Tap Into Unstructured Data
Contents

Introduction .......................................................................................................................... 1
Text Exploration ................................................................................................................. 2
Background for the Example Data Set ............................................................................. 5
Part 1 - JMP® Text Explorer ............................................................................................ 6
Part 2 - JMP® Pro Text Explorer ..................................................................................... 13
Summary ........................................................................................................................... 18
References ......................................................................................................................... 18
Acknowledgements ........................................................................................................... 19
Introduction

In many of the training classes I teach, I like to introduce the participants to platforms by posing these questions:

*How many of you are collecting unstructured, text data?* The majority are collecting this type of data.

*Does it represent a large proportion of the data you are collecting in your organization?* The majority answer yes, with many saying up to 90 percent of the data collected.

*What are you doing with this data?* Silence.

*Why aren't you doing anything with this data?* I have had one person tell me because they did not think they would get any useful insights from this data. One.

*If you are not doing anything with this data, then why are you collecting and storing it?* Silence.

Is this you?

If you are collecting unstructured text data, then explore it, analyze it, and use it to find useful insights. Today.

This paper uses an easily understood example from a collection of National Science Foundation (NSF) research grant abstracts to illustrate Text Explorer. Because some users have JMP and others have JMP Pro, this paper separates the tools into two different parts: the first highlighting JMP Text Explorer capabilities in JMP, the second JMP Pro.

**Part I highlights capabilities in JMP Text Explorer.** The simplicity of string processing is presented first, specifically focusing on common obstacles: stop words, terms and (multi-word) phrases, misspelled words, and synonyms. Next, the paper demonstrates the usefulness of the graphical capabilities of Text Explorer to quickly uncover previously unknown information in unstructured data. Lastly, Part I shows how the information uncovered in Text Explorer can interact with other platforms to capitalize on the power of JMP.

**Part II highlights the capabilities in JMP Pro Text Explorer.** JMP Pro upgrades the capabilities of text exploration by providing a decomposition of the matrix containing documents and terms/phrases. The paper demonstrates the usefulness of topic extraction to delineate common themes in the document collection. In addition, Part II illustrates the use of JMP graphical capabilities to visualize how these themes depend on other (structured) variables like time or region of the country. Lastly, Part II shows how modeling of the information from topic extraction can provide answers to investigation questions from predictive analytics.
Text Exploration

Text exploration involves a series of steps to retrieve useful insights from a collection of unstructured text data.

The first step is to collect unstructured data into a JMP data table. In text mining, a document is a string of words, terms or ‘tokens’, while a collection of documents is often referred to as a “corpus.” These documents can come from multiple, disparate sources: text files, Word or Excel files, PDFs, internet, social media, etc. The JMP data table should be organized with each document as a row, each column containing either structured (a continuous or categorical variable) or unstructured data (text).

The next step to apply string and natural language processing to the column of unstructured text data. JMP Text Explorer uses what is called a ‘bag of words’ approach. This approach, which disregards the ordering of the words in each document as well as their ontological properties, is simplistic but effective. For string processing, JMP isolates individual words, removes punctuation, normalizes case (to lowercase) and removes numbers.

Natural language processing utilizes a pattern that occurs in the word counts of most corpora that is described by Zipf’s Law. Zipf’s Law states that the frequency of terms in a corpus is inversely proportional to its rank. The practical implication of Zipf’s Law is that a few terms will occur frequently while most terms will occur infrequently. Words that appear either quite frequently or quite infrequently are uninformative; frequently occurring words, such as the 5 million occurrences of the in the example below, are indiscriminate; while words that appear in only a small percentage of documents may not be indicative of common themes in the corpus.
JMP uses natural language processing tools to focus on the terms (or phrases) that analysts have the most interest in exploring. Words that may not provide any information about common themes in the corpus (like *the* and *an* and *in*) as well as words that may appear in every document, with no discriminating power (like the word *research* in the NSF grant example), are often referred to as *stop words*. Since these words can be uninformative, JMP will automatically remove the most common words (up front) and allow you to choose the minimum term frequency for text exploration. JMP has made it convenient to add stop words for a specific user, data table (local) or even column within a data table.

In addition, JMP can stem the text of different words that are conveying the same idea (*save*, *saves* and *saved* would be stemmed to *save*). Since certain terms should be analyzed together or grouped into phrases (*differential equations* should not be isolated into *differential* and *equations*), JMP can add these common phrases to the term list. Lastly, JMP can remove very short (single letters) or long words (that may be website addresses) and words that appear in only a few documents. This cuts down on the number of words, allowing you to focus on the most interesting terms. While some of these text exploration options are called within the analysis, many are in the dialog for the Text Exploration platform.
Next, JMP constructs a numerical representation of the text in the documents in what is called a document-term matrix (DTM). The DTM represents each document as a row and each term in the documents as a column.

There are various transformations of the DTM. The DTM shown above uses binary weighting: The entries represent if the word appears (1) or does not appear (0) in the document. In addition, JMP provides the ability to use other weighting schemes:

- **Ternary:** 0 if the term does not appear, 1 if the term appears once, 2 if the term appears more than once.
- **Term frequency:** How many times a term appears in a document.
- **Log:** The log of the term frequency.
- **Term-frequency inverse document frequency (tf-idf):** How many times a term appears in a document relative to the number of times it appears in the entire corpus.

There is no optimal weighting scheme; each provides different benefits based upon your particular corpus though binary and tf-idf seem to be most useful in practice.

Thanks to Zipf’s law, DTMs are usually very large and sparse (most entries are “0”). In the example described below, the DTM is a 304,725 (documents) by 359,326 (terms) matrix. This matrix would require more than 400GB of RAM to open in a JMP table. Therefore, text analysis utilizes dimension reduction techniques to reduce the DTM to a dense matrix with many fewer columns, attempting to preserve as much as the structure of the original DTM as possible without losing any information. JMP Pro uses singular value decomposition (SVD) to achieve this. The SVD of the DTM (a principle components analysis of the DTM, if the columns of the DTM are centered first) constructs a matrix which provides a rank-reduced description of the documents as well as a matrix which provides a rank-reduced description of the terms. These rank-reduced matrices can then be used to find relationships between documents as well as terms. Behind the scenes, JMP uses a “sparse representation” of the DTM that ignores entries of “0” to perform these calculations. This representation requires less than 1GB of RAM in the following example (as opposed to 400GB). It is the combination of the
memory-saving sparse representation of the DTM and the dimensionality-reducing SVD (along with some clever multi-threaded routines implemented by the JMP developers) that enables the seemingly magic ability of Text Explorer to automatically extract topics and themes from large collections of large documents.

The graphical and modeling capabilities of JMP can be used on the DTM as well as the matrices from SVD to provide useful insights from the unstructured text data. The JMP Partition platform can model a response variable as a function of the words in the DTM while Graph Builder can be used to visualize the results. Scatterplots can display relationships between both terms and documents. JMP Pro Text Explorer can also cluster terms and documents, and conduct topic analysis to see natural themes represented in the corpus. Modeling techniques can then be employed on these topics to determine if or which topics elicit effects on the response of interest.

All text exploration and analysis shown in this paper can be accomplished in the Text Explorer platform in JMP Pro. A subset of the analysis, the analysis that does not utilize SVD, can be accomplished in the Text Explorer platform in JMP.

Background for the Example Data Set

This paper uses a familiar example, abstracts from 304,725 NSF grant awards from 1988 to 2015. From the NSF website: “The National Science Foundation funds research and education in most fields of science and engineering. It does this through grants, and cooperative agreements to more than 2,000 colleges, universities, K-12 school systems, businesses, informal science organizations and other research organizations throughout the United States. The Foundation accounts for about one-fourth of federal support to academic institutions for basic research. NSF receives approximately 40,000 proposals each year for research, education and training projects, of which approximately 11,000 are funded.” https://www.nsf.gov/funding/aboutfunding.jsp

For this paper, the primary investigation questions are to determine the highest-frequency themes in grant research abstracts, if these themes are (US) region- or time-dependent, and which themes are awarded the most (and least) grant funding. JMP Text Explorer can find the most common words and phrases in the abstracts. Graph Builder can determine if the frequency of these terms change over time and determine if these are related to a specific area of the US Partition can model the amount of grant funding versus individual words and phrases to see if particular words/phrases elicit higher grant awards. JMP Pro Text Explorer can use information from the matrices constructed using the SVD to conduct topic analysis, capturing underlying themes in the corpus. Graph Builder can determine if these topics change over time or are related to a specific area of the US Fit Model can then model the grant fund award versus the topics to see if certain topics bring in more money than others.
Part 1 - JMP® Text Explorer

The first step is to collect unstructured data into a JMP data table. The NSF awards were downloaded from [https://www.nsf.gov/awardsearch/](https://www.nsf.gov/awardsearch/) as Excel files, then imported into JMP.

After changing the data type of Abstract to Unstructured Text, Text Explorer is launched.

Notice specific defaults are changed to make the analysis more efficient:

- Minimum Characters per Word is changed from 1 to 3.
- Maximum Characters per Word is changed to 20.
- All terms are stemmed.

After running the analysis, the Term and Phrase lists appear.
The most frequent term is research, while the most frequent phrase is graduate students. Many of the high-frequency words provide no information about common themes in the corpus; these stop words can be removed from the analysis. The dot next to words in the term list is indicative of stemming. Many of the phrases are relevant to common themes in the corpus; therefore, these phrases should be added to the analysis. Notice two phrases, computer science and united states, are colored red. These are indicative of common phrases recognized by JMP; hence they are already added to the analysis. The Term and Phrase Lists can quickly convey information about the most frequent themes in the corpus. Notice terms such as model and program as well as phrases such as large scale and computer science are frequent in the corpus.

Since research provide no discriminatory power in a list of abstracts, it can be easily removed from the analysis by adding it to the stop word list.
The top term in the list is now **project**, while the second is **use**. The term **use** is a result of stemming of ten words.

The phrase **large scale** frequently occurs in the corpus. The context in which the phrase is used may be unknown. Fortunately, the contextual use of that phrase can easily be shown in JMP Text Explorer with the Show Text command.
There are many phrases that provide information on common themes in this corpus of abstracts. Therefore, these phrases can easily be added to the text analysis.

Many times, analysts need to correct misspelled words or ‘tell’ the text analysis that two words are synonyms. These updates to the unstructured text can easily be completed using the Recode function in JMP. The Recode function is a local function in JMP Text Explorer.

Two terms are selected for Recode, techniqu- and approach-. Notice you can recode individual words to recode; you do not have to recode all words that were stemmed.

A word cloud is a graphical representation of the frequency of terms used in the text analysis. In this example, the word cloud (below) is colored by the amount awarded to date. Both the layout and coloring of the Word Cloud are options in Text Explorer.
Next, a subset of the document-term matrix (with some of the most frequent words) can be saved back to the data table for analysis. Notice the binary weighting scheme is used for this analysis. Alternatively, you can save indicators for specific terms and/or phrases; indicators are the same as the binary weighting scheme.

The corpus contains abstracts from 304,725 National Science Foundation (NSF) grant awards from 1988 to 2015. Remember, one of the primary investigation questions was to determine if the highest-frequency themes in research are time-dependent. Graph Builder can be used to graphically explore this research question.
Notice the phrase *big data* became prevalent around 2013, while the prevalence of *gene expression* has been declining (as a percentage of total grants in a given year) since the early 2000s. However, there has been no significant change in the prevalence of the phrase *material science* over time.

Additionally, one of the primary investigation questions was to determine if the highest-frequency themes in research are (US) region dependent. Graph Builder can be used to graphically explore this research question.

Notice *sea ice* seems to be much more prevalent in Alaska than other states.

Another one of the primary investigation questions was to determine which themes were awarded the most grant funding. Once the DTM is saved back to the data table, JMP’s Partition platform can easily be used to answer this question. For this particular investigation question, researchers may be interested in finding specific terms or phrases that are not subject-dependent. For example, if I am doing research in ecology, knowing that the term *nanotubes* leads to higher awards does not help me write an effective proposal. However, knowing the phrase *broader impacts* leads to higher awards does help me.
Certain terms lead to higher (average) awards: *partnership·*, *integr·* and *rapid·*. It would be imperative to analyze the contextual use of these particular terms. For example, *integr·* is the stemmed word for 27 different words. However, the effectiveness of the term *partnership·* is obvious.

Graph Builder can be used to graphically display information gained from modeling in Partition. Notice the use of *partnership·* in an abstract is correlated with much higher award amounts.

JMP Text Explorer can be used to answer all three of the primary investigation questions: the highest-frequency themes in research, if these themes are (US) region- or time-dependent, and which themes are awarded the most grant funding. However, the additional capabilities in JMP Pro can provide much more insight into these questions.
Part 2 – JMP® Pro Text Explorer

In a corpus, many themes will emerge, some dominant, others not. These themes will contain multiple terms that are used together frequently in the corpus. For example, terms like delivery, good command, above-average velocity and fastball will appear in scouting reports for baseball pitchers, while terms like instincts, tools, bat speed and throws easy will appear in scouting reports for baseball fielders.

It is useful to classify these common themes into subsets for understanding of a large corpus.

In JMP Pro, this can be accomplished by clustering terms and documents. Clustering terms provides an indication of words that are frequently used together within the same document. Clustering documents provides an indication of documents that have common, frequently occurring terms. A more efficient method to classifying common themes can be accomplished using topic analysis.

Topic analysis can be used to answer our primary investigation question, which is to determine the highest-frequency themes in the NSF abstracts from 1988 to 2015. The meaning of the ordering of the topics depends on both the weighting of the underlying DTM and on the use of scaling when calculating the SVD. In the next screenshot, the topics produced from the binary DTM with no centering and no scaling of the SVD roughly correspond to the prevalence of those topics across the entire corpus. These may be summarized as dominant general themes. In the 304,725 NSF abstracts, these include computing performance, devices, geometry, genes, conferences, organic chemistry and networks.
When the tf-idf transformation is applied to the DTM and “centering” is selected for the SVD, the topic order no longer corresponds to frequency of occurrence throughout the corpus. General themes descend further down while dominant specific themes rise to the top. In the NSF example, these specific themes include algebraic geometry, gene mutation, political elections, nanoparticle materials, volcanic magma and eruptions, and climate change.

The two other investigation questions can be easily answered with Text Explorer. The first is which themes were awarded the most (and least) grant funding. The second is if these common themes were (US) region- or time-dependent.

Additionally, the topic scores plots can be used to isolate the most active documents from the corpus for that topic.
To determine which themes were awarded the most (and least) grant funding, use the modeling capability of JMP with amount awarded to date as the response with topics as the factors.

The Sorted Parameter Estimate report not only provides the most significant topics, it also provides a direction. Notice topic 70 and 20 have positive effects on the amount of money awarded to the grant.

The theme in topic 70 is STEM education and collaborations between industries and universities. The theme in topic 20 is sediment in Pacific Ocean shelf.

Notice topics 40 and 15 have negative effects on the amount of money awarded to the grant.

The theme in topic 40 is developing curriculum and instructional (teaching) material for students in the area of math. The theme in topic 15 is attending conferences, workshops and symposiums as a participant or speaker.
Graph Builder can be used to determine if these common themes are (US) region- or time-dependent.

First examine topics by state. It is not surprising that topic 11 (volcano eruptions) is prevalent in Hawaii.

Now, evaluate topics over time with states. Topic 36 is about scientific and engineering principles in biology and ecology. This is a general, familiar topic.
Evaluating topic 36 over time shows it has slowly grown in prevalence from 1988 to present day in all regions.

Although there are many dominant themes, many themes seem to be related. For example, examine topics 48 and 54.

The theme for both topics is around atmospheric conditions. JMP can cluster these terms to even further delineate the highest-frequency themes in research.
Evaluating the four topics in cluster 1 shows the theme for this cluster is the atmosphere. Notice this cluster includes both Topics 48 and 54 from above.

<table>
<thead>
<tr>
<th>Term</th>
<th>Score</th>
<th>Term</th>
<th>Score</th>
<th>Term</th>
<th>Score</th>
<th>Term</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ocean-sea</td>
<td>0.67420</td>
<td>aerosol</td>
<td>0.4141</td>
<td>cloud</td>
<td>0.7822</td>
<td>convect-</td>
<td>0.27622</td>
</tr>
<tr>
<td>marin-</td>
<td>0.18936</td>
<td>atmospher-</td>
<td>0.3513</td>
<td>aerosol</td>
<td>0.2674</td>
<td>radar</td>
<td>0.23591</td>
</tr>
<tr>
<td>water</td>
<td>0.14896</td>
<td>oxide</td>
<td>0.2607</td>
<td>ice</td>
<td>0.1586</td>
<td>storm</td>
<td>0.23497</td>
</tr>
<tr>
<td>mix</td>
<td>0.13967</td>
<td>cloud</td>
<td>0.1661</td>
<td>ice</td>
<td>0.1490</td>
<td>forecast</td>
<td>0.22269</td>
</tr>
<tr>
<td>flux</td>
<td>0.12874</td>
<td>transport-</td>
<td>0.1502</td>
<td>transport-</td>
<td>-0.1014</td>
<td>weather</td>
<td>0.22193</td>
</tr>
<tr>
<td>sedimen-</td>
<td>0.11997</td>
<td>air</td>
<td>0.1409</td>
<td>droplet</td>
<td>0.0987</td>
<td>precipitat-</td>
<td>0.20444</td>
</tr>
<tr>
<td>shelf</td>
<td>0.11139</td>
<td>gas</td>
<td>0.1297</td>
<td>ecosystem</td>
<td>0.0976</td>
<td>observ-</td>
<td>0.17100</td>
</tr>
<tr>
<td>ridg-</td>
<td>0.10372</td>
<td>emiss</td>
<td>0.1219</td>
<td>comput-</td>
<td>0.0931</td>
<td>cloud</td>
<td>0.16870</td>
</tr>
<tr>
<td>pacif-</td>
<td>0.10207</td>
<td>nitrogen</td>
<td>0.1219</td>
<td>hurricane</td>
<td>0.16133</td>
<td>wind</td>
<td>0.15538</td>
</tr>
<tr>
<td>circl-</td>
<td>0.10207</td>
<td>cluster</td>
<td>-0.1078</td>
<td>heat</td>
<td>0.13126</td>
<td>tropic</td>
<td>0.12645</td>
</tr>
<tr>
<td>ozon-</td>
<td>0.1055</td>
<td>ozon-</td>
<td>0.0931</td>
<td>ionosphere</td>
<td>0.12409</td>
<td>i-</td>
<td>0.12409</td>
</tr>
</tbody>
</table>

It is evident that Text Explorer in JMP Pro provides deep insight into all three of the primary investigation questions: the highest frequency themes in research, if these themes are (US) region- or time-dependent, and which themes are awarded the most (and least) grant funding.

Summary

Most (if not all) organizations are collecting unstructured text data. If you are collecting this type of data, explore it and analyze it to quickly discover practical insights contained in the corpus. JMP Text Explorer not only has extraordinary capabilities, but also is extremely easy to use. Additionally, the information provided by Text Explorer can be further exploited in other JMP platforms.

The Text Explorer platform is in both JMP and JMP Pro. While JMP provides capabilities for both string and natural language processing, JMP Pro upgrades the capabilities of text exploration by providing information contained in the decomposition of the DTM for Topic Extraction and Document Clustering.

References

The NSF data set was downloaded from https://www.nsf.gov/awardsearch/.

Acknowledgements

Many thanks to the JMP team for developing the Text Explorer platform. Special thanks to two of my team members, Jim Wisnowski and Andrew Karl. Jim had a vision for the value of text mining to our customers. He was correct. In 2013, Andrew developed a JMP script that allowed us to achieve that vision. This tool was the precursor to JMP Text Explorer.

About the Author

HEATH RUSHING, co-founder and Principal Consultant at Adsurgo, is a spirited problem solver and leader with a wealth of expertise in quality engineering, applied statistical methods, Lean Six Sigma, operations research and training. He leads the Training and Consulting, Quality Engineering and Medical Applications divisions at Adsurgo. Much of his recent consulting engagements have been with the biopharma and medical device industries, integrating Quality by Design principles across enterprises. He also recently led development of Adsurgo’s highly acclaimed applied text mining tool that makes analysis of unstructured data accessible to a broader range of engineers and scientists. Previously, he was the Six Sigma Training Manager at SAS, earning awards as top external contributor to annual JMP software sales. Rushing was a Principal Quality Engineer for Amgen, where he developed and implemented numerous innovative statistical methods to advance corporate risk management, process characterization, technology transfer and analytical method validation. He also served as an officer in the US Air Force as a test and intelligence operations analyst and Air Force Academy professor.

Comments and questions:
W. Heath Rushing at Adsurgo LLC.
Email: Heath.Rushing@adsurgo.com
Phone: 720-536-0852 ext. 12