ANALYZING UNSTRUCTURED DATA: TEXT ANALYTICS IN JMP

Volker Kraft
SAS Institute – JMP Division, JMP Academic Team
volker.kraft@jmp.com

As much as 80% of all data is unstructured but still has exploitable information available. For example, unstructured text data could result from comment fields in surveys or incident reports. You want to explore this unstructured text to better understand the information that it contains. Text Mining, based on a transformation of free text into numerical summaries, can pave the way for new findings.
This example of the new text mining feature in JMP starts with a multi-step text preparation using techniques like stemming and tokenizing. This data curation is pivotal for the subsequent analysis phase, exploring data clusters and semantics. Finally, combining text mining results with other structured data takes familiar multivariate analysis and predictive modeling to a next level.

INTRODUCTION

In the digital world of today, the cost and the limitation barriers in storing and accessing data have been pretty much removed. In 2015, IDC Research estimated that 90% of this vast amount of digital data is *unstructured* data, including responses to open-ended survey questions, social media, email, maintenance reports and HTML web pages (Vijayan, 2015).

The new *Text Explorer* platform in JMP 13 does not only allow to learn from unstructured text data without programming, but can make the tedious cleanup of inherently messy text data quick, easy, and fun. As soon as the data are ready for analysis, you are typically interested in questions like

- Can we identify groups of documents that are similar to one another as a way of summarizing content or uncover themes?
- Are there any unusual documents ("outliers")?
- Can we discover relationships between structured and unstructured data?
- Can we improve predictive models by combining text data with other data sources?

We all know how to tackle these questions with more traditional, structured data. Text exploration in JMP amounts to transforming text data into a more traditional rectangular data format, which then supports classical multivariate analysis techniques like regression, CART or GLM.

It is amazing to see how much you can learn from text data just by using "simple" data mining techniques, without worrying about higher linguistic levels like syntax analysis or natural language understanding. Text Explorer is typical JMP and makes text analytics easily accessible, very interactive and visual. Statistical discoveries from unstructured text data can also be a lot of fun, especially in teaching data mining or multivariate analysis courses.

Let us introduce the basic terminology and workflow of text analytics in JMP, by looking at a real-world example.

BASIC TERMINOLOGY AND WORKFLOW

The JMP table "NTSBAccidentReports2001-2003.jmp" (Fig. 1) provides data about 3,235 accidents in air traffic, collected between 2001-2003 by the National Transportation Safety Board in the US. One column, 'narr_cause', is formatted as "unstructured text". This column is called a *corpus*, and each cell stores a *document*. One task in this example is to explore relationships between the (unstructured) reports and other (structured) information like month of year or weather conditions incl. temperature and wind. Another goal is predicting fatality (YES/NO) of an accident based on the text information stored in accident reports.

In: A Molnar (Ed.), *Teaching Statistics in a Data Rich World*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2017, Rabat, Morocco. iase-web.org/Conference_Proceedings.php

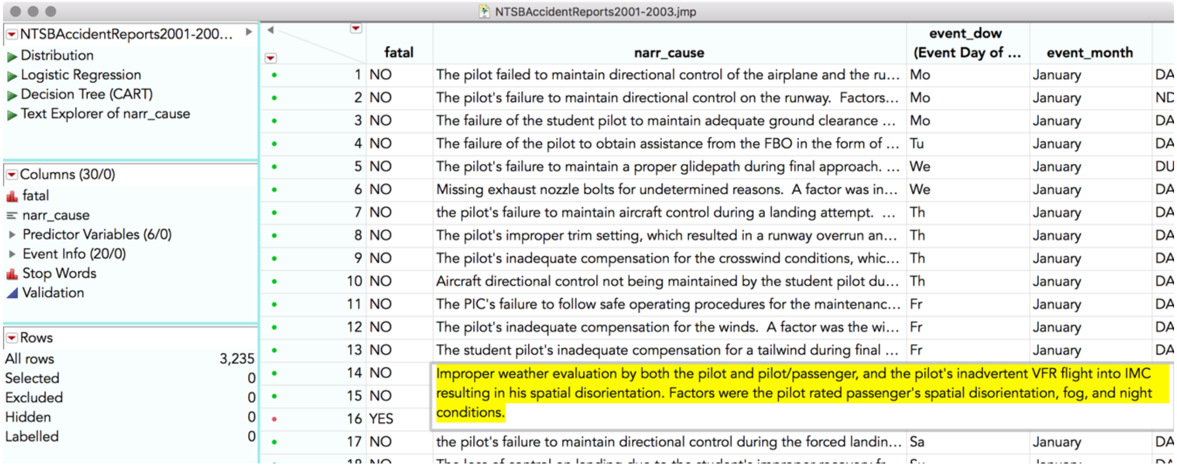


Fig. 1: NTSB Accident Reports, with unstructured text column 'narr_cause' (=corpus), containing 3,235 cells (=documents)

According to Klimberg et al. (2016) the process of text analysis can be split into three phases:

1. *Term Creation:* This phase does all text cleanup and develops the so-called *document term matrix* (DTM). The DTM is a set of indicator variables that represent the *terms* (columns) in the documents (rows). Terms are all words and phrases which will be considered in the analysis phase. Techniques like tokenizing, phrasing and terming are used to initially develop the DTM. Subsequently, you explore the set of variables and curate the DTM, by grouping words or removing infrequent words, until you are satisfied.
2. *Text Analysis:* Text visualization and the text multivariate techniques of clustering, principal components, and factor analysis are used to understand the composition of the DTM.
3. *Explore relationships and predict outcomes:* If a dependent variable exists (here: 'fatal'), you can use the text multivariate analysis results, along with other structured data, as independent variables in a predictive technique.

PART I: TERM CREATION

We launch Text Explorer from the Analyze menu in JMP and select 'narr_cause' as the Text Column for analysis (Fig. 2). All other settings are (carefully preconfigured) defaults. Both stemming and regular expressions for tokenizing are advanced options and not used here.

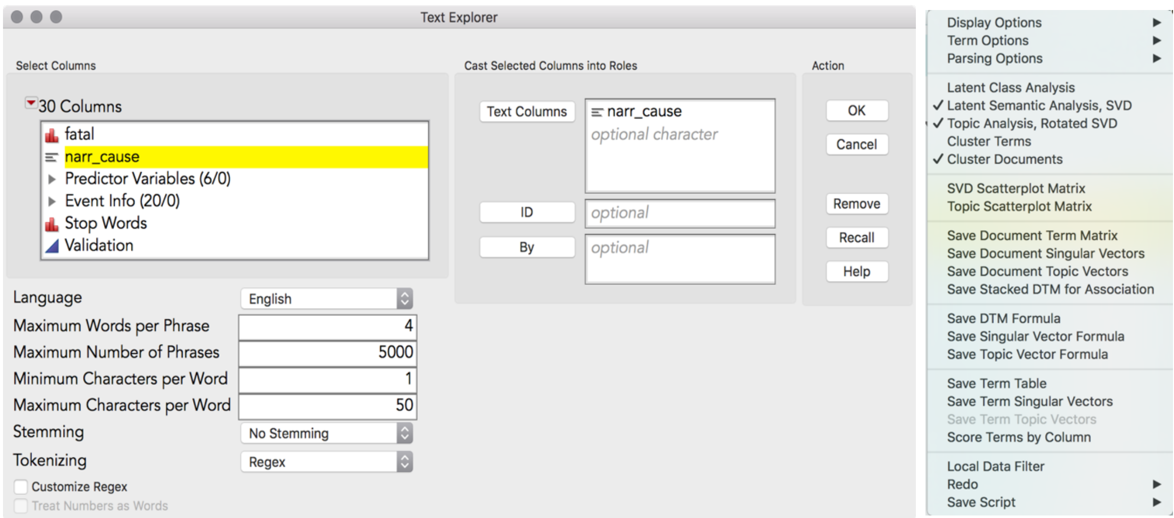


Fig. 2: Launching Text Explorer to analyze text column 'narr_cause' (left). Top red triangle menu in Text Explorer report (right).

The summary in Fig. 3 tells us that 3235 documents were parsed into 85675 tokens, representing 1328 different terms. The terms used most often are ‘landing’ and ‘failure’. Some terms were manually removed from the term list (right-click > “Add Stop Word”).

Phrases are word sequences which occur more than once in the corpus. The phrase used most often is ‘failure to maintain’, occurring 844 times and consisting of three words. The grey phrases have been added to the term list (right-click > “Add Phrase”), since they represent important concepts. ‘Landing gear’ in red was already added as a system phrase.

NTSBAccidentReports2001-2003 - Text Explorer of narr_cause

Text Explorer for narr_cause

Number of Terms	Number of Cases	Total Tokens	Tokens per Case	Number of Non-empty Cases	Portion Non-empty per Case
1328	3235	85675	26.4838	3235	1.0000

Term and Phrase Lists

Term	Count
landing	1467
failure	992
factor	989
resulted	853
failure to maintain	844
flight	831
factors	810
contributing	711
terrain	701
accident	638
inadequate	562
resulting	547
control	506
loss of engine power	451
conditions	441

Phrase	Count	N
failure to maintain	844	3
engine power	479	2
loss of engine	457	3
loss of engine power	451	4
directional control	396	2
contributing factor	339	2
forced landing	334	2
maintain directional control	306	3
maintain directional	306	2
contributing factors	248	2
failure to maintain directional	238	4
undetermined reasons	223	2
landing gear	185	2
inadvertent stall	154	2
landing roll	154	2

Fig. 3: Top terms and phrases after removing some terms and adding phrases

The term list can be visualized by a word cloud (Fig. 4). Top terms and phrases (added to the term list) are emphasized by font size. Everywhere where you see a term or phrase (e.g. in Fig. 3 and Fig. 4) you can right-click and choose “Show Text” to see the contextual information. This is very helpful if you want to understand how a single term or phrase is used.

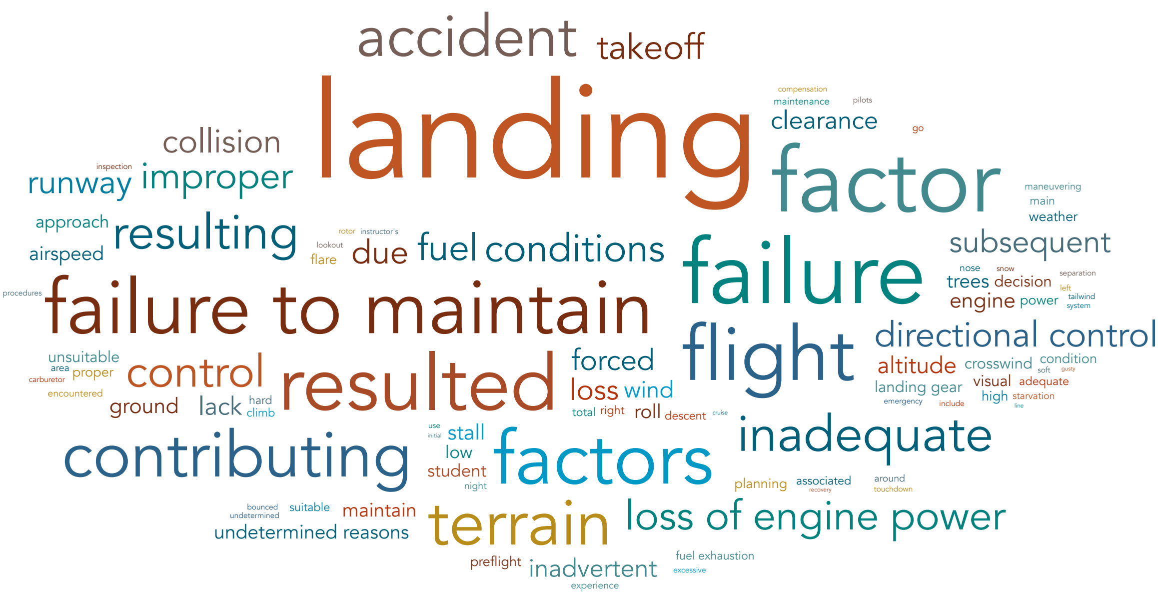


Fig. 4: Word Cloud

PART II: TEXT ANALYSIS

Figure 5 shows the results from the dimension reduction in the document space (left, DTM rows) and term space (right, DTM columns), called *Latent Semantic Analysis*. Similar to Principal Component Analysis in multivariate analysis, Singular Value Decomposition (SVD) is used in text analysis to take advantage of the sparse nature of the DTM. With only two dimensions both related documents and terms group nicely. Results like SVD vectors or SVD Matrix can be exported for further analysis by other JMP platforms.

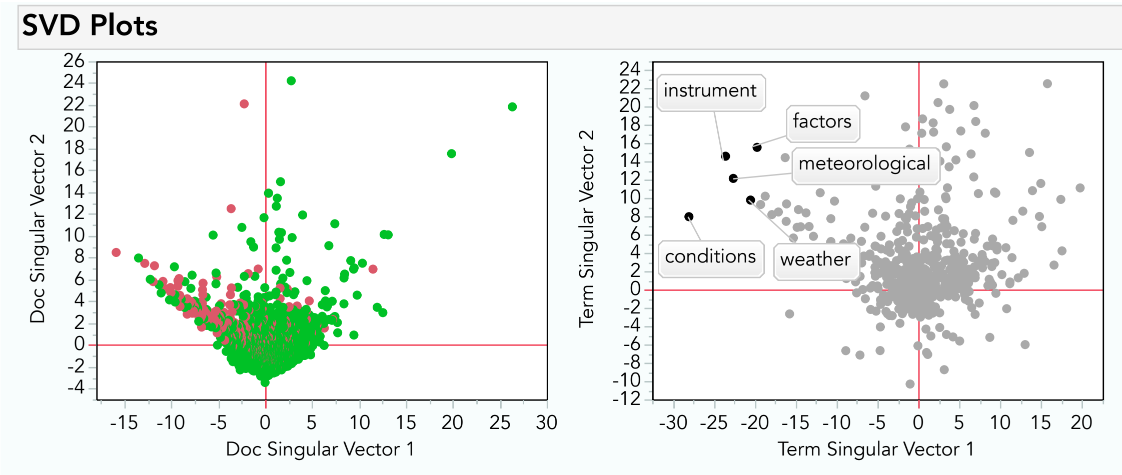


Fig. 5: Two-dim. document and term vector spaces after Latent Semantic Analysis (SVD)

Topic Analysis performs a varimax rotated singular value decomposition of the DTM to produce groups of terms called topics. This corresponds to Factor Analysis in multivariate analysis. In this example, Topic1 seems to represent the weather theme, while Topic6 is about maintenance.

Documents can be visually selected based on their topic scores, and more detailed relationships can be discovered by a combined scatterplot matrix of the rotated SVD vectors (topics) in the term and document space.

Topic Words

Topic1		Topic2		Topic3		Topic4		Topic5	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
conditions	0.26629	n89803	0.42364	forced	0.2690	midair	0.36219	directives	0.32031
instrument	0.26562	rolling	0.41177	fuel	0.2547	way	0.34063	local	0.28624
meteorological	0.25200	set	0.31753	loss of engine power	0.2336	yield	0.34063	controller	0.26957
continued	0.21365	cockpit	0.30131	starvation	0.1981	atc	0.27992	faa	0.25078
vfr	0.21248	insure	0.26164	suitable	0.1819	lower	0.24830	follow	0.23959
weather	0.20342	front	0.25694	due	0.1674	lookout	0.24036	procedures	0.23777
factors	0.19526	forward	0.23763	terrain	0.1610	airplanes	0.22820	taxiing	0.20765
flight	0.19308	brakes	0.21348	fuel exhaustion	0.1498	radio	0.22003	cross	0.16843
fog	0.18825	attention	0.21054	unsuitable	0.1461	visual	0.21812	md	0.15902
ceilings	0.17450	parked	0.20675	failure to maintain	-0.1445	extra	0.19409	tower	0.15873
imc	0.16480	colliding	0.18631	lack	0.1323	instructions	0.17031	company	0.13894
night	0.16343			preflight	0.1258	collision	0.15169	personnel's	0.13476
adverse	0.15879			selector	0.1253				
low	0.15622			tank	0.1227				
dark	0.15023								

Topic6		Topic7		Topic8		Topic9		Topic10	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
maintenance	0.21821	pattern	0.37247	supervision	0.26775	stall	0.2194	runway	0.21757
personnel	0.20892	entry	0.28841	remedial	0.23798	airspeed	0.2098	go	0.17635
main	0.20867	traffic	0.28284	action	0.22390	landing	-0.2008	proper	0.17585
landing gear	0.20566	using	0.26881	instructor's	0.21273	altitude	0.1692	around	0.17463
service	0.16827	without	0.25605	instructor	0.21045	compensation	-0.1453	point	0.16389
company	0.15817	executing	0.25288	student	0.19544	inadvertent	0.1433	factors	0.15093
bulletin	0.15699	insure	0.24649	student's	0.18438	roll	-0.1412	touchdown	0.14621
right	0.15016	manual	0.24269	dual	0.18007	adequate	0.1310	wrong	0.12354
actuator	0.14899	information	0.23415	certified	0.16579	crosswind	-0.1297	end	0.12053
left	0.13669	correct	0.22607	inadequate	0.16435	low	0.1286	accident	0.11897
inspection	0.13135	pilots	0.22601	cfi	0.15321	wind	-0.1271	attain	0.11781
collapse	0.12798	operating	0.19662	delayed	0.14908	directional control	-0.1269	tailwind	0.11619
assembly	0.12746	adequately	0.17632	flight	0.13495	spin	0.1201	decision	0.11271
attach	0.12503					conditions	-0.1102	overrun	0.11254

Fig. 5: Topic Words for 10 topics

Other analysis options include *Clustering Analysis* of terms and documents. The number of clusters can be set and cluster membership can be saved to the existing (Cluster Docs) or a new data table (Cluster Terms).

PART III: DISCOVER RELATIONSHIPS AND PREDICTIVE MODELING

The Text Explorer top triangle menu (Fig. 2, right) shows several options for saving the DTM or any analysis results. This allows you to explore relationships between your unstructured and structured data, or to use numeric representations of your text data as independent variables in predictive modeling.

The result from the term creation phase, the DTM, can be directly added to the data table. The number of terms can be chosen in order to reduce the dimensionality of the matrix. An interesting option is the type of weighting factor stored in indicator columns: Compared to 'Binary' (term is used or not), settings like 'TF-IDF' (Term-Frequency Inverse-Document-Frequency) can be very powerful, see Help Text Explorer (2017).

Figure 6 shows all documents over time (X-axis) and split by 'fatal' (Y-axis) and cluster 1-10 assigned by Cluster Documents. This result suggests for instance to investigate some special cases like Cluster 4 or 8. For cluster 9 fatality could be reduced over time.

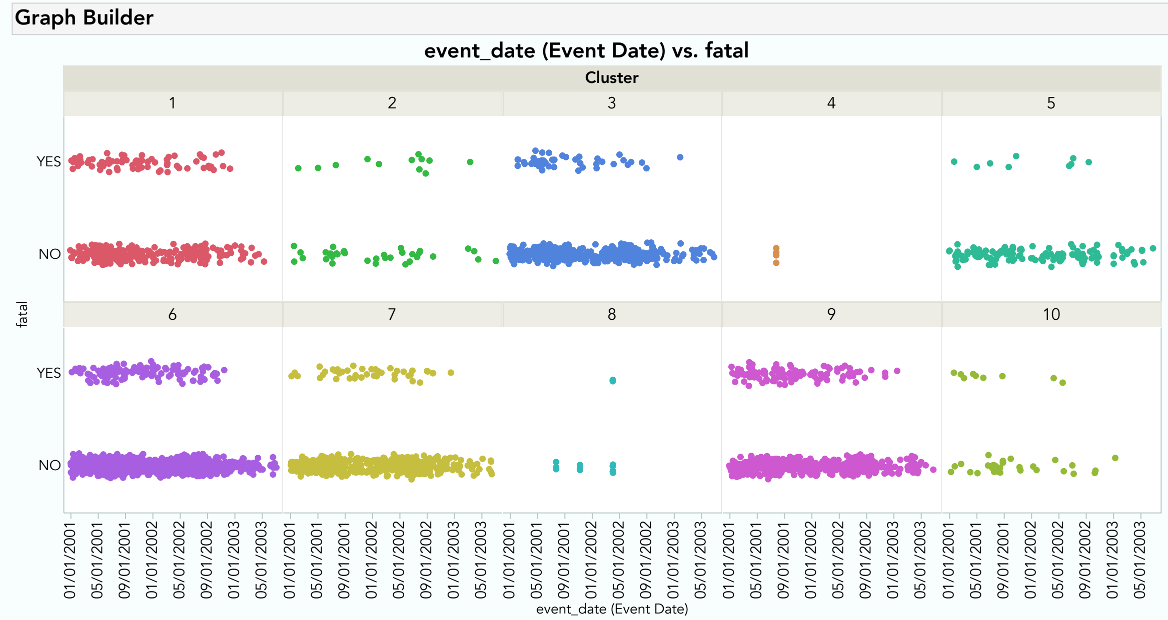


Fig. 6: Combining text analysis results (here wrap by 'Cluster' from Cluster Documents) with other data (Here Y='fatal' and X='event_date')

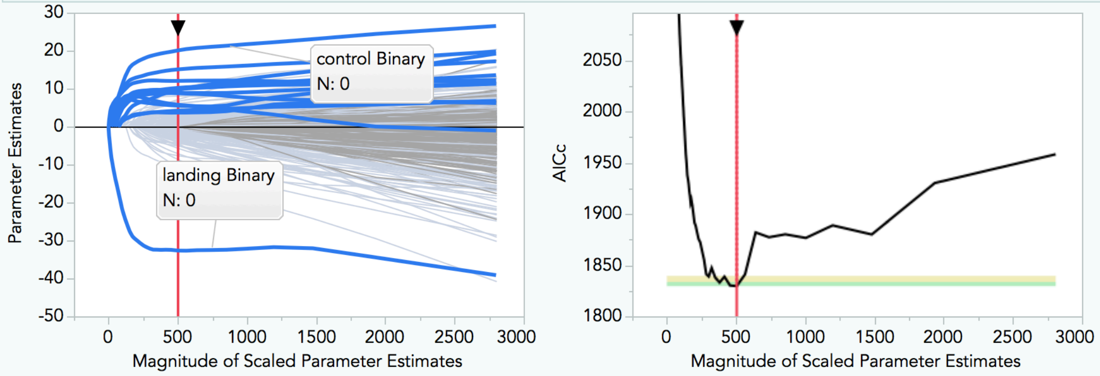
Figure 7 shows an example for a predictive model solely based on text information: Using the Generalized (Penalized) Regression personality in Fit Model, we built a model predicting 'fatal' using all(!) indicator variables from the DTM. The LASSO method helps with variable selection: By moving the red slider on the left towards zero, the penalty is increased and more variables are removed from the model. The validation plot on the right helps to select the "best" model, here based on AIC corrected.

The model tells us that the term 'landing' (with binary weights) is highly significant and negative. Probably this means that we don't need to worry too much about fatal accidents during the landing phase.

A next step could be to improve the model by adding structured data which is also available, maybe about time or weather conditions. All of this takes you minutes with JMP, rather than days of programming and tedious text processing. Interactive visual outputs are made directly available to communicate your findings.

Lasso with AICc Validation

Solution Path



Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%	Singularity Details
Intercept	-2.507953	0.134383	348.29741	<.0001*	-2.771339	-2.244567	
landing Binary	-1.181923	0.1660677	50.653357	<.0001*	-1.50741	-0.856437	
failure Binary	0	0	0	1.0000	0	0	
factor Binary	-0.093249	0.1279109	0.5314684	0.4660	-0.34395	0.1574514	
resulted Binary	0	0	0	1.0000	0	0	
failure to maintain Binary	0.3893494	0.1764678	4.8679732	0.0274*	0.0434789	0.7352198	
flight Binary	0.3782425	0.1545093	5.9928205	0.0144*	0.0754098	0.6810751	
factors Binary	0	0	0	1.0000	0	0	
contributing Binary	0	0	0	1.0000	0	0	
terrain Binary	0.0162515	0.1870853	0.0075458	0.9308	-0.350429	0.382932	

Fig. 7: Using text analysis results in predictive modelling (here: Generalized Regression)

CONCLUSION

Making discoveries and visualizing information from unstructured text data has never been easier and more fun. Using version 13 of JMP (and of JMP Pro for even more powerful text analysis) allows you to add unstructured text data to your data mining scenario. Combining the results from your text analysis with other structured data can take your predictive models to the next level.

Ideas for future developments include more options to import text data into JMP (see also JMP Addin, 2016), adding other data formats like audio or video and to further streamline the integration of text analysis and predictive modeling tools.

REFERENCES

JMP Addin - Adsurgo TM Tools (2016): *Text Importer - Text, PDF, Word Documents, and Powerpoint*. JMP Community. http://bit.ly/2srow3s

JMP Help Text Explorer (2017): Text Explorer - Explore Unstructured Text in Your Data. http://bit.ly/2rh2vok

JMP Home (2017): http://www.JMP.com

Karl A., Wisnowski J., Rushing W.H. (2015). *A practical guide to text mining with topic extraction. WIREs Comput Stat* 2015, 7:326–340. doi: 10.1002/wics.1361

Klimberg, Ron and B.D. McCullough (2016). *Fundamentals of Predictive Analytics with JMP, Second Edition*. Cary, NC: SAS Institute Inc. Chapter 15: Text Mining.

Singh V. and Zhang Q. (2016). *Text Mining Online Reviews with JMP 13*. JMP Discovery Summit, Cary, NC: http://bit.ly/2t5CWTs

Vijayan, J. (2015). *Solving the Unstructured Data Challenge*. CIO. http://www.cio.com/article/2941015/big-data/solving-the-unstructured-data-challenge.html