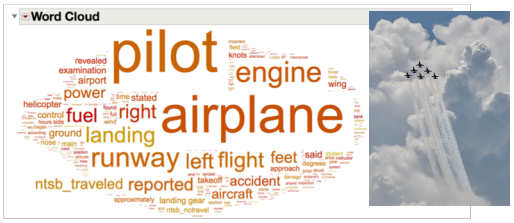
CHINA DISCOVERY SUMMIT
EXPLORING DATA INSPIRING INNOVATION

Flying Through the Word Cloud into Modeling with JMP Pro Text Explorer

Laura A. Higgins, Ph.D.
JMP Global Technical Enablement Engineer

Abstract

The world is full of unstructured text, and most of it goes unexplored. Further complicating this, extracting meaning and value from text until recently required special tools. JMP Pro's Text Explorer has many powerful options for bringing insight into text data and using the curated term list. Let's fly beyond the word cloud and explore the modeling tools of Text Explorer.



What else can you do?

Once you have curated text, don't stop at the word cloud – try one of these other techniques to uncover meaning hidden in text data.

•Are there underlying themes in the data? Find which documents (rows) cluster together.

•What conceptual topics and themes occur across documents with terms found together?

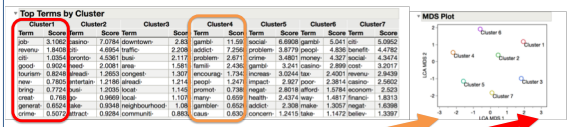
•Create stable variables from text analysis to use in modeling. Understand if topics contribute in a positive or negative way to an outcome variable.

•Create output for market basket/association analysis.

Which documents cluster together: Latent Class Analysis

Are there underlying themes in the data?
Find which documents (rows) cluster together.

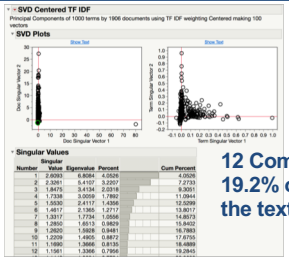
How to do this:
How many clusters? Iterative process
•Find meaningful clusters with Top Terms and MDS Plot
•Compare BIC – lower is better
•Check out Mixture Probability and read the text



Latent Semantic Analysis: Step 1 for Further Analysis

Create the document term matrix (SVD).
First step for many other analyses.

How to do this:
Latent Semantic Analysis is Principal Components
•Use SVD Plots and read text for interpretation
•Singular Values show how much signal = Cumulative Percent of Variation
•Use one of the following techniques for further analysis

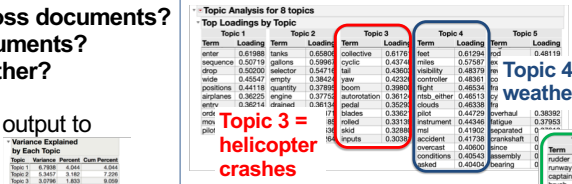


12 Components explain 19.2% of the variation in the text data

Which terms occur together: Topic Analysis

What conceptual topics and themes occur across documents?
What terms occur together across documents?
Also what terms are *not* found together?

How to do this:
•Explore different number of topics – use Variance output to quantify how much signal.
•Each topic shows how terms load
•Negative terms are not found with positive terms
•Read text to understand. Try using in modeling for further interpretation with other variables.



Use Topic Analysis Output in Models

Create stable variables from text analysis to use in modeling.
Understand if topics contribute in a positive or negative way to an outcome variable.

How to do this:
•Use Document Topic Vectors in GenReg with an outcome variable
•Topics 4, 1, 5 contribute to Fatal Accidents
•All other topics contribute to non-fatal accidents
•Topics 4, 6 are the most important for understanding fatal vs non-fatal accidents