is used with regard to the steps that companies are taking to mitigate and address the breach and it is also used in terms of amount stolen from victims at gas pumps using skimmers. Based on this initial exploratory evaluation of the data, it may be useful to analysts to create a new column to track mitigation strategies incorporated, and another regarding value of data lost. In using the text explorer platform, analysts can quickly elaborate and separate out fields of valuable information which can be used for analyses regarding data breaches, in this case settlements of breaches as a new column of information within the datafile. This same method can be applied to view key phrases alphabetically which provides additional insight into information that could be important for analysis (e.g. if researchers were only interested in data breaches that specifically dealt with social security numbers). JMP allows users a more efficient means of identifying which information may be parsed into a new column as a separate indicator (Right click phrase/term>>>Save Indicators) or which key phrases should be considered as words themselves (e.g. identity theft).



Word Clouds

JMP allows analysts to visually see the key terms and phrases within word clouds (size represents greater frequency). They are often used as a visual option for tag clouds which became popular within community-oriented websites and are considered a low-cost option for visual displays (Heimerl, et al., 2014). Word clouds are efficient in showing essential terms found within the data (words that capture attention or "pop"), and word clouds are engaging (they tend to have some impact and generate interest among viewers). Word clouds are heavily relied on to show results without getting into the technicalities.



Capitalizing on JMP's Interactive Term Options

Stemming

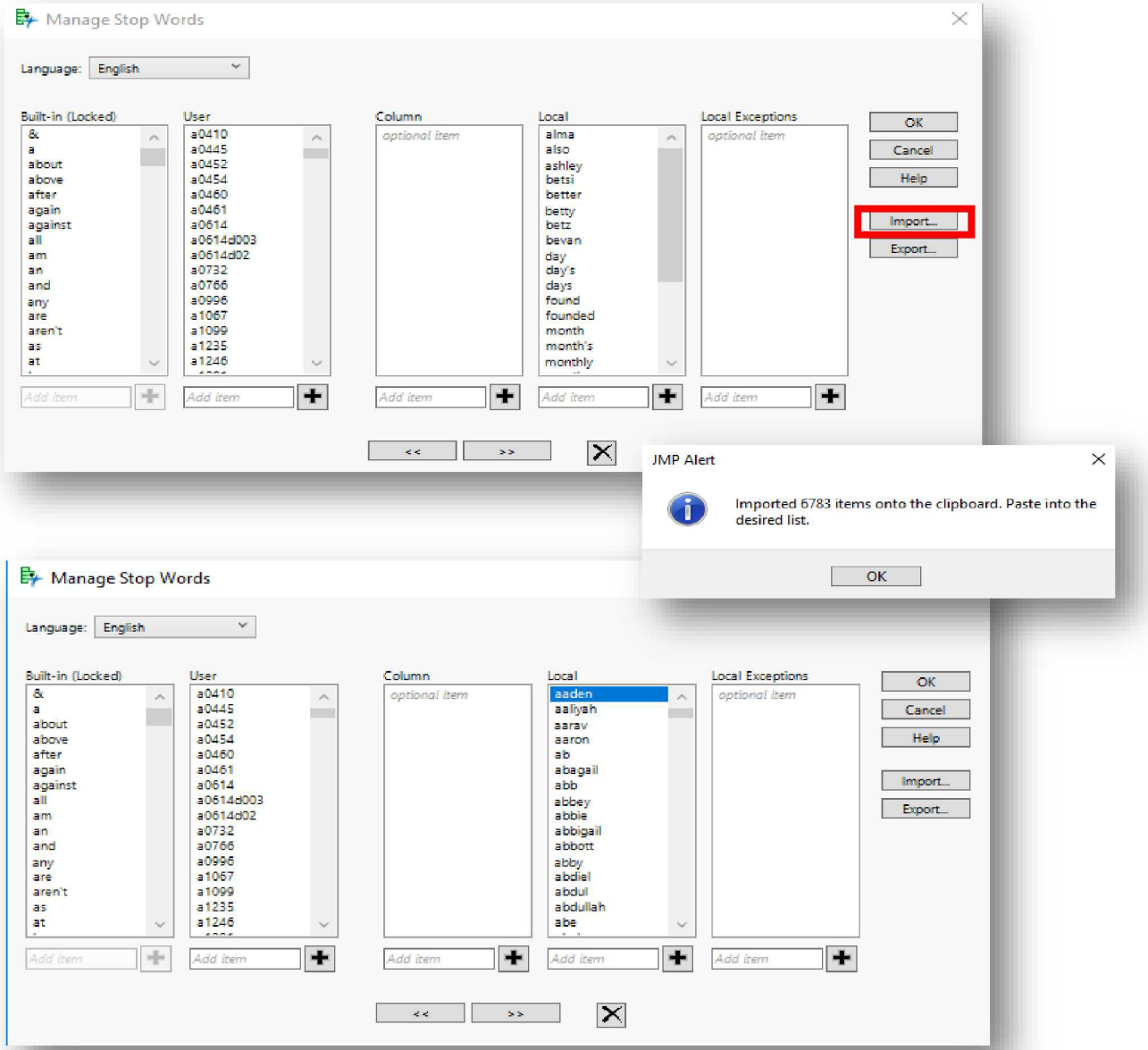
The interactivity provided by JMP allows users to scroll through the terms provided and note whether stemming (Term Options>>>Stemming) is an option and if so, whether to stem for combining (combining same word stems such as manage, managing, managed) or stem for all (mainly recommended when merging with other datafiles after cleaning).

Managing Stop Words, Recodes, Phrases and Stem Exceptions

JMP's ability to interactively modify on the fly in various ways, affords analysts the ability to focus more on the data and less on coding and subsequent debugging. Analysts add stop words and stem exceptions with a simple right click of a computer mouse. Applying filters are useful and speed up pre-processing time by allowing analysts to focus quickly on information that will answer a research question.

Another option available for JMP users is to import files when available that help streamline pre-processing when looking for stop words or phrases. This option allows for a quick import of lists (words) that can be used as stop words, stem exceptions or phrases which should or should not be considered important for answering the research question(s) with analysis. For example, the U.S. data breaches data from 2013-2018 has individual person names written within the summary text file that should not be considered as important terms for analysis in answering the research

question what are the risks and issues for data breaches in the United States? It would take a much larger portion of time for a researcher to manually go through and check the terms to eliminate person names. Even though JMP's interactive window allows users to quickly find words of interest such as names, to manually write code to omit the terms from analysis or to interactively search through tens of thousands of terms (e.g. the data breach data set), it requires a lot of time. This is true for any statistical platform one chooses, however the beauty of JMP allows users to quickly import without worrying about code. To solve this problem within the data breach file, a baby names csv was retrieved from GitHub (e.g. https://raw.githubusercontent.com/hadley/data-baby-names/master/baby-names.csv). The baby names csv was imported into JMP by first creating a new data file and because each year from 1880 to 2008 had duplicate popular names (e.g. John), duplicate rows were first selected (Rows>>>Row Selection>>>Select Duplicate Rows) and then deleted (Right click>>> Delete Rows). This takes our example file from GitHub from 258,000 names down to 6,782 unique names.

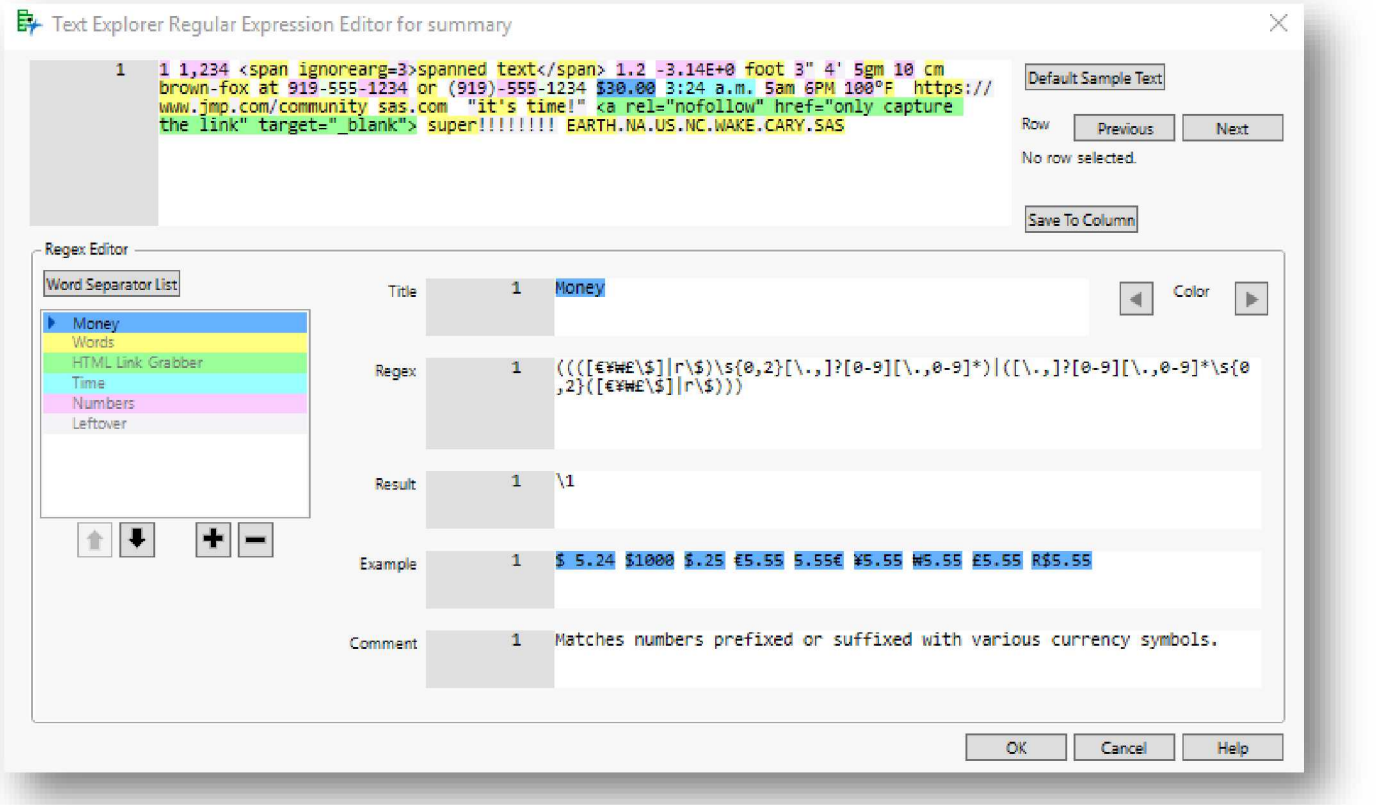


Capitalizing on JMP's Interactive Parsing Options

Besides the interactive term options covered, another interactive and highly useful group of options within JMP are the parsing options available. The powerhouse for pre-processing text comes with JMP's unique display for customizing regular expressions (Parsing Options>>>Customizing Regex). Programmers can save numerous amounts of time coding or manually omitting and cleaning data by using regular expressions. With a regular expression, one can perform powerful string parsing with only a handful of lines of code, or maybe even just a single line. A regular expression is faster to write and easier to debug and maintain compared to dozens or hundreds of lines of code to achieve the same result or manual cleaning by point and click. Thanks to JMP's interactive regular expressions window debugging and tweaking is streamlined and highly efficient.

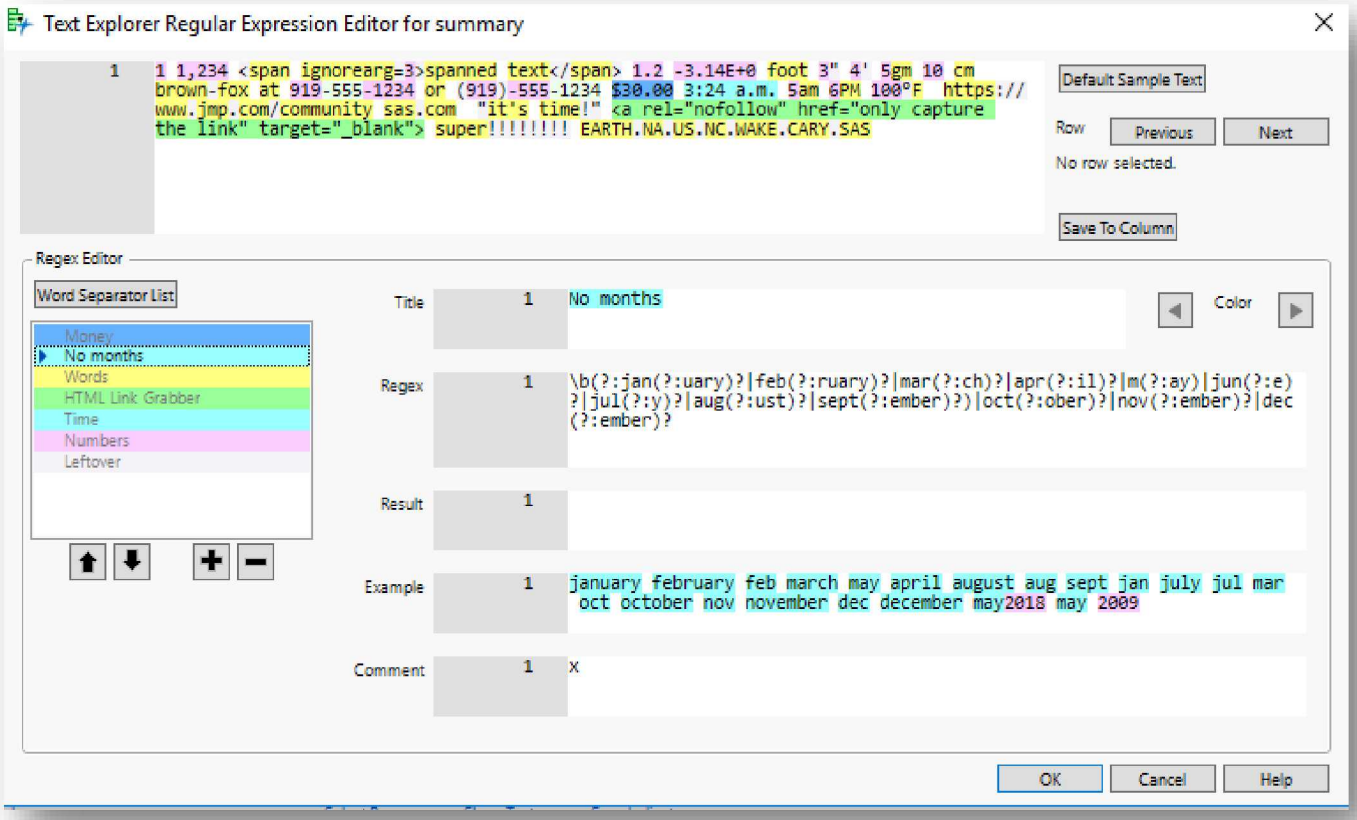
JMP's unique display for customizing regular expressions is a powerhouse!

JMP comes out of the box with built-in regular expressions and basic words (Parsing Options>>>Tokenizing>>>Regex/Basic Words). JMP automatically color codes each customized regular expression within the Regex Editor and shows which words are being selected under which custom regular expression. The ones listed here are JMP's built in regular expressions.



JMP also allows users to write their own regular expression and desired result (\1=keep; an empty result means omit) along with specific examples to make sure the regular expression is capturing what the user wants to capture by either omitting or keeping the result.

In using the data breach example, regular expressions were created to remove information that would not be considered as an issue or risk found within the summary which included months, years, web addresses, etc. Ordering these regular expressions (simply moving their location on the left navigation window) allows analysts to capture key information first and then omit or keep prior to running subsequent regular expressions that would have the opposite effect. For example, words are kept in the default regular expression, but to get rid of months which are in fact words, we would want to move the “No Months” regular expression up before the “Words” regular expression and have results become an empty cell.



While regular expressions can be difficult to tackle, JMP’s interactive layout for regular expression creation along with additional tools such as Regex Buddy and Regex Crossword, allow analysts to get a firm grasp on regular expressions by essentially pulling back the curtain on the mystery that is regular expressions and showing how efficient regular expressions are in pre-processing steps for cleaning data to get quality results.

Conclusion

JMP is a powerhouse in text analytics due to its visualizations, interactivity, and unique display for customizing regular expressions. The art of text analytics is enhanced with JMP by the fluid interaction between the analysts and the data and by the flexibility offered. JMP Text Explorer is an efficient and effective means to pre-processing as well as becoming familiar with any nuances of the data. Analysts can dive right into what stories the data tell from the unstructured text, without needing to worry about coding.

JMP minimized the time it takes to pre-process unstructured text.

Unstructured text is often difficult to navigate while pre-processing data. Using JMP's display options, term options and parsing options, especially using the import feature for stop words and phrases and regular expressions, analysts can effectively pre-process the data, significantly minimizing the time it when compared to other popular and widely used programs such as R or Python. Because JMP comes with built-in regular expressions and basic words, the software makes text analytics attainable to analysts without a coding background. While regular expressions can be difficult to tackle, JMP's interactive layout for regular expression creation along with additional tools available, allow analysts to get a firm grasp on regular expressions and how they can be used in pre-processing to clean data for quality results.

"JMP" into text analytics even without a coding background!

References

Goyvaerts, J. (2019). Regex Buddy. Retrieved from https://regular-expressions.info

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii International Conference on System Sciences* (pp. 1833-1842). IEEE.

IBM Watson Developer Cloud. (2016). *Tone Analyzer*. Retrieved from https://tone-analyzer-demo.ng.bluemix.net/

JMP® PRO, Version *14.3*. SAS Institute Inc., Cary, NC, 1989-2019.

JMP User Community. (2019). Retrieved from https://community.jmp.com/

Michelsen, M. H., &Michelsen, O.B. Regex Crossword. Retrieved from https://regexcrossword.com/

SAS Institute Inc. 2018. *JMP® 14 Basic Analysis*. Cary, NC: SAS Institute Inc. Retrieved from JMP Help>>>Books.

SAS Institute Inc. 2018. *Discovering JMP® 14*. Cary, NC: SAS Institute Inc. Retrieved from JMP Help>>>Books.

SAS Institute Inc. 2018. *JMP® 14 Scripting Guide*. Cary, NC: SAS Institute Inc. Retrieved from JMP Help>>>JMP Help

Utlaut, T. L., Morgan, G. Z., & Anderson, K. C. (2018). *JSL Companion: Applications of the JMP Scripting Language* (2nd ed.). SAS Institute Inc.

Wickam, Hadley. (2019). "US Baby Names 1880-2018." *US Baby Names 1880-2018.* GitHub, retrieved from www.github.com (https://hadley.github.io/babynames/)

*In addition to the references listed above, internal Sandia references include:

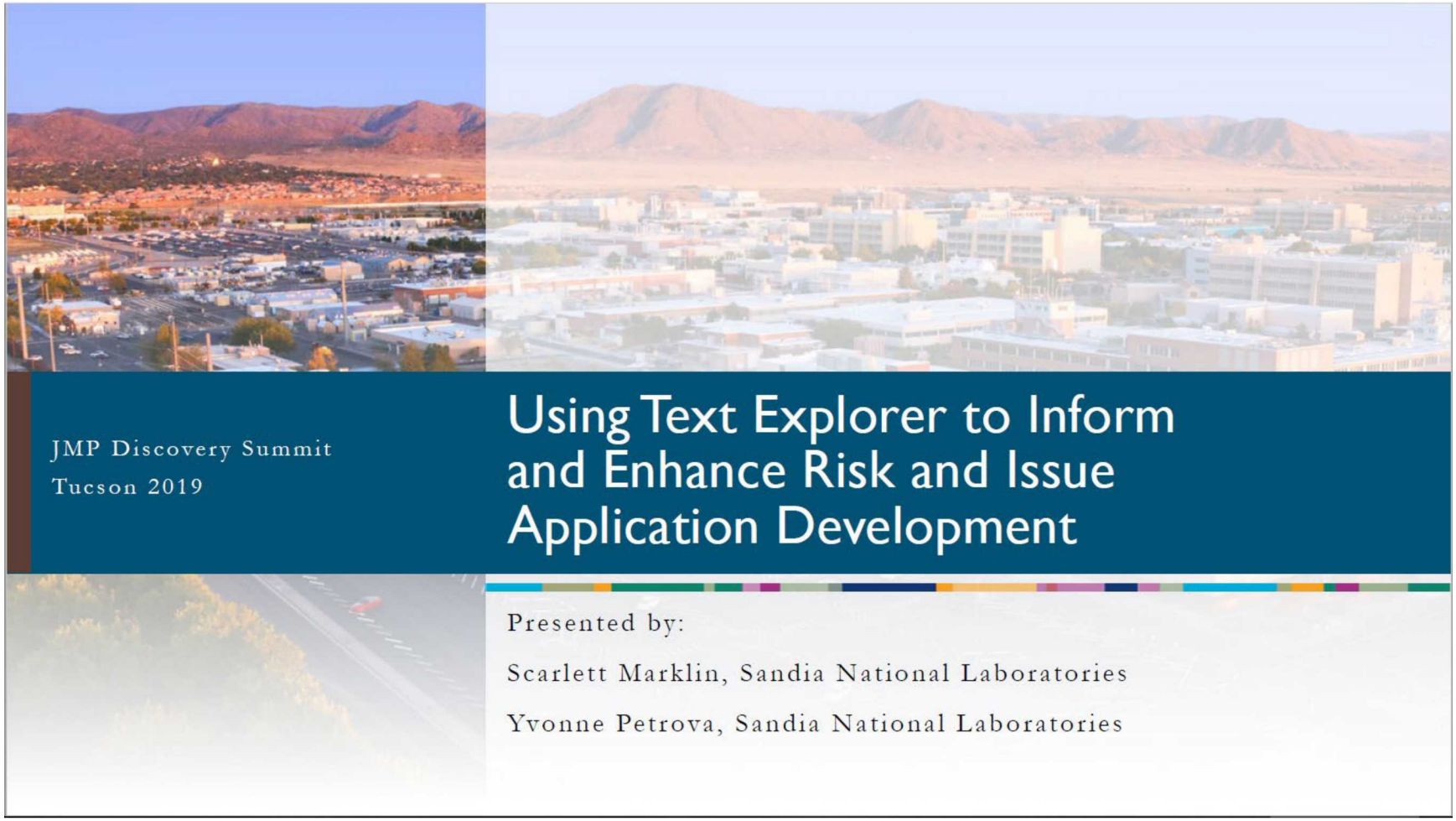
AIS User Community. (2014). Retrieved from https://sharepoint.sandia.gov/sites/AIS/SitePages/AIS%20home.aspx

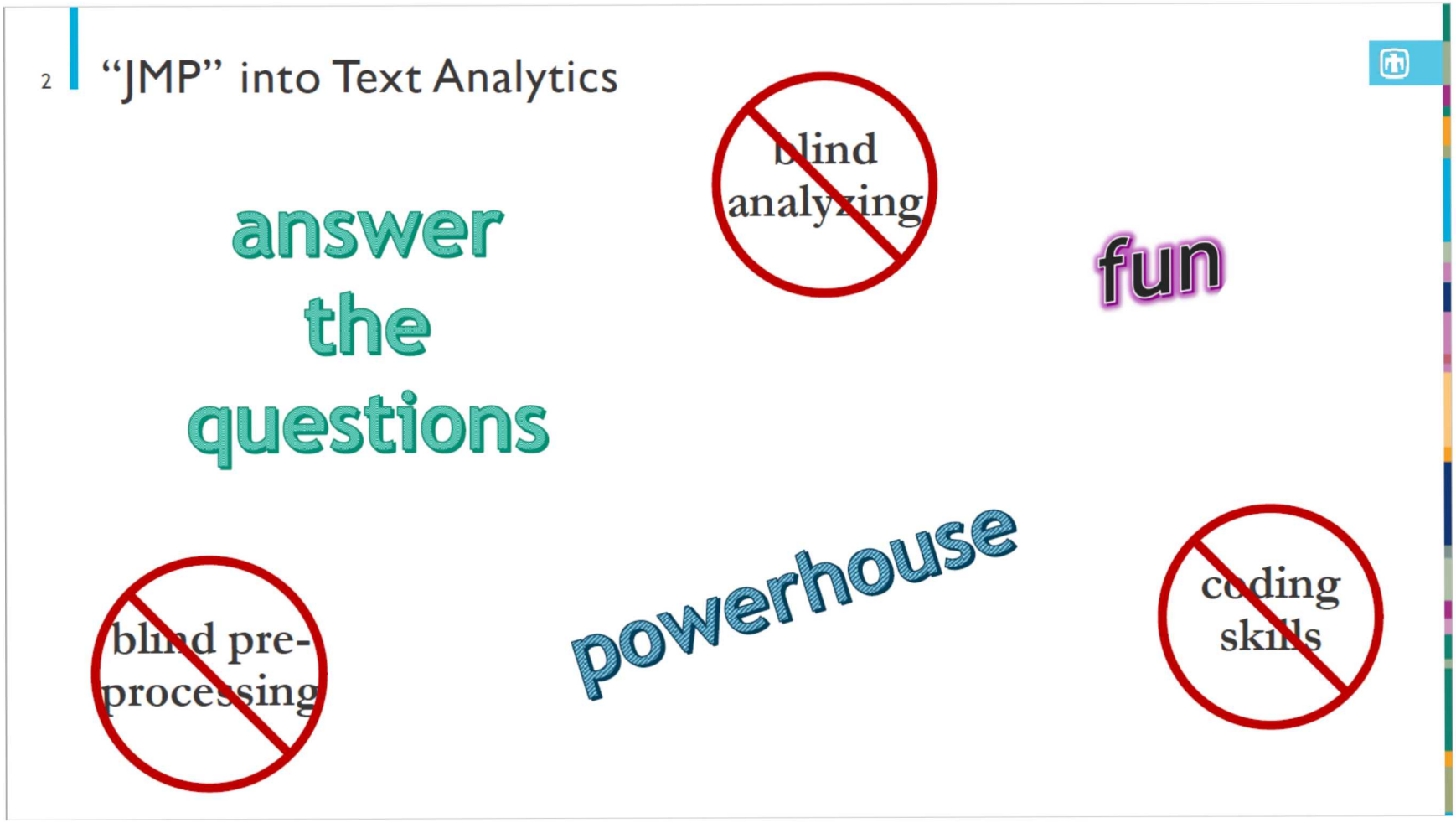
Marklin, S. (2018). *Bridging Lessons Learned from AIS to Oasis: Insights from Text Analysis Using AIS.* Sandia National Laboratories (Internal Paper).

Marklin, S., & Petrova, Y. (2018). *Review of Assurance Information System for Text Analytics.* Sandia National Laboratories (Internal Paper).

Appendix

JMP Discovery Summit, Tucson 2019 Presentation




2
"JMP" into Text Analytics
blind analyzing
answer the questions
fun
blind pre-processing
powerhouse
coding skills

3

The Problem at Sandia

- More consistent and accurate records of risks and issues
- Clear understanding of text from previous seven years
- Enhancements and more streamlined data capture
- Classify and identify categories of issues, risks, actions and results
  - Analyze each field separately
  - Analyze causal and not causal

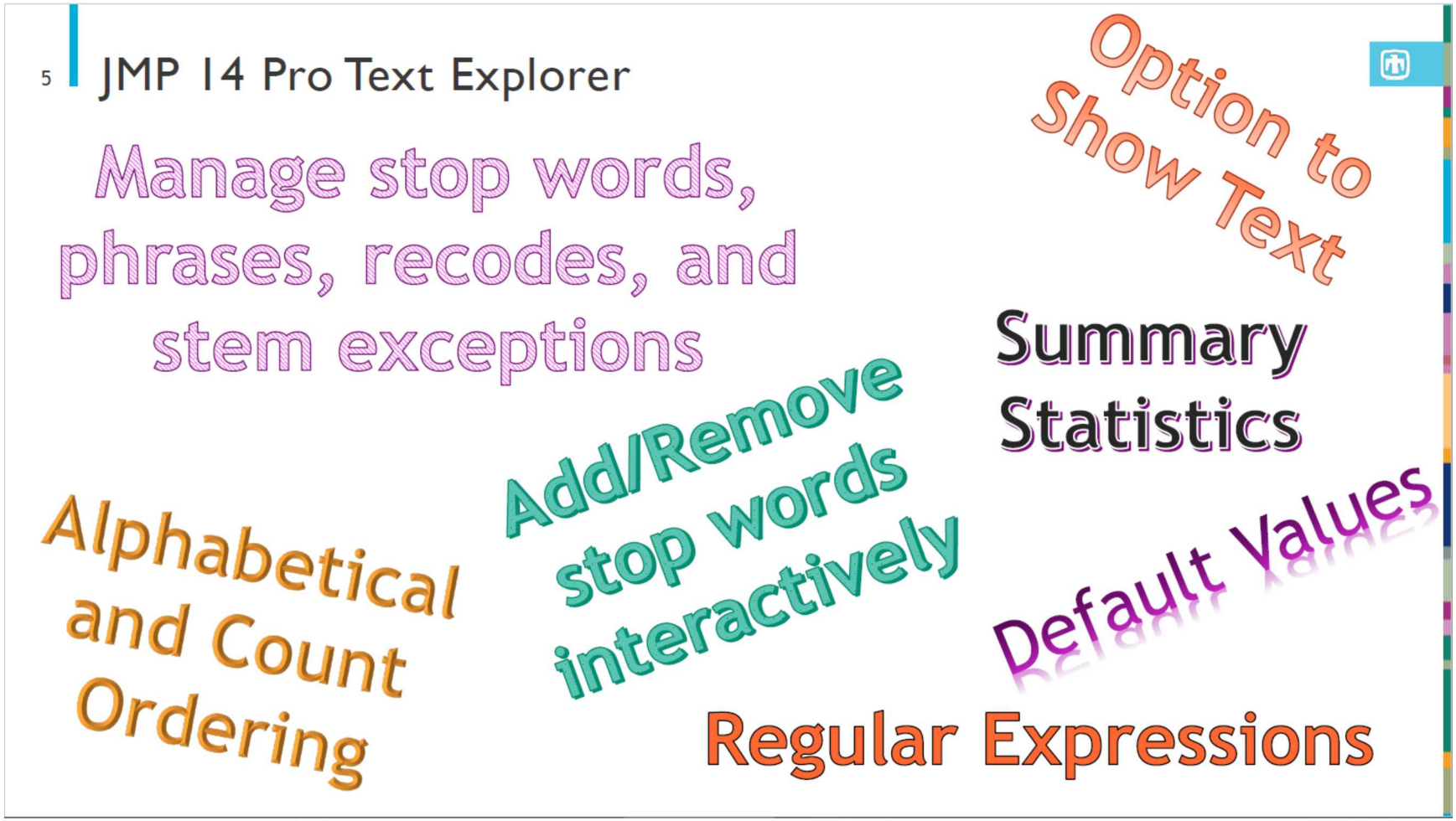
Now what?!?!

4

Data Quality

- Unstructured
- Difficult to navigate
- Empty fields, excessive text fields, references to other documents
- No baseline from which to start
- Automate pre-processing!

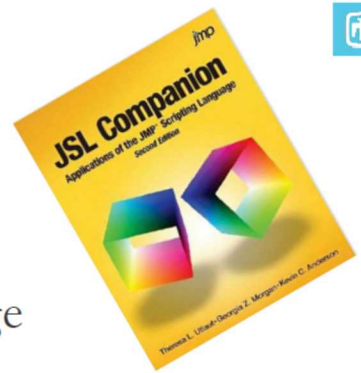
JMP 14 Pro Text Explorer


5
JMP 14 Pro Text Explorer
Manage stop words, phrases, recodes, and stem exceptions
Option to Show Text
Summary Statistics
Add/Remove stop words interactively
Default Values
Alphabetical and Count Ordering
Regular Expressions

6

Resources

- JMP Community https://community.jmp.com/
- JMP books within JMP itself
- JSL Companion: Application of the JMP Scripting Language
- Regex Buddy https://regular-expressions.info
- Regex Crossword https://regexcrossword.com/




7
Outcomes
Recommendations
More robust data
Categories
Improvements
Gaps & Inconsistencies
Stronger results
Enhancements