Harness the Power of Text Mining: Using FDA Data to Analyze Medical Device Recalls

W. Heath Rushing
Adsurgo LLC

ADSURGO
ascend with analytics

W. Heath Rushing

Principal Consultant, Adsurgo LLC

Heath Rushing is the cofounder of **Adsurgo** and author of the book *Design and Analysis of Experiments by Douglas Montgomery: A Supplement for using JMP*. Previously, he was the JMP and Six Sigma training manager at SAS. He led a team of nine technical professionals designing and delivering applied statistics and quality continuing education courses. He created tailored courses, applications, and long-term training plans in quality and statistics across a variety of industries to include biotech, pharmaceutical, medical device, and chemical processing. Mr. Rushing has been an invited speaker on applicability of statistics for national and international conferences. As a Quality Engineer at Amgen, he championed statistical principles in every business unit. He designed and delivered a DOE course that immediately became the company standard required at multiple sites. Additionally, he developed and implemented numerous innovative statistical methods advancing corporate risk management, process capability, and validation acceptance criteria. He won the top teaching award out of 54 instructors in the Air Force Academy math department where he taught several semesters and sections of operations research and statistics. Additionally, he designs and delivers short courses in statistics, data mining, and simulation modeling for SAS.

Objectives

A student who successfully completes this course will:

- Know what text mining and natural language processing is and how the differ from data mining and predictive analytics.
- Understand the role of enabling technologies in the evolution of text mining methodologies.
- Know the different areas of text analytics.
- Be able to apply data mining techniques such as decision trees, cluster analysis, and logistic regression to translate intermediate text mining data to decision quality results.
- Understand text mining is a subset of natural language processing.
- Understand the role of latent semantic analysis using singular value decomposition.
- Understand how to use a standard text mining software product.

Outline

- Demonstration: FDA Recall Data
- Introduction to Text Mining
- Demonstration: Inspection Observations
- Text Mining
  - String Processing
  - Natural Language Processing
  - Statistical Approaches
  - Clustering
- Application Examples
- Appendix
  - References

Text Mining Example – FDA Recall Data

- Data: Medical Device Recall Data from fda.gov
- Objective: Use text mining to summarize issues in medical device recalls.
- Software used: SAS/JMP script with R.

INTRODUCTION

What is Text Mining?

- Text mining relies on several disciplines
  - Statistics
  - AI and Machine Learning
    - Data Mining
  - Database Management
  - Library and Information Sciences
  - Computational Linguistics
  - Information Technology

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Extracting Numerical Representations of Text

- In order to analyze text in a systematic and structured way, we first need to develop a numerical representation of the text.
- Obviously, there is not a unique solution to this problem. The appropriate mapping of text->numbers depends on the goal of the study.

How is Text Mining Similar to Data Mining

- Once we have extracted numerical summaries of the documents, we will rely on existing statistical and machine learning methods to process the information.
- However, these summaries have special properties that need to be recognized. As such, text mining is a deeper topic than just finding a mapping of text to numbers.

How is Text Mining Different from Data Mining

- Focusing on syntax instead of semantics, the standard approach for representing a collection of documents is with a matrix with documents as rows and terms as columns.
  - As an analogy to other data mining problems, think of documents as observations and terms as variables.
  - This matrix will typically be extremely large, yet sparse.
  - The nonzero entries of the matrix are all positive.
  - Overlaps of nonzero components in two documents are more informative than overlaps of zero components.
  - There will not be any missing data.
  - Long documents will have larger term counts than short documents.
  - Different transformations of this matrix are used in practice.

Evolution of Text Mining

- Early motivation: cataloging library books and articles.
  - Dewy Decimal System (1876)
  - Summarizing scientific documents with abstracts (1898)
  - Computer-generated abstracts (1958)
  - Discussion of classifying library books by word frequencies (1961)

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Development of Enabling Technology

- Graphical capabilities.
  - Software is now widely available for plotting labeled points while minimizing overlap in the labels in 2- or 3-D.

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Development of Enabling Technology

- Reduction of Dimensionality and Feature Selection.
  - Processing text is a high-dimensional problem, even in applications where grammatical structure is ignored.
  - Computational developments have improved the scalability of text mining applications.
    - Sparse matrix methods have improved memory requirements and sped up computations.
    - The singular value decomposition has been extremely useful for reducing the dimensionality of problems.

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Development of Enabling Technology

- Statistical Approaches
  - Document clustering via hierarchical and k-means algorithms.
  - Document classification via:
    - Linear models
    - CART

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Seven Practice Areas of Text Mining

1. Search and information retrieval (IR)
2. Document clustering
3. Document classification
4. Web mining
5. Information extraction (IE)
6. Natural language processing
7. Concept extraction

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Focus of Our Course

- Information retrieval
- Concept extraction
- Document clustering
- Document classification
- Sentiment analysis
- Some tools from natural language processing

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Problems TM Can Address

- Predicting the probability an insurance claim is fraudulent based on the text in the claim
- Filtering spam
- Producing a list of documents (e.g. emails, error reports) that are most similar to one of interest. Might be useful in an audit, or if you work for the NSA
- Obtaining a fast, yet representative, summary of the topics in a collection of documents
- Finding which types of aviation accidents are most strongly associated with the presence of fatalities
- Putting out-of-spec incident reports from process engineers to use (instead of not using them at all)
- Evaluating customer sentiment about new product releases (Twitter, focus groups, complaints, etc.)
- Determining authorship
- Sentiment analysis
- Trend analysis: what are the most common themes in the abstracts at a major statistics conference this year? What were they in 2003?

Text Mining Example – Inspection Observations

- Data: Inspection observations from fda.gov.
- Objective: Determine the most frequent themes in inspection observations for a particular industry (medical device, drugs, biologics).
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

DSURGO
ascend with analytics

TEXT MINING

STRING PROCESSING

DSURGO
ascend with analytics

What can text mining do for you?

- Data mining techniques require numbers. They don't know what to do with text.
  - Some categorical factors will be expressed using text, .e.g., "pass/fail". However, this is not text mining.
  - What if we could also express free-form text with numbers?

DSURGO
ascend with analytics

Possibilities

- If we could represent text with numerical indices, we could use those indices as input to
  - Supervised learning methods (target variable)
    - Linear and logistic regression
    - CART
  - Unsupervised methods (no target variable)
    - Hierarchical Clustering
    - K-means clustering

NTSB Example

- In this section, we will occasionally use data from a collection of National Transportation Safety Board aviation accident reports to illustrate a concept
- The documents in this corpus consists of short descriptions of the cause of each accident.
- The data are available from Weiss, S., et al. (2009) *Text Mining: Predictive Methods for Analyzing Unstructured Information*

Text Mining Example – NTSB Aircraft Accident Reports

- Data: NTSB Aircraft Accident Reports.
- Objective: Determine what factors contributed to fatal accidents.
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

String Processing

Car Accidents
Slid on ice into a curb.
Driving too fast in a dust storm, hit the curb.
Low-budget tires failed after bumping curb.

- We will use the three car accident descriptions above to illustrate text processing.

Bag of Words Approach

- Using a “bag of words” approach, we disregard the ordering of the words in each document as well as their grammatical properties.
- While this may seem simplistic, it has been shown to give excellent results in many applications.

Vocabulary

- Document: a string of words.
- Corpus: a collection of documents.
- In the text mining literature, “words,” “terms,” and “tokens” all describe roughly the same idea. There are some subtleties to their use: we will use them interchangeably to mean words that have been extracted from a document and processed.

Processing Text

- Within each document, we will first
  - Isolate individual words
  - Remove punctuation
  - Normalize case (convert all characters to lowercase)
  - Remove numbers
- Later, we will discuss further processing of the words.

Isolate Words

Document 1	Document 2	Document 3
Slid	Driving	Low-budget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb.	dust	curb.
	storm,	
	hit	
	the	
	curb.	

- Notice that punctuation is concatenated to adjacent terms.

Remove Punctuation

Document 1	Document 2	Document 3
Slid	Driving	Lowbudget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb	dust	curb
	storm	
	hit	
	the	
	curb	

DSURGO
ascend with analytics

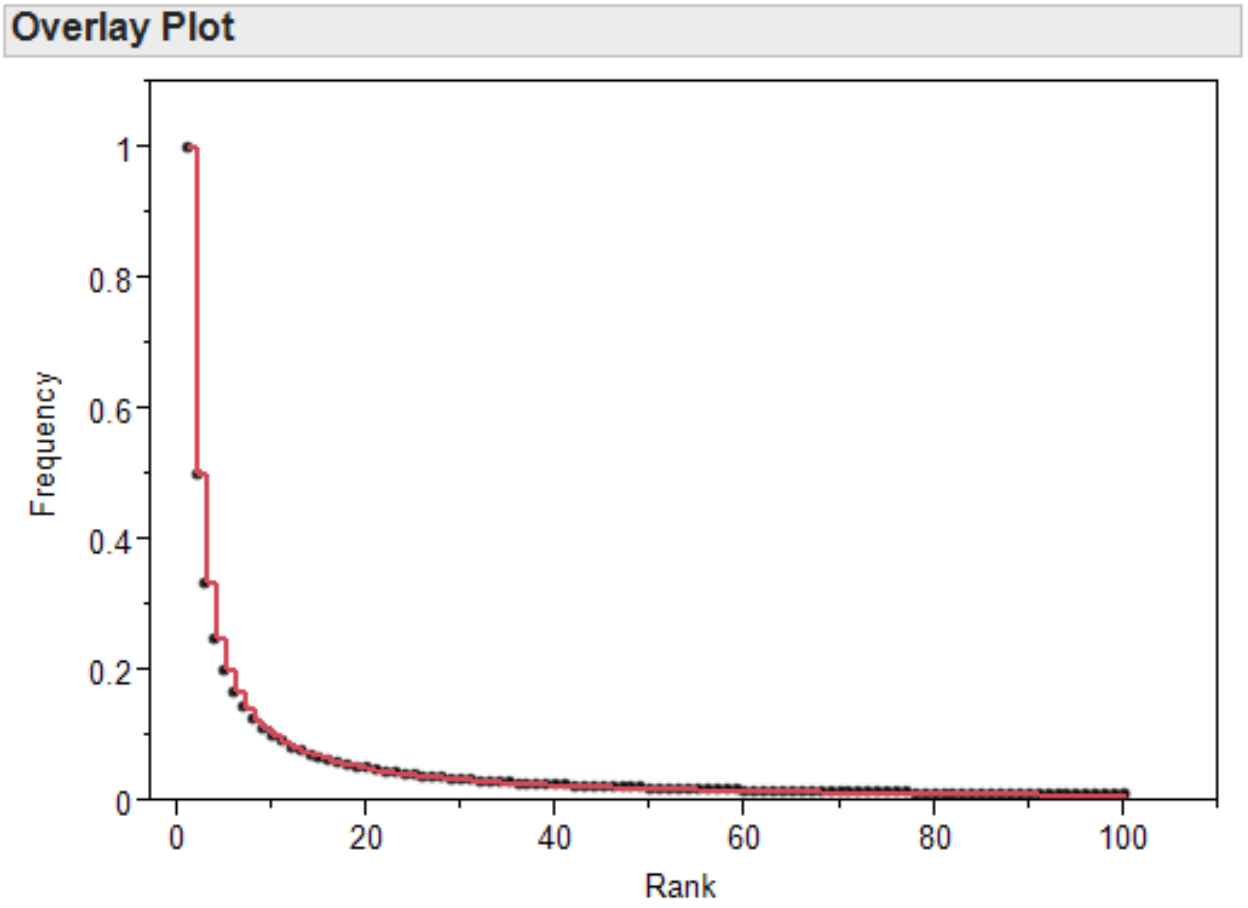
Normalize Case

Document 1	Document 2	Document 3
slid	driving	lowbudget
on	too	tire
ice	fast	failed
into	in	after
a	a	bumping
curb	dust	curb
	storm	
	hit	
	the	
	curb	

NATURAL LANGUAGE PROCESSING

Zipf's Law and Term Frequency Counts

- When counting frequency of terms in a corpus, the frequency of a word will be roughly proportional to its rank.



Practical Implications of Zipf's Law

- The matrix that contains the documents (in rows) and terms (in columns) is called the document-term matrix (DTM).
- The DTM is sparse: with the majority of terms only appearing a few times, most entries in the DTM will be 0.
- A small collection of words occur so frequently (and likely in so many documents) that they do not give any discriminating power. These are referred to as stopwords.

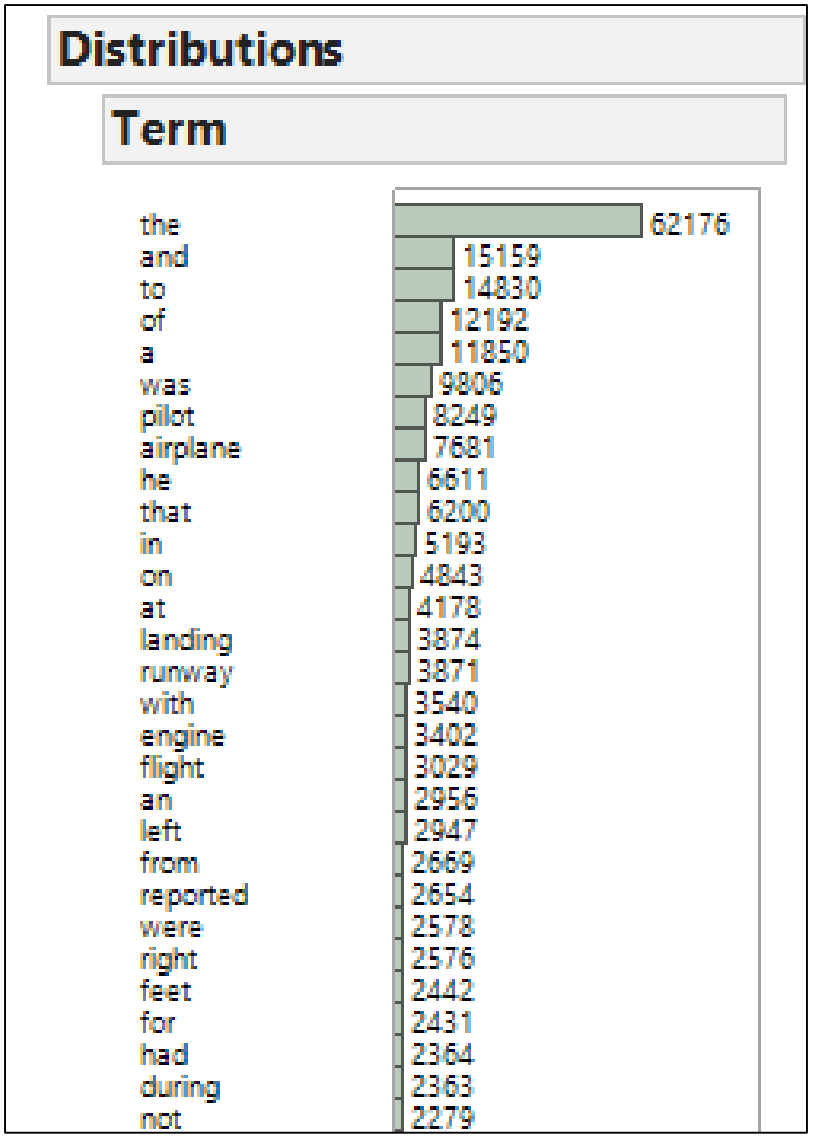
Finding the Important Terms

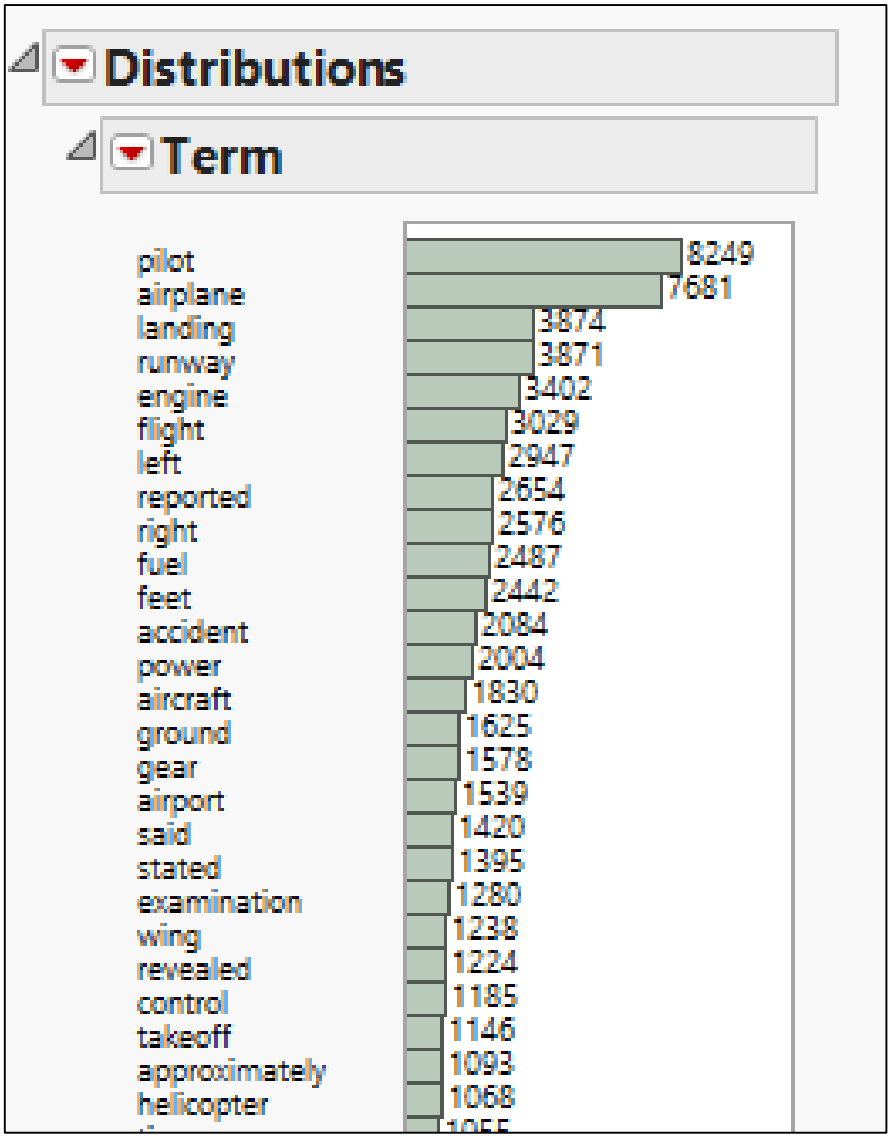
- Stopwords occur so frequently that they are uninformative.
- A large proportion of infrequent words will either be typos or appear in so few documents that they are not useful in detecting patterns.
- Typically, there will be a handful of medium frequency words that provide the most flexibility in differentiating between the document themes.

Natural Language Processing

- After extracting the tokens from a document, it is typically useful to
  - Remove stopwords (most frequent words).
  - Stem the text.
  - Remove words with character length below a minimum or above a maximum.
  - Remove words that appear in only a few documents (most infrequent words).

Stopwords


Distributions
Term
the 62176
and 15159
to 14830
of 12192
a 11850
was 9806
pilot 8249
airplane 7681
he 6611
that 6200
in 5193
on 4843
at 4178
landing 3874
runway 3871
with 3540
engine 3402
flight 3029
an 2956
left 2947
from 2669
reported 2654
were 2578
right 2576
feet 2442
for 2431
had 2364
during 2363
not 2279

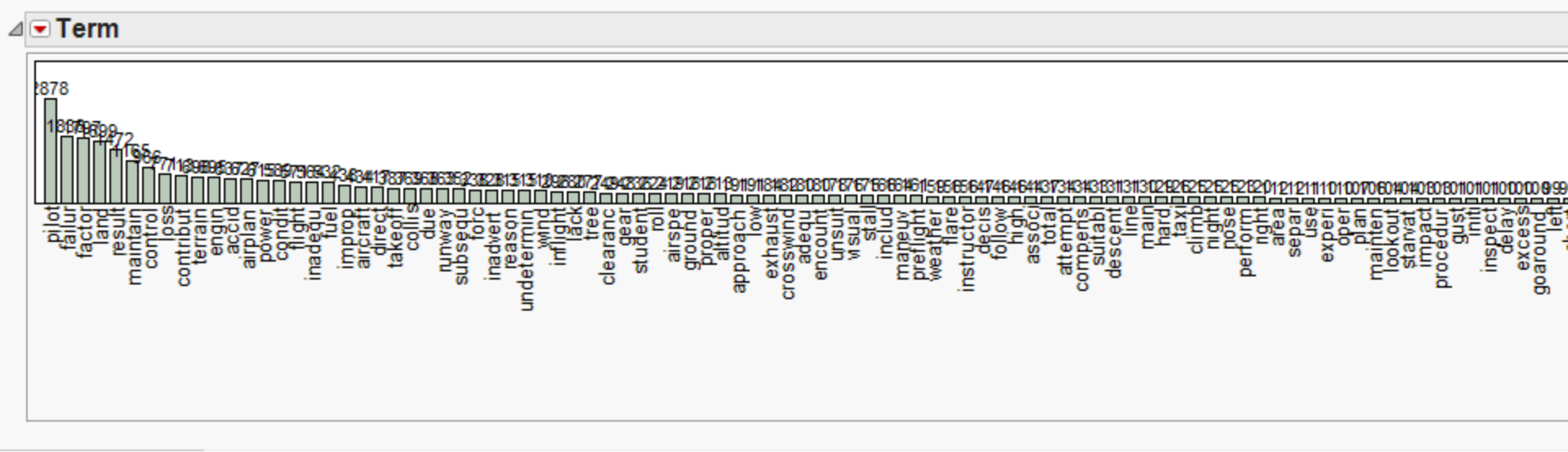

Distributions
Term
pilot 8249
airplane 7681
landing 3874
runway 3871
engine 3402
flight 3029
left 2947
reported 2654
right 2576
fuel 2487
feet 2442
accident 2084
power 2004
aircraft 1830
ground 1625
gear 1578
airport 1539
said 1420
stated 1395
examination 1280
wing 1238
revealed 1224
control 1185
takeoff 1146
approximately 1093
helicopter 1068

Stopwords

A partial list of common stopwords appears below. All text mining software packages will have a list of common English stopwords that may be used.

"i"	"me"	"my"	"myself"	"we"
"our"	"ours"	"ourselves"	"you"	"your"
"yours"	"yourself"	"yourselves"	"he"	"him"
"his"	"himself"	"she"	"her"	"hers"
"herself"	"it"	"its"	"itself"	"they"
"them"	"their"	"theirs"	"themselves"	"what"
"which"	"who"	"whom"	"this"	"that"
"these"	"those"	"am"	"is"	"are"
"was"	"were"	"be"	"been"	"being"
"have"	"has"	"had"	"having"	"do"
"does"	"did"	"doing"	"would"	"should"
"could"	"ought"	"i'm"	"you're"	"he's"
"she's"	"it's"	"we're"	"they're"	"i've"
"you've"	"we've"	"they've"	"i'd"	"you'd"
"he'd"	"she'd"	"we'd"	"they'd"	"i'll"
"you'll"	"he'll"	"she'll"	"we'll"	"they'll"
"isn't"	"aren't"	"wasn't"	"weren't"	"hasn't"
"haven't"	"hadn't"	"doesn't"	"don't"	"didn't"
"won't"	"wouldn't"	"shan't"	"shouldn't"	"can't"
"cannot"	"couldn't"	"mustn't"	"let's"	"that's"
"who's"	"what's"	"here's"	"there's"	"when's"
"where's"	"why's"	"how's"	"a"	"an"
"the"	"and"	"but"	"if"	"or"
"because"	"as"	"until"	"while"	"of"
"at"	"by"	"for"	"with"	"about"
"against"	"between"	"into"	"through"	"during"
"before"	"after"	"above"	"below"	"to"

Term Frequency



- The drop-off in term frequencies will not be as pronounced as the one predicted by Zipf's law if stopwords have been removed, as they have been in this example.

Wordcloud

A wordcloud displays the most frequent terms in a corpus in the center of the cloud. Terms get smaller and move away from the center (and are color-coded) as they become less frequent. The orientation of the term is irrelevant in this example.

Custom Stopwords

- In each application, there will likely be a set of frequent words that appear in most documents.
- For example, in a collection of reports about airplane accidents, the words “airplane”, “plane”, and “aircraft” may be specified as custom stopwords.
- In other cases, the generic stopword list may remove too many terms.

Custom Stopwords

- Optionally, a column of custom stopwords may be given to the script.
- These do not need to be individual words: a phrase of words separated by spaces may be entered into a single cell of the custom stopwords column
- The script searches for and removes words from this list at three separate points during the execution of the script
  - Before the documents have been processed
  - After the punctuation has been removed and the text converted to lower case
  - After the terms have been stemmed

Remove Stopwords

Document 1	Document 2	Document 3
slid	driving	lowbudget
ice	fast	tire
curb	dust	failed
	storm	bumping
	hit	curb
	curb	

Stemming Text

- “The computer is saving the table.”
- “The computers saved the tables.”
- After removing stopwords, these documents are reduced to
  - computer saving table
  - computers saved tables
- Even though these documents are conveying the same idea, they have no words in common and would not be recognized as similar.

Stemming Text

- Stemming effectively chops the endings off of words. Plural and singular nouns are reduced to the same token, and conjugated verbs are also reduced.
- After stemming, the sentences from the last slide become:
  - comput save tabl
  - comput save tabl
- The documents are now recognized as identical.
- As a further example, the next slide will contain the stemmed text from this slide (with stopwords removed).

Stemming Text

- stem effect chop end word plural singular noun reduc token conjug verb also reduc
- stem sentenc last slide becom
  - comput save tabl
  - comput save tabl
- document now recogn ident
- exampl next slide will contain stem text slide stopword remov

Stem Text

Document 1	Document 2	Document 3
slid	drive	lowbudget
ice	fast	tire
curb	dust	fail
	storm	bump
	hit	curb
	curb	

Other Processing Options

- Remove numbers: may not want to if part numbers or important identifiers are present.
- Filtering out words by character length
  - With the exception of chemical names, the longest non-coined, non-technical English word is 28 letters (Wikipedia).
  - Anything longer is likely to be a web URL or a string of garbled text (common when converting PDF to text).
- Removing words that appear in only a few documents.
  - Words that appear in only 1 document give no discriminating power. Removing them reduces computational requirements. May also want to remove words that appear in only a handful of documents.

Representing Text with Numbers

- To find clusters of documents or to use the information present in the documents in a predictive model, we need a numerical representation of the text.
- Using the bag of words approach, we create a document term matrix (DTM). Each document is represented by a row, and each token is represented by a column. The components of the matrix represent how many times each token appears in each document.

Document Term Matrix

Doc	bump	curb	drive	dust	fail	fast	hit	ice	lowbudget	slid	storm	tire
1	0	1	0	0	0	0	0	1	0	1	0	0
2	0	1	1	1	0	1	1	0	0	0	1	0
3	1	1	0	0	1	0	0	0	1	0	0	1

DSURGO
ascend with analytics

Properties of the DTM

- The DTM will typically be very sparse (most entries are 0).
- Even for modestly sized applications, the full DTM will be too large to hold in memory.
- Since most entries are 0, multiplying the matrix results in several multiplications by 0, which could be omitted.
- Special software and algorithms are available for storing and manipulating sparse matrices.

Transformations of the DTM

- Various transformations of the term-frequency counts in the DTM have been found to be useful.

Transformations of the DTM

- Frequency (local) weights
  - Binary: Useful if there is a lot of variance in the lengths of the documents in the corpus.
  - Ternary/Frequency: Some researchers have found that distinguishing between terms that appear only once in a document vs. those that appear multiple time can improve results.
  - Log: Dampens the presence of high counts in longer documents without sacrificing as much information as the binary weighting scheme.

Transformations of the DTM

- Term (global) weights
  - Term Frequency - Inverse Document Frequency (tf-idf)
    - Shrinks the weight of terms that appear in many documents while also inflating the weight of terms that appear in only a few documents
    - Sometimes makes interpretation of results more difficult, but can give better predictive performance. In practice, it is best to try different weighting schemes: there is no need to pick only one!

Transformations of the DTM

- Normalizing each document
  - The term frequency weights in each document may be normalized so that the sum of each document vector is 1. This is done by dividing the term counts in each document (each row of the DTM) by the total number of words in each document (the row sums of the DTM)
  - This can be useful when the documents are of different lengths. An illustration of how this can help: if a document D' is created by pasting two copies of a document D together, D and D' will be identical after normalization.

Normalized Term-Frequency Document Term Matrix

Doc	bump	curb	drive	dust	fail	fast	hit	ice	lowb udget	slid	storm	tire
1	0	0.333	0	0	0	0	0	0.333	0	0.333	0	0
2	0	0.167	0.167	0.167	0	0.167	0.167	0	0	0	0.167	0
3	0.2	0.2	0	0	0.2	0	0	0	0.2	0	0	0.2

Term Frequency-Inverse Document Frequency

- Terms that appear in most documents give little discriminating power (hence the removal of stopwords).
- To address this, we multiply the term frequencies by an inverse document frequency, which down-weights words that appear in many documents.

Inverse Document Frequency

- idf down-weights terms that appear in many documents. The idf for term t is

$$idf_t = \log_2 \left(\frac{D}{df_t} \right)$$

- D is the number of documents in the corpus.
- df_t is the number of documents containing term t.
- If a term appears in every document, its idf is 0.

Term Frequency – Inverse Document Frequency

- After calculating the inverse document frequencies (one for each term), we multiply the term frequency entries of the DTM by the appropriate idf.
- For example, the tf entry for *curb* in Document 1 is 1, since *curb* appears once in Document 1. However, the inverse document frequency of *curb* is $log\left(\frac{3}{3}\right) = 0$. Thus the tf-idf entry for *curb* in Document 1 is 0.

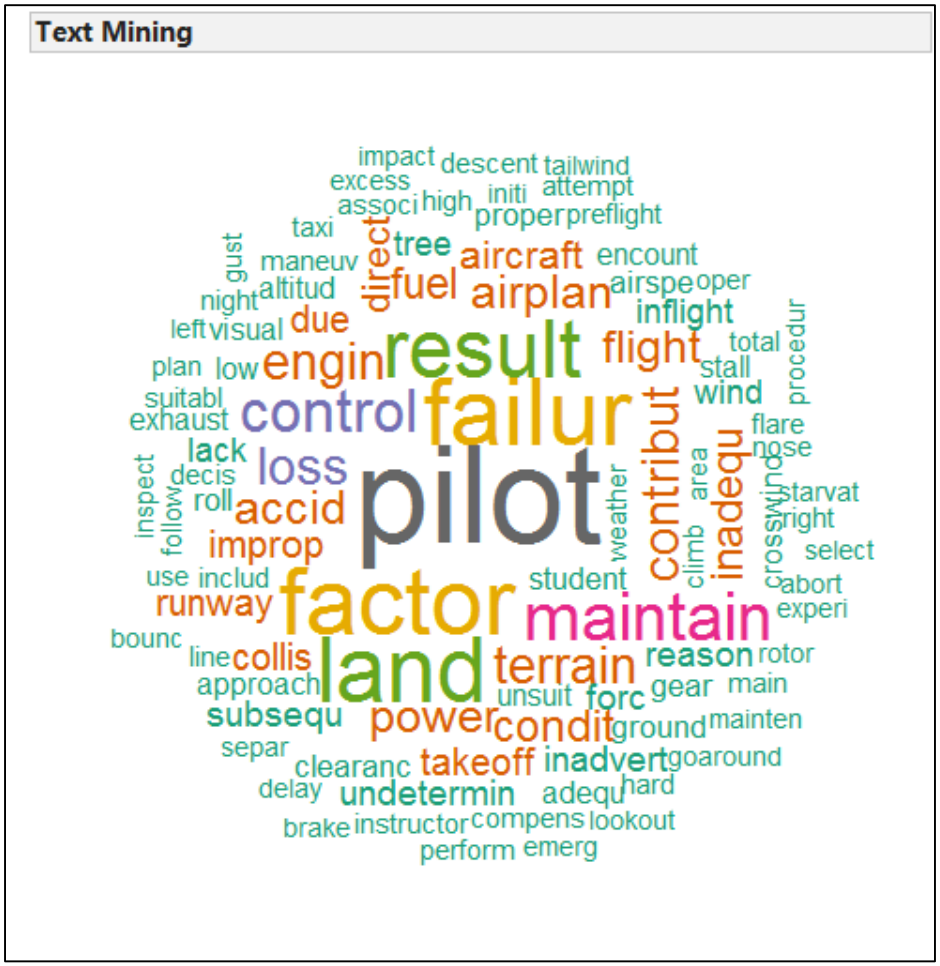
tf-idf

Doc	bump	curb	drive	dust	fail	fast	hit	ice	lowbudget	slid	storm	tire
1	0	0	0	0	0	0	0	1.585	0	1.585	0	0
2	0	0	1.585	1.585	0	1.585	1.585	0	0	0	1.585	0
3	1.585	0	0	0	1.585	0	0	0	1.585	0	0	1.585

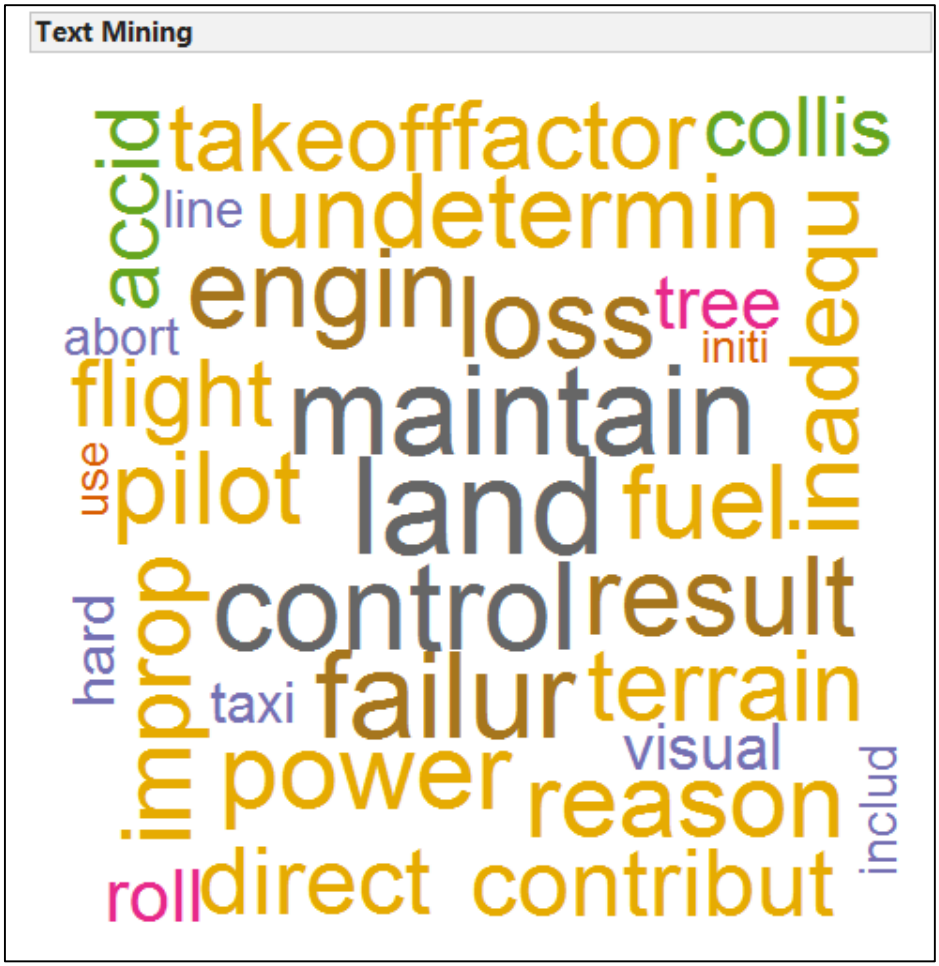
Wordcloud

A wordcloud displays the most frequent terms in a corpus in the center of the cloud. Terms get smaller and move away from the center (and are color-coded) as they become less frequent. The orientation of the term is irrelevant in this example.

Wordcloud depends on Frequency Weighting



Term Frequency



tf-idf

Frequency Weighting Summary

- There is no universally best weighting: take time to try different options.
- The following slides provide an additional example of how the different weighting schemes compare to the raw frequency counts.

Binary

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	1	0	0	1	1
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	1	0	0	1	1
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	1	0	0	1	1
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

Ternary

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	2	2
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	2	2
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	2	2
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

Log

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	1.5849625007	2
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	1.5849625007	2
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	2	0	0	1.5849625007	2
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

tf(normalized)-idf

	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
2	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
3	one, eleven	0	1	0	0	1	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
5	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
6	one, eleven	0	1	0	0	1	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	3	0	0	2	3
8	one, two, three, four, five, six, seven, eight	1	0	1	1	1	1	1	1	1
9	one, eleven	0	1	0	0	1	0	0	0	0

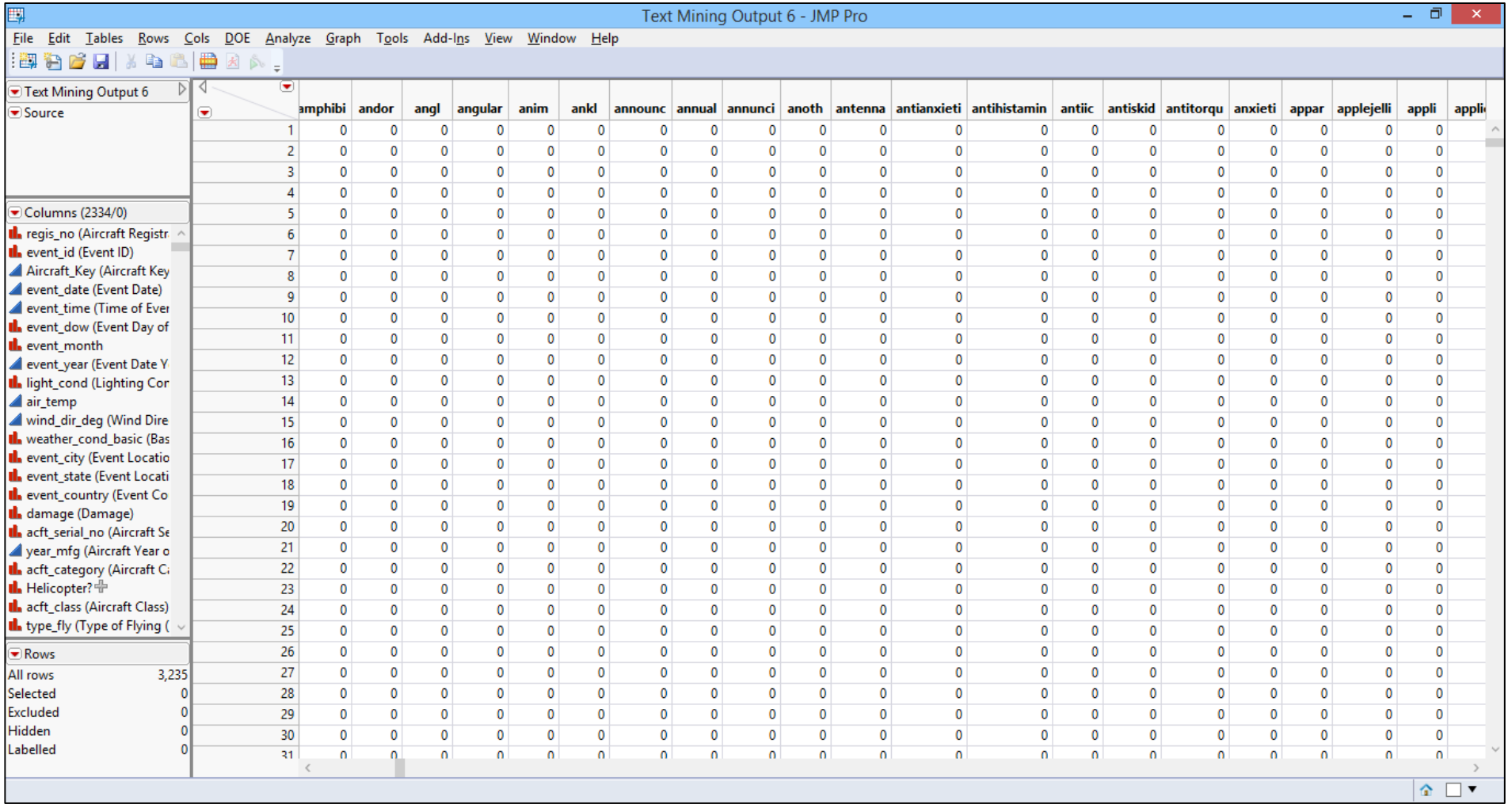
	Column 1	eight	eleven	five	four	one	seven	six	three	two
1	one, two, three, one, two, three, one, two	0	0	0	0	0	0	0	0.1462406252	0.2193609378
2	one, two, three, four, five, six, seven, eight	0.1981203126	0	0.1981203126	0.1981203126	0	0.1981203126	0.1981203126	0.0731203126	0.0731203126
3	one, eleven	0	0.7924812504	0	0	0	0	0	0	0
4	one, two, three, one, two, three, one, two	0	0	0	0	0	0	0	0.1462406252	0.2193609378
5	one, two, three, four, five, six, seven, eight	0.1981203126	0	0.1981203126	0.1981203126	0	0.1981203126	0.1981203126	0.0731203126	0.0731203126
6	one, eleven	0	0.7924812504	0	0	0	0	0	0	0
7	one, two, three, one, two, three, one, two	0	0	0	0	0	0	0	0.1462406252	0.2193609378
8	one, two, three, four, five, six, seven, eight	0.1981203126	0	0.1981203126	0.1981203126	0	0.1981203126	0.1981203126	0.0731203126	0.0731203126
9	one, eleven	0	0.7924812504	0	0	0	0	0	0	0

STATISTICAL APPROACHES

Singular Value Decomposition

- The DTM is usually very large, though sparse.
- Working directly with the DTM requires software capable of performing sparse matrix algebra.
- Even then, most of the terms represent noise variables. This presents a complication for regression methods.

Full DTM is Sparse

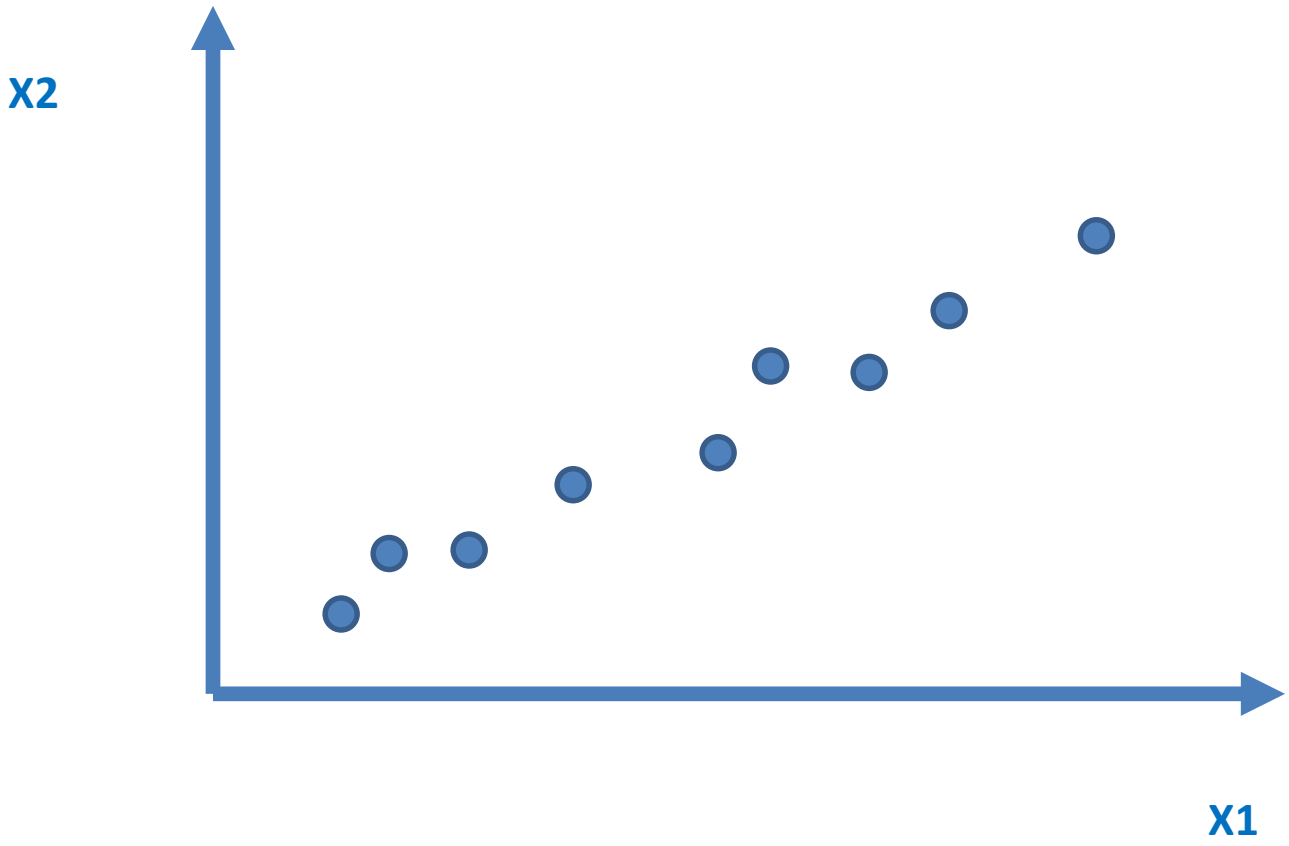


Singular Value Decomposition

- The reduced-rank singular value decomposition (SVD) provides us with a dimensionality reduction technique.
- The SVD reduces the DTM to a (dense) matrix with fewer columns. The new (orthogonal) columns are linear combinations of the rows in the original DTM, selected to preserve as much of the structure of the original DTM as possible.

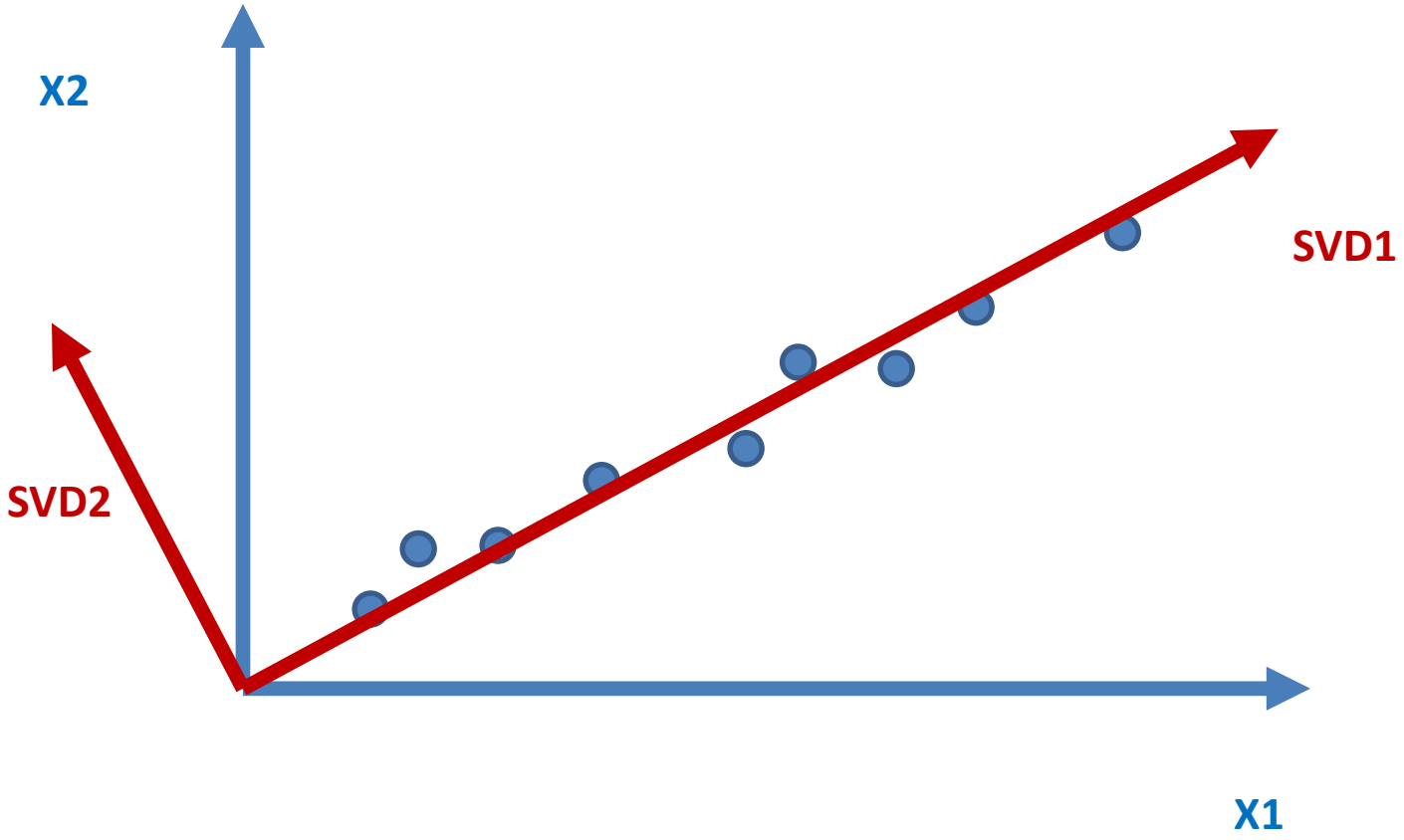
SVD Example

X1 and X2 describe the location of these points.
However, they appear to fall mostly along a line.



SVD Example

Roughly, the SVD finds a new set of orthogonal basis vectors such that each additional dimension accounts for as much of the variation of the data as possible.



SVD Example

SVD1

- This lower-dimensional representation of the data preserves as much of the original structure as possible.
- In addition to memory and computational benefit of storing and manipulating fewer variables, predictive models will be working with fewer noise factors.

Singular Value Decomposition

- For a DTM *X,* the SVD factorization is

$$X \approx UDV^t,$$

where

- *U* is a dense *d* by *s* orthogonal matrix ***U* gives us a new rank-reduced description of *documents***
- *D* is a diagonal matrix with nonnegative entries (the singular values).
- V^t is a dense *s* by *w* orthogonal matrix, where *s* is the rank of the SVD factorization (*s=1,…,min(d,w)*), and the superscript *t* indicates "transpose." ***V* gives us a new rank-reduced description of *terms*.**
- *d* is the number of documents
- *w* is the number of words
- *s* is the rank of the SVD factorization (*s=1,…,min(d,w)*).

Singular Value Decomposition

- The appropriate value of *s* is a matter of debate, and is application dependent. Smaller values of *s* represent a greater extent of dimensionality reduction at the cost of a loss of structure of the original DTM. Values from 30 to 250 are commonly used.

Latent Semantic Analysis

- In natural language processing, the use of a rank-reduced SVD is referred to as latent semantic analysis (LSA).
- A popular LSA technique is to plot the corpus dictionary using the first two vectors resulting from the SVD.
- Similar words (words that either appear frequently in the same documents, or appear frequently with common sets of words throughout the corpus) are plotted together, and a rough interpretation can often be assigned to dimensions appearing in the plot.

Latent Semantic Analysis

- Example: the next example studies online car reviews for Lexus, BMW, Mercedes, and an anonymous brand, carzz.
- The data for this example may be found in: Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Demonstration – Cars

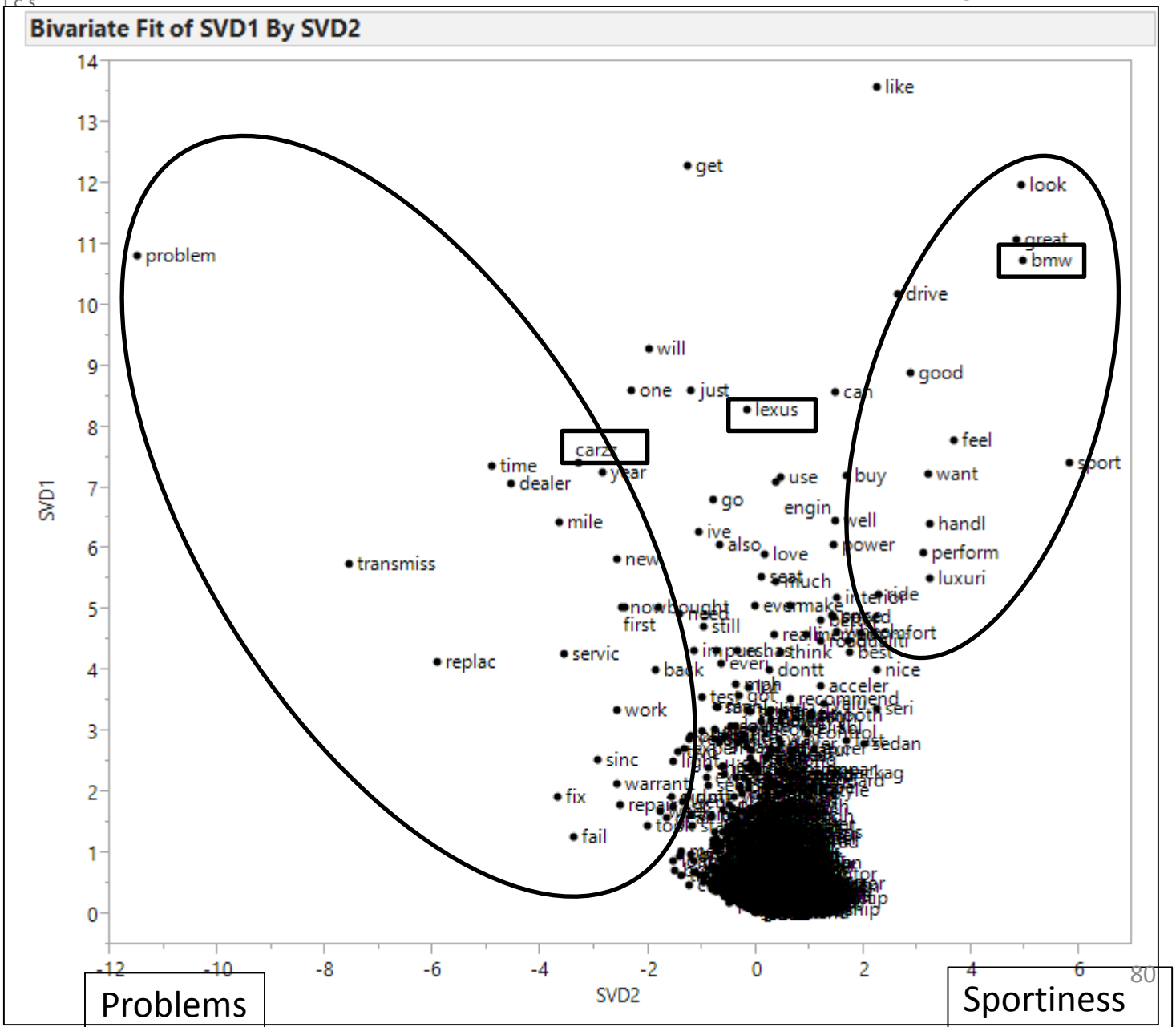
- Data: 638 car reviews written by owners.
- Objective: Determine general sentiment about automobiles. Determine if general sentiments can be 'attached' to certain types of automobiles.
- Software used: R statistical software package

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

Latent Semantic Analysis

General

Specific



Problems

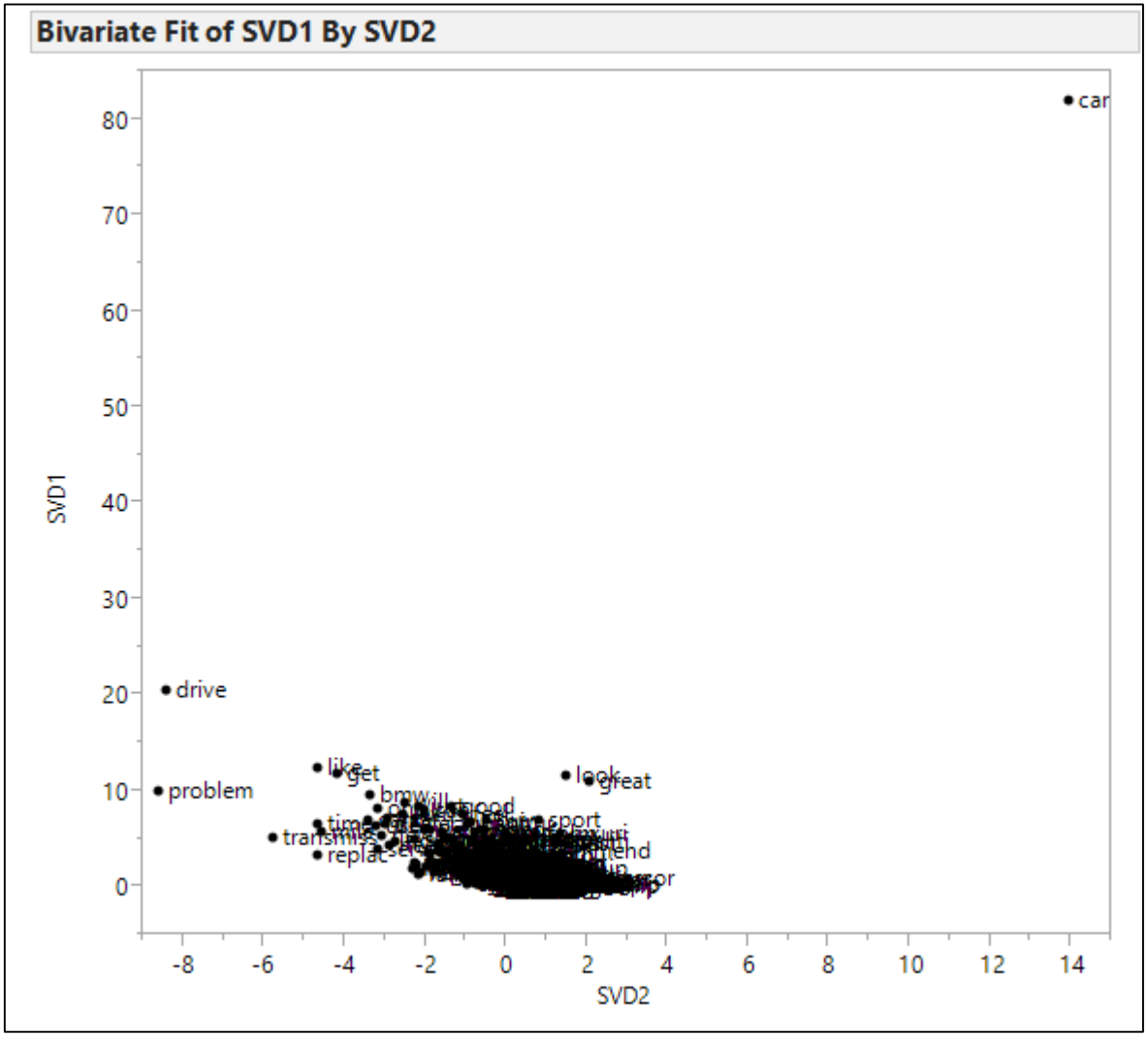
Sportiness

DSURGO
ascend with analytics

Latent Semantic Analysis

- We obtained the previous plot after removing custom stopword: “car”.
- Notice on the next slide how this term dominates the plot (due to its appearance in nearly every document and our use of a term frequency weighting versus tf-idf).
- This is a good illustration of the need for the ability to remove custom stopwords.

Latent Semantic Analysis

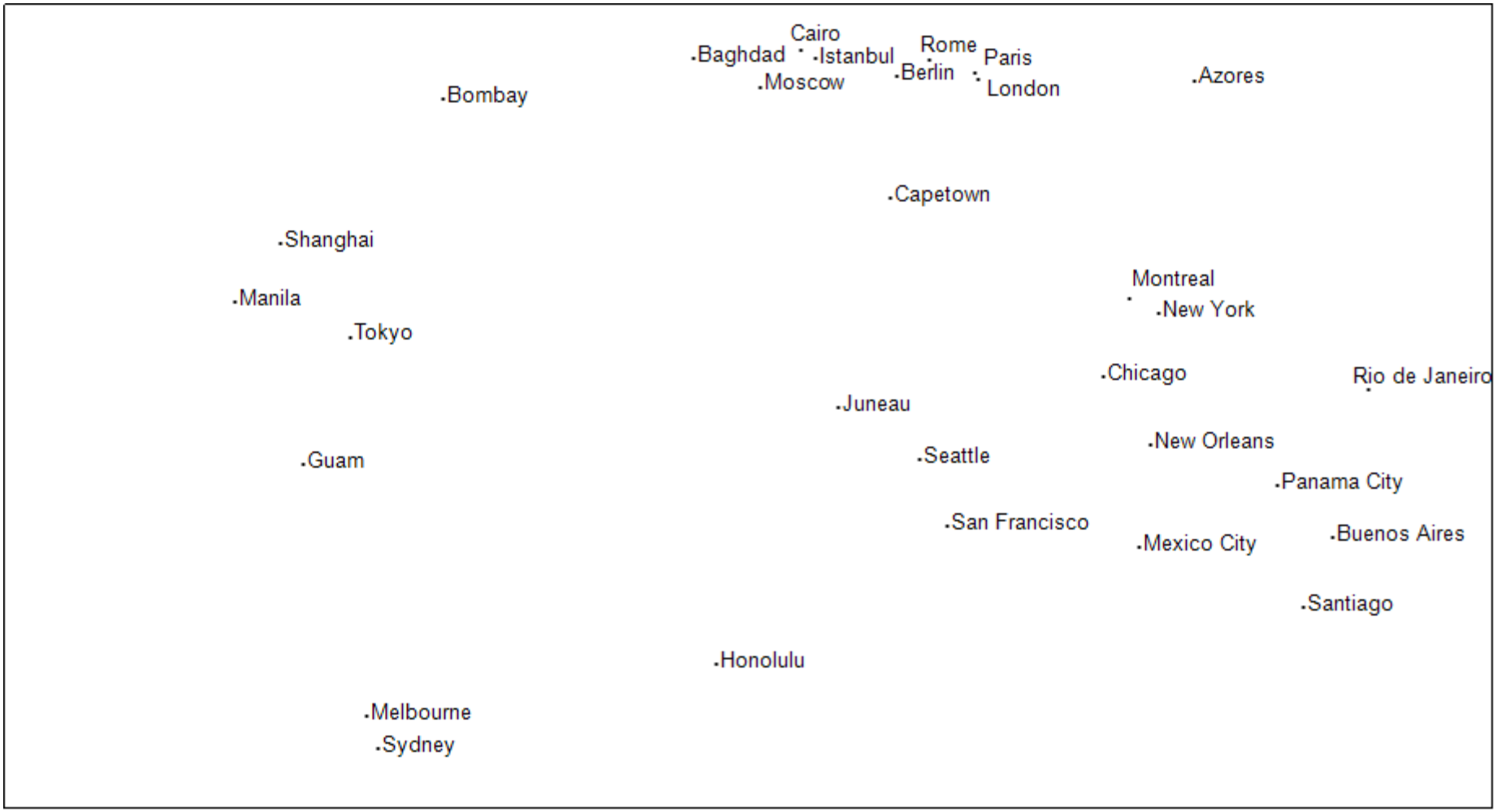


A Related Application

- A related problem (called “multidimensional scaling”) that uses a similar approach: given the pairwise distances between 30 cities, how could you best represent these cities on a piece of paper? This reduction from 3D to 2D will result in some amount of error.
- How long would this take you to construct by hand?

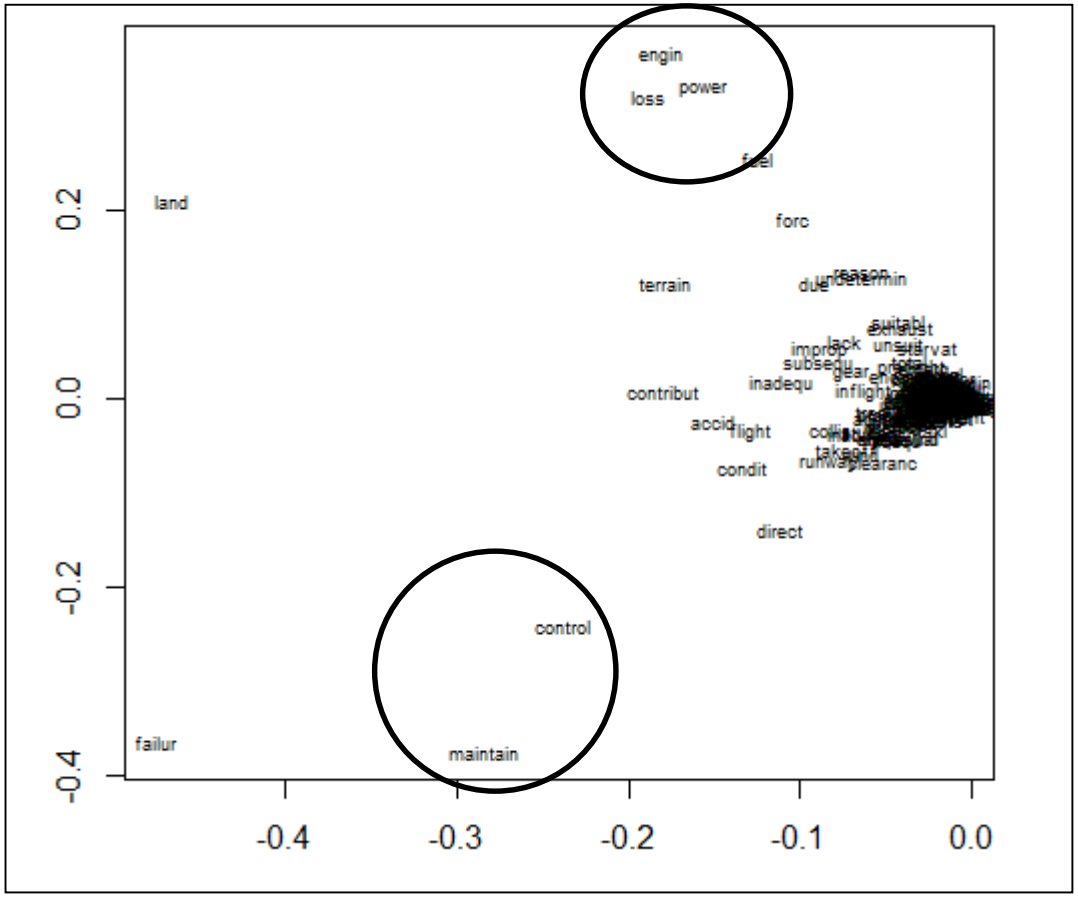
	Azores	Baghdad	Berlin	Bombay	Buenos Aires	Cairo	Capetown
Azores	0	39	22	59	54	33	57
Baghdad	39	0	20	20	81	8	49
Berlin	22	20	0	39	74	18	60
Bombay	59	20	39	0	93	27	51
Buenos Aires	54	81	74	93	0	73	43
Cairo	33	8	18	27	73	0	45
Capetown	57	49	60	51	43	45	0

What's the point?



- The SVD is capable of intelligently scaling down the dimensionality of your data.

SVD1 vs. SVD2



- The words appearing close to each other appear together frequently (or appear independently with a common set of words) in documents in the corpus. We also look for themes describing the spread of terms in this plot (latent semantic analysis).

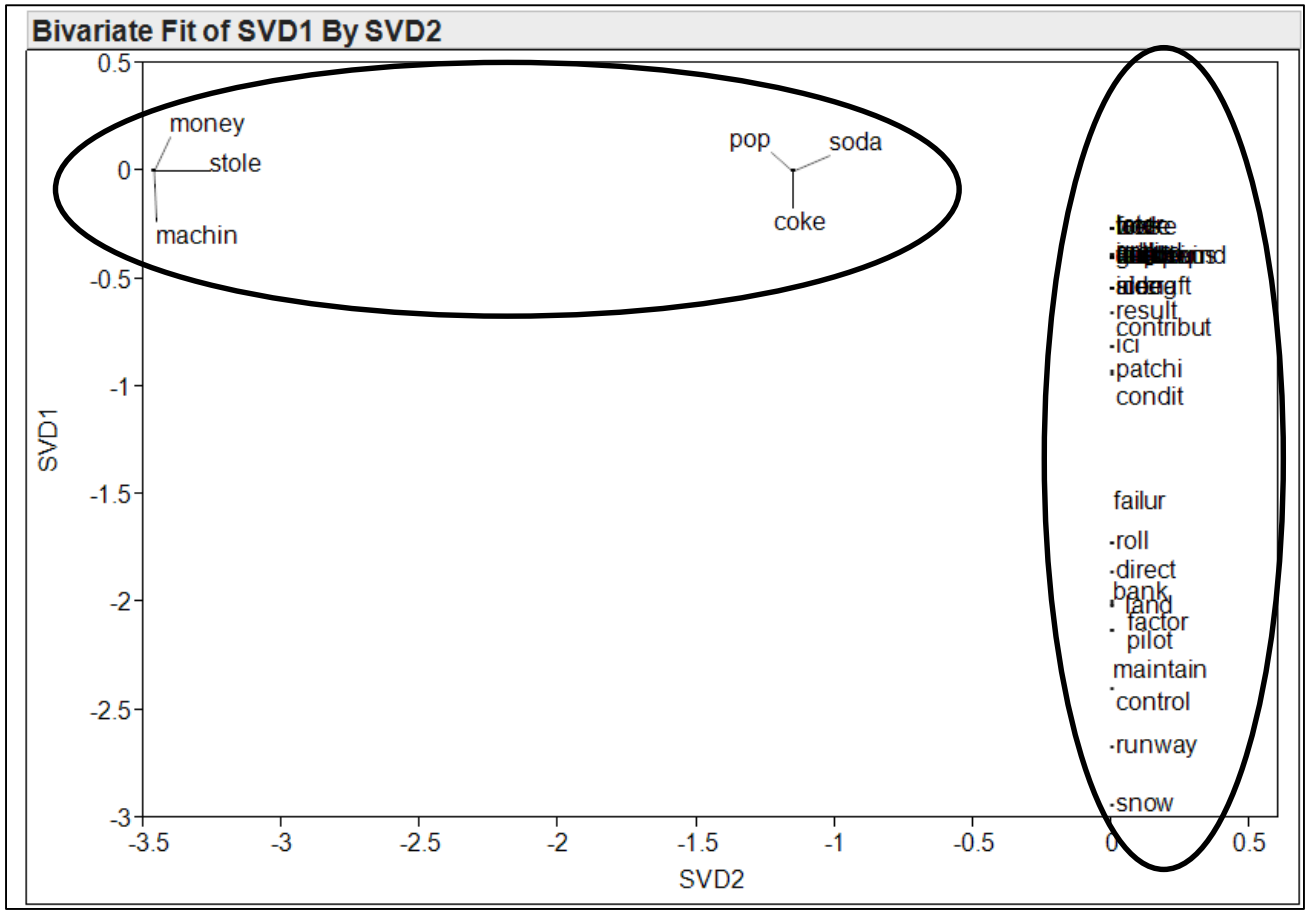
SVD AND SYNONYMS

SVD and Synonyms

	Column 1
1	The soda machine stole my money.
2	The soda machine stole my money.
3	my money, the soda machine stole!
4	The soda machine stole my money.
5	The pop machine stole my money.
6	The pop machine stole my money?
7	The pop machine stole my money.
8	The pop machine stole my money.
9	The Coke machine stole my money.
10	money! coke! machine stole!
11	The Coke machine stole my money.
12	The Coke machine stole my money.
13	The pilot failed to maintain directional control of the airplane due to inadequate compensation fo
14	the pilot's failure to maintain directional control during landing roll which resulted in a collision wit
15	The pilot's failure to maintain directional control during the landing roll. A factor included the pat
16	a wet brake that later froze resulting in a loss of directional control on landing. A contributing fac
17	The pilot's failure to maintain directional control during the landing roll. An icy runway and a sno
18	the pilot's failure to maintain aircraft control during landing roll. Contributing factors were the run

Soda, Coke, and pop are three different words to describe the same idea. In addition to the soda theme, there is an unrelated group of NTSB documents.

SVD and Synonyms



- SVD allows us to recognize that soda, pop, and coke are synonyms, even though these terms never appear in the same document.
- The plot of the first two singular vectors shows that there is a soda machine "theme" and an aviation "theme."

CLUSTERING

Clustering

- Once we have produced either a DTM or an SVD of a DTM, we may use the resulting numeric columns with clustering algorithms to answer questions such as
  - Which groups of documents are most similar?
  - Which documents are most similar to a particular document?
  - Which groups of terms tend to appear either together in the same documents or together with the same words?
  - Which terms are most similar to a particular term?
  - Are certain clusters of documents more strongly related to other variables (e.g. income, cost, fraudulent activity) than other clusters?

Stratifying the Corpus

- Given a new collection of documents to analyze, a first reaction might be to read through a small sample to get a picture of the topics.
- If someone gives you a collection of 20,000 paragraph-long reports and asks you to summarize them in an hour, what would you do?

Stratifying the Corpus

- If you decide you have time to read 40 reports, you could group the documents into 40 clusters (using hierarchical or k-means clustering).
- Then, sample one document from each cluster.
- This will give a more representative subset of documents than would be found by taking a simple random sample.

Clustering Algorithms

- Hierarchical clustering
  - Not practical as the number of documents, *d*, grows since it needs to calculate and manipulate a triangle of the $d \times d$ distance matrix
  - Ward's method: tends to be a best choice for most applications (gives similar results to average clustering).
  - Single link: often leads to chaining of unrelated documents
  - Complete link: better than single link, but sensitive to outliers
  - http://nlp.stanford.edu/IR-book/pdf/17hier.pdf

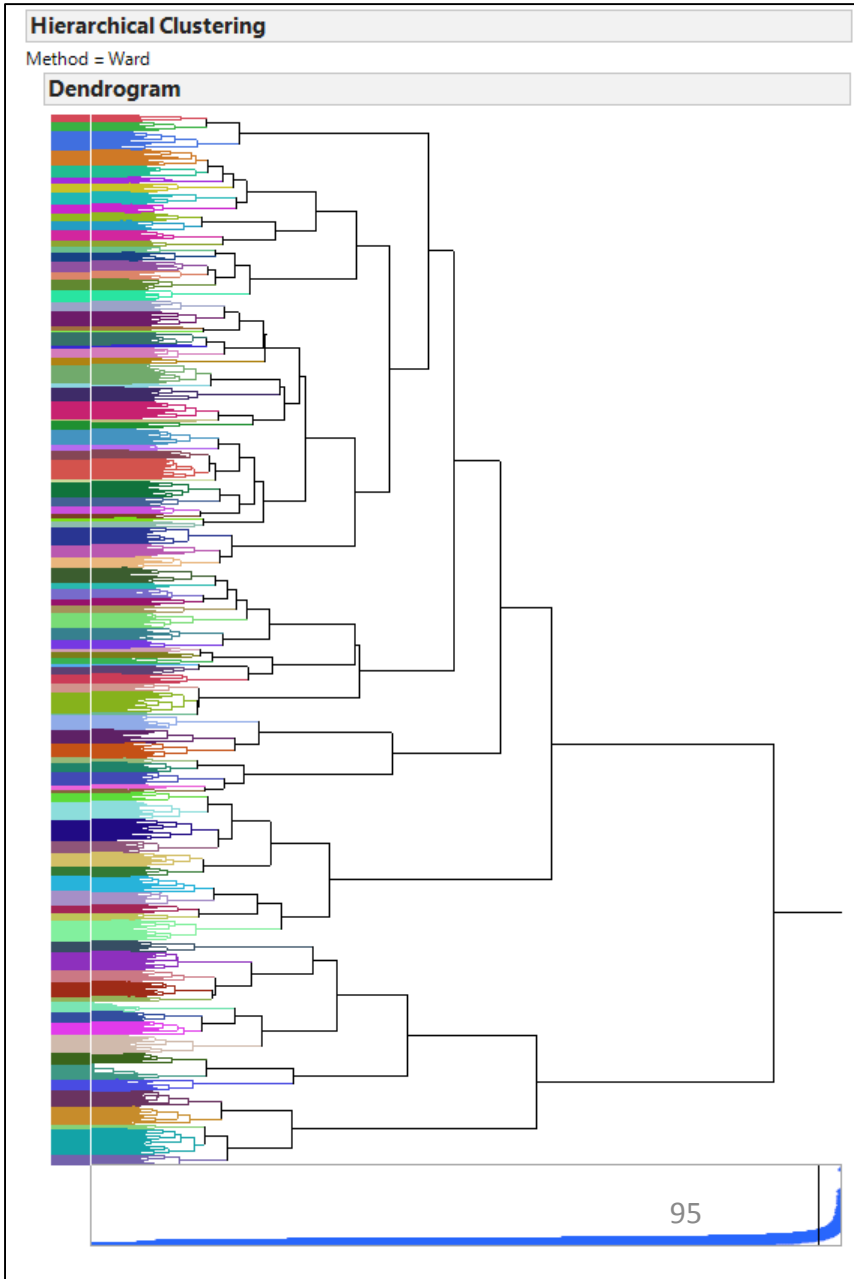
Clustering Algorithms

- K-means
  - Computationally simpler than hierarchical clustering
  - More difficult to select the appropriate number of clusters

Document Clustering

- At the right is a dendogram for a hierarchical clustering using Ward's method in JMP on the 300 vectors returned by the rank-reduced SVD.

- The joining distances, shown at the bottom, guide our selection of the number of clusters (100 in this case).



Concentration of Fatal Accidents in Clusters

Cluster	N(NO)	Row %(NO)	N(YES)	Row %(YES)
188	38	100.00%	0	0.00%
169	35	100.00%	0	0.00%
189	33	100.00%	0	0.00%
236	31	100.00%	0	0.00%
193	29	100.00%	0	0.00%
176	26	100.00%	0	0.00%
180	26	100.00%	0	0.00%
192	26	100.00%	0	0.00%
60	25	100.00%	0	0.00%

- None of the accidents described in these clusters were fatal.

Bounced Landings: Not Fatal

Cluster	fatal	narr_cause
188	NO	The loss of control on landing due to the student's improper recovery from a bounced landing, and the resulting nose over on the grass runway.
188	NO	The student's failure to maintain control of the aircraft during landing due to his improper landing flare height and improper recovery from a bounced lar
188	NO	the student pilot's failure to recover from a bounced landing, which resulted in porpoising and subsequently a nose over.
188	NO	The pilot's premature flare, which resulted in an inadvertent stall and a bounced landing. A factor was the improper recovery from a bounced landing.
188	NO	the student pilot's improper recovery from a bounced landing.
188	NO	The pilot's improper flare, and improper recovery from a bounced landing.
188	NO	The pilot's improper flare and his improper recovery from a bounced landing.
188	NO	The pilot's improper recovery from a bounced landing.
188	NO	The student pilot's failure to maintain aircraft control during the landing, her failure to recover from the bounced landing, and the nose gear overload.
188	NO	The student pilots improper flare, and improper recovery from a bounced landing. A factor was the student pilot's lack of total experience.
188	NO	The pilot's inadequate recovery from a bounced landing. A factor associated with the accident was a crosswind.
188	NO	The pilot's improper recovery from a bounced landing.
188	NO	An inoperative airspeed indicator and the pilot's improper recovery from the bounced landing.
188	NO	The pilot's improper flare, and improper recovery from a bounced landing.
188	NO	The student pilot's improper flare and recovery from a bounced landing.
188	NO	The pilot's inadequate recovery from a bounced landing which resulted in a hard contact with the runway. Factors associated with the accident were t
188	NO	The pilot's improper recovery from a bounced landing. A factor in the accident was the pilot improper flare.
188	NO	The student pilot's improper flare and failure to recover from a bounced landing resulting in the subsequent collapse of the nose gear during the landing
188	NO	The pilot's improper recovery from a bounced landing. A factor was the pilot's failure to flare during initial touchdown.
188	NO	The pilot's misjudgment of distance, his subsequent improper recovery from a bounced landing, and the failure to maintain airspeed which resulted in
188	NO	the student pilot's failure to properly recover from a bounced landing which resulted in the airplane porpoising. A contributing factor was the student pi

Soft Terrain: Not Fatal

Cluster	fatal	narr_cause
169	NO	The pilot's failure to maintain a proper glidepath during final approach. A factor associated with the accident was soft terrain.
169	NO	The pilot's failure to maintain directional control during the landing roll. Factors were the crosswind and soft terrain condition.
169	NO	The pilot's inadequate preflight planning/preparation, and his selection of unsuitable terrain for landing. A factor in the accident was snow-covered terrai
169	NO	The pilot's selection of unsuitable terrain for takeoff, and his inadequate preflight planning/preparation resulting in a collision with trees during the initial c
169	NO	The pilot's inadvertent stall while maneuvering. A factor associated with the accident was soft, snow-covered terrain.
169	NO	The pilot's selection of unsuitable terrain for landing and subsequent nose over during the landing flare. Factors in the accident were soft, snow-covered
169	NO	the pilot's failure to maintain directional control during the takeoff initial climb. Contributory factors were the pilot's lack of experience with the aircraft ar
169	NO	the rocker assembly failure during low level maneuvering. Factors were the soft and sandy terrain and the unsuitable terrain the pilot encountered during
169	NO	A soft area in the turf runway, which resulted in a loss of directional control during the landing rollout.
169	NO	The pilot's selection of unsuitable terrain for landing. Factors in the accident were a soft area of runway, and sunglare.
169	NO	The pilot's selection of unsuitable terrain for takeoff. Factors in the accident were soft terrain, and the pilot's delay in aborting the takeoff.
169	NO	The pilot's selection of an unsuitable landing area. A factor associated with the accident was soft terrain.
169	NO	The selection by the pilot of an unsuitable precautionary landing site on soft, uneven terrain, which resulted in a rollover.
169	NO	The inadequate preflight planning by the pilot, the pilot initiating the flight with an inadequate fuel supply, and the unsuitable terrain encountered during t
169	NO	The inadequate fuel supply for the flight which resulted in fuel exhaustion. A factor associated with the accident was the low altitude and the soft terrai
169	NO	The pilots failure to maintain directional control during the landing. Factors were the crosswind and the soft terrain.
169	NO	the pilot's failure to maintain directional control during the landing roll, which resulted in the airplane departing the runway, impacting with a windsock, a
169	NO	the unsuitable terrain for landing selected by the pilot. A factor was the soft terrain.
169	NO	the pilot's improper rotation and failure to maintain directional control during takeoff. Additional factors were the crosswind and the soft terrain.
169	NO	A loss of engine power for undetermined reasons, which resulted in a forced landing and subsequent nose over during landing roll. A factor was the soft
169	NO	The student pilot's failure to maintain directional control of the airplane during the landing roll. A contributing factor was the soft terrain.
169	NO	The improper planning/decision in runway selection. The soft runway condition and wet snow were contributing factors.
169	NO	Loss of engine power for undetermined reasons. Soft terrain was a factor.
169	NO	The pilot's decision to continue the takeoff. A factor in the accident was the soft wet runway.
169	NO	The pilot's use of unsuitable terrain (landing surface) at his privately owned landing site. A contribtuing factor was the soft area which the aircraft's right
169	NO	The pilot did not maintain directional control and executed improper use of the brakes. A factor associated with the accident was the soft terrain.

Concentration of Fatal Accidents in Clusters

Cluster	N(NO)	Row %(NO)	N(YES)	Row %(YES)
196	3	42.86%	4	57.14%
27	2	40.00%	3	60.00%
139	7	36.84%	12	63.16%
82	11	35.48%	20	64.52%
155	1	33.33%	2	66.67%
208	7	31.82%	15	68.18%
206	3	30.00%	7	70.00%
81	2	28.57%	5	71.43%
83	2	28.57%	5	71.43%
77	4	25.00%	12	75.00%
50	1	20.00%	4	80.00%
187	7	19.44%	29	80.56%
53	2	18.18%	9	81.82%
205	2	10.00%	18	90.00%
222	0	0.00%	1	100.00%

- The clusters at the bottom of this table have a higher concentration of fatal accidents.

Spatial Disorientation: Fatal

Cluster	fatal	narr_cause
205	NO	Improper weather evaluation by both the pilot and pilot/passenger, and the pilot's inadvertent VFR flight into IMC resulting in his spatial disorientation.
205	YES	The pilots decision not to fly to the alternate airport, his decision to continue the flight in known adverse weather conditions, spatial disorientation by tl
205	YES	The pilot's failure to maintain control due to spatial disorientation.
205	YES	The pilot flying at an altitude insufficient to clear surrounding terrain. Contributing factors were the pilot becoming lost/disoriented, his subsequent spat
205	YES	The pilot's spatial disorientation due to a night visual illusion. A factor was the dark night condition.
205	YES	the pilot's spatial disorientation, which led to his failure to maintain aircraft control. A contributing factor was the pilot's decision to intentionally fly into
205	YES	the pilot's continued VFR flight into IMC, which resulted in spatial disorientation and the ensuing loss of aircraft control while in cruise flight. Contribut
205	YES	the pilot's VFR flight into IMC, which resulted in spatial disorientation and a loss of aircraft control. A contributing factor to the accident was the pilot's
205	YES	The pilot experienced spatial disorientation, which resulted in an in-flight loss of control and subsequent collision with trees and terrain. A factor was tl
205	YES	The pilot's failure to maintain a proper climb rate while taking off at night, which was a result of spatial disorientation. Factors in the accident were the
205	YES	The pilot's loss of control in flight due to spatial disorientation, and his subsequent overstress of the airplane during a recovery attempt. A factor in the
205	YES	The pilot initiated a VFR flight into known IMC conditions which resulted in a loss of control of the airplane due to spatial disorientation. Factors were t
205	NO	Pilot's failure to maintain adequate separation from terrain during the initial climb. Factors include spatial disorientation and a dark moonless night.
205	YES	The pilot's becoming lost and disoriented and his failure to maintain control of the airplane while flying over an unpopulated area on a dark night, which
205	YES	the pilot's failure to maintain aircraft control and his inadvertent flight into known adverse weather conditions. Factors relating to this accident were the
205	YES	The pilot experienced spatial disorientation that resulted in the loss of control.
205	YES	Flight into known adverse weather conditions by the pilot and the spatial disorientation of pilot. Contributing factors were the lack of certification by th
205	YES	The pilot's failure to follow operating procedures and, experienced spatial disorientation while attempting a night landing to an offshore platform. A fact
205	YES	The pilot's spatial disorientation, which resulted in his subsequent loss of control of the airplane. A factor was the dark night, over water visual conditi
205	YES	The pilot's spatial disorientation during a missed approach, which resulted in a loss of control, and the airplane's subsequent impact with water. Fact

Drugs: Fatal

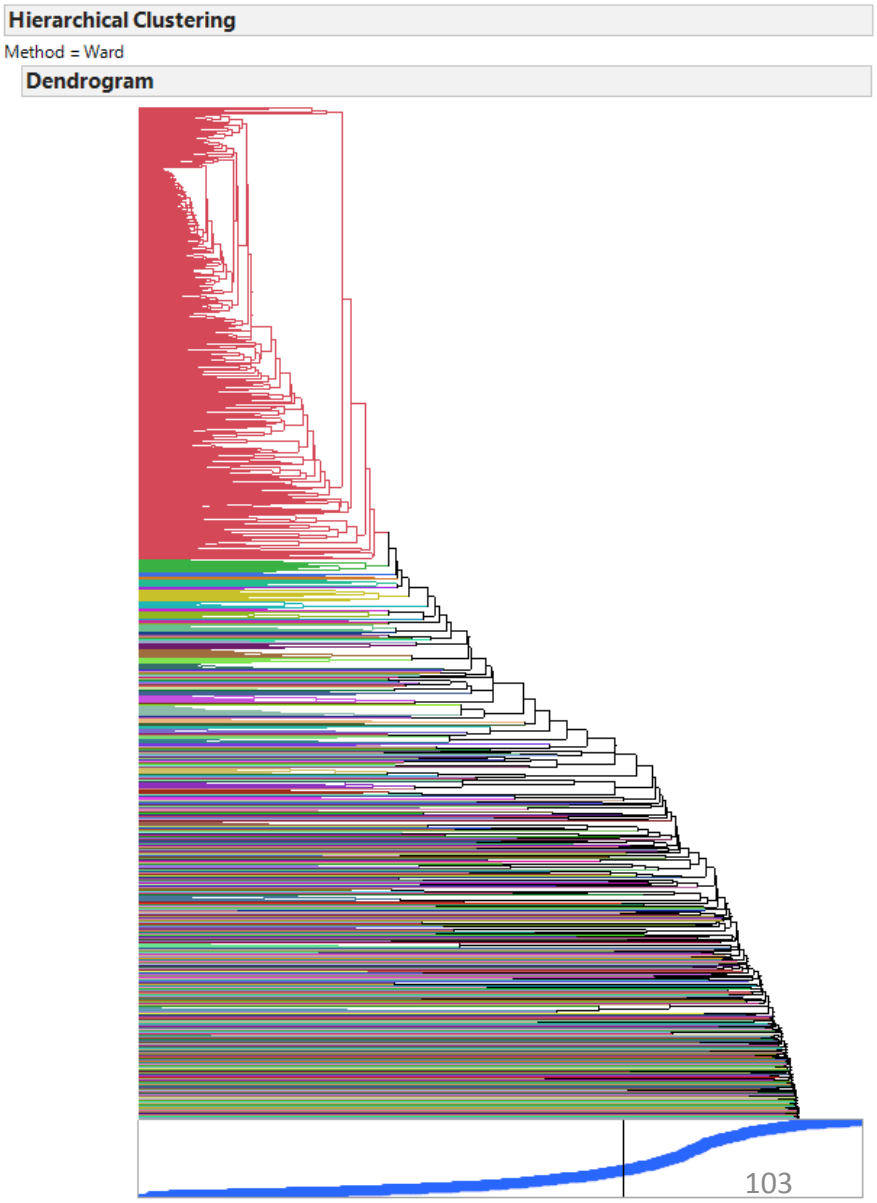
Cluster	fatal	narr_cause
53	YES	The airplane flightcrew's failure to maintain adequate distance/altitude from mountainous terrain during a departure climb to cruise flight, and the captain's impairment from drugs. Factors ir
53	YES	The pilot's inadequate altitude clearance above water while conducting low level flight maneuvers. A factor related to the accident was the pilot's impairment of judgment due to alcohol cons
53	YES	The pilot's failure to maintain aircraft control during takeoff. A factor was the pilot's impairment due to a narcotic painkiller and antihistamine.
53	YES	The pilot's failure to maintain aircraft control. A factor in the accident was the physiological impairment of the pilot due to the consumption of alcohol.
53	YES	The pilot's unsuccessful recovery from an intentional aerobatic stall/spin maneuver. Contributing to the accident were the pilot's impairment (alcohol), and his psychological condition.
53	YES	The pilot inadvertently stalled the airplane. A factor was the impairment due to marihuana.
53	YES	The inadvertent flat spin of the airplane by the flightcrew resulting from the flight instructor's inadequate supervision. A contributing factor was the impairment (drugs) of the private pilot.
53	YES	The pilot's unsuccessful corrective action (recovery) from an inverted spin. A contributing factor was the pilot's encounter with the inverted spin maneuver.
53	NO	The pilot's failure to maintain control of the aircraft which resulted in an uncontrolled descent and an in flight collision with water. Contributing to the accident was the impairment of the pilot
53	YES	The pilot's failure to maintain adequate airspeed which resulted in an inadvertent stall, and subsequent collision with terrain. A contributing factor was the pilot's impairment from the effects
53	NO	The pilot's physical impairment due to a previous head injury which resulted in his becoming disoriented. A contributing factor was the lack of suitable terrain for the precautionary landing.

Failure to Maintain Airspeed: Fatal

Cluster	fatal	narr_cause
187	YES	The pilot's failure to maintain airspeed, which resulted in an inadvertent stall/spin while on base leg.
187	YES	the inadvertent stall/spin. Additional factors included the aerobatic manuevers, low altitude, and the procedures not followed.
187	YES	the pilot in command inadvertently allowing the airplane to stall/spin. Contributing factors were the pilot's total lack of experience in airplane make/mode
187	YES	The pilot not maintaining aircraft control during the initial climb after takeoff and the inadvertent stall/spin. A factor to the accident was the pilot's total la
187	YES	the pilot's failure to maintain aircraft control due to his incapacitation for an undetermined reason. A contributing factor was the subsequent inadvertent
187	YES	the pilot's failure to maintain aircraft control following a loss of engine power while maneuvering, which resulted in an inadvertent stall/spin. Contributing
187	YES	The student's failure to maintain adequate airspeed during the crosswind climb that resulted in a stall/spin at low altitude and the airplane's subsequent
187	YES	The pilot's failure to maintain control of the airplane resulting in the inadvertent stall/spin. A factor was the pilot's unfamiliarity with the airplane.
187	YES	The pilot's failure to maintain airspeed during an aerobatic maneuver, which resulted in an inadvertent inverted spin.
187	YES	The pilot's improper use of the flight controls while turning to base, which resulted in a stall/spin and subsequent impact with the ground.
187	YES	the failure of the pilot to maintain airspeed, which resulted in an inadvertent stall/spin, and subsequent impact with the terrain.
187	YES	The pilot not performing an aborted takeoff and the inadvertent stall he encountered on his inadvertent initial climb. Factors were his inadvertent lift-off, t
187	YES	the failure of the pilot to maintain airspeed, while attempting a forced landing following a loss of engine power for undetermined reasons, which resulted i
187	YES	The inadvertent stall/spin by the pilot.
187	YES	the pilot's failure to maintain aircraft control during the base turn, which resulted in an inadvertant stall/spin.
187	NO	loss of engine power due to both piston rings failing, and the subsequent inadvertent stall/spin during the attempted forced landing. A contributing facto
187	NO	The inadvertent stall/spin encountered by the pilot during a slow flight maneuver. Factors relating to this accident were the low airspeed and the trees.
187	YES	the failure of the pilot to maintain airspeed, which resulted in an inadvertent stall/spin, and subsequent impact with trees, while at a low altitude.
187	YES	The pilot's failure to maintain airspeed during a low-altitude aerobatic maneuver, which resulted in an inadvertent stall/spin and subsequent uncontrolled
187	YES	The loss of engine power for undetermined reasons, and the pilot's failure to maintain airspeed which resulted in an inadvertent stall/spin.
187	YES	The pilots failure to maintain airspeed while maneuvering in instrument flight conditions resulting in an inadvertence stall/spin (vertical descent) and subs
187	YES	the pilot's failure to maintain control of the airplane while maneuvering resulting in an inadvertent stall/spin.
187	YES	The pilot's failure to maintain airspeed after a loss of engine power, which resulted in an inadvertent stall/spin. Also causal, was the loss of engine powe
187	YES	The pilot's failure to maintain adequate airspeed during the turn to final, which resulted in an inadvertent stall/spin. Factors included low ceilings and nig
187	YES	the pilot's failure to maintain control of the airplane resulting in the airplane entering a flat spin from which the pilot did not recover.
187	YES	The pilots' failure to maintain airspeed, which resulted in an inadvertent stall/spin. The continued spin to the ground was a result of the pilots' failure to c

Clustering Terms

- Often, there will be a large cluster (seen at right) of unimportant terms



Clustering Terms

- When using vectors returned by the SVD, terms in clusters may be synonyms: they may never appear in the same document, but may appear with a common collection of words

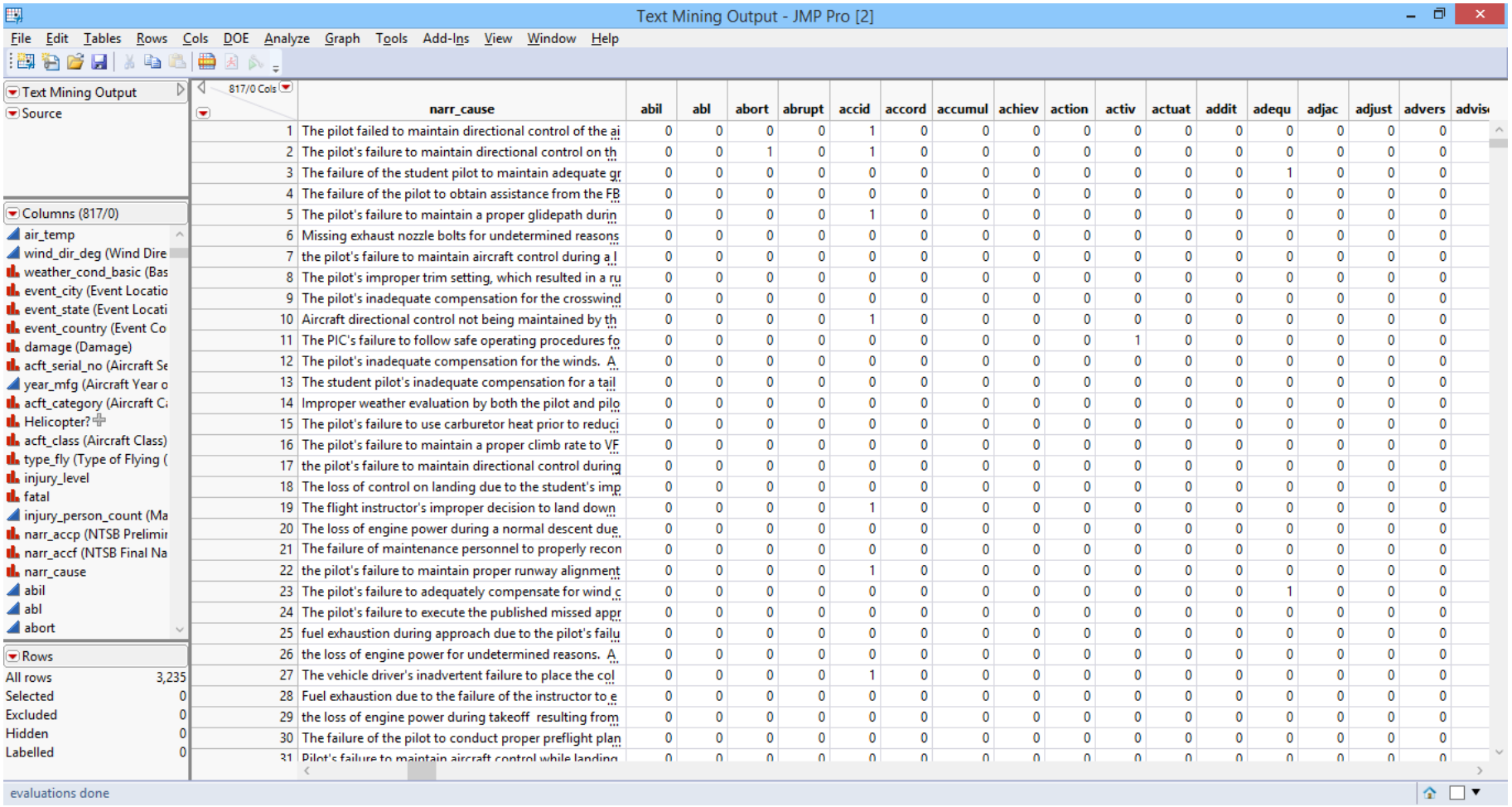
	Label	Cluster
1	reason	441
2	undetermin	441
3	attent	332
4	divert	332
5	disorient	311
6	spatial	311
7	pattern	305
8	traffic	305
9	defici	284
10	known	284
11	rough	275
12	uneven	275
13	pole	262
14	util	262
15	hing	233
16	spring	233
17	stud	233
18	tab	233
19	tension	233
20	worn	233

DSURGO
ascend with analytics

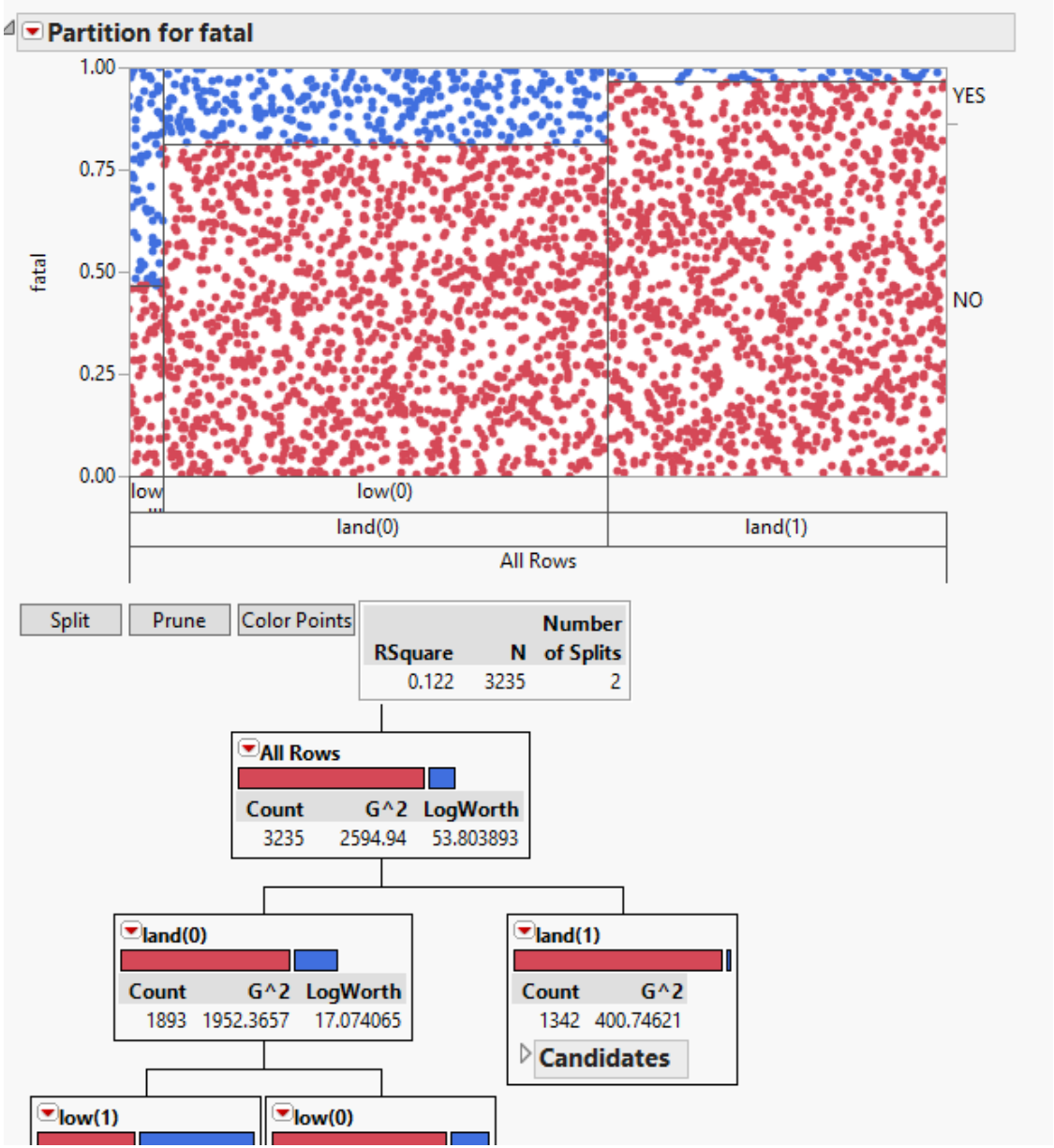
Modeling a Target Variable

- In some applications, there will be one or more variables of interest that are reported along with the text. For example, an error report may contain a text report from an engineer, along with a measure of processing time/cost.
- Using the DTM or its SVD, we can
  - study the relationship between individual terms and the target variable
  - predict future values of the target variable using text input

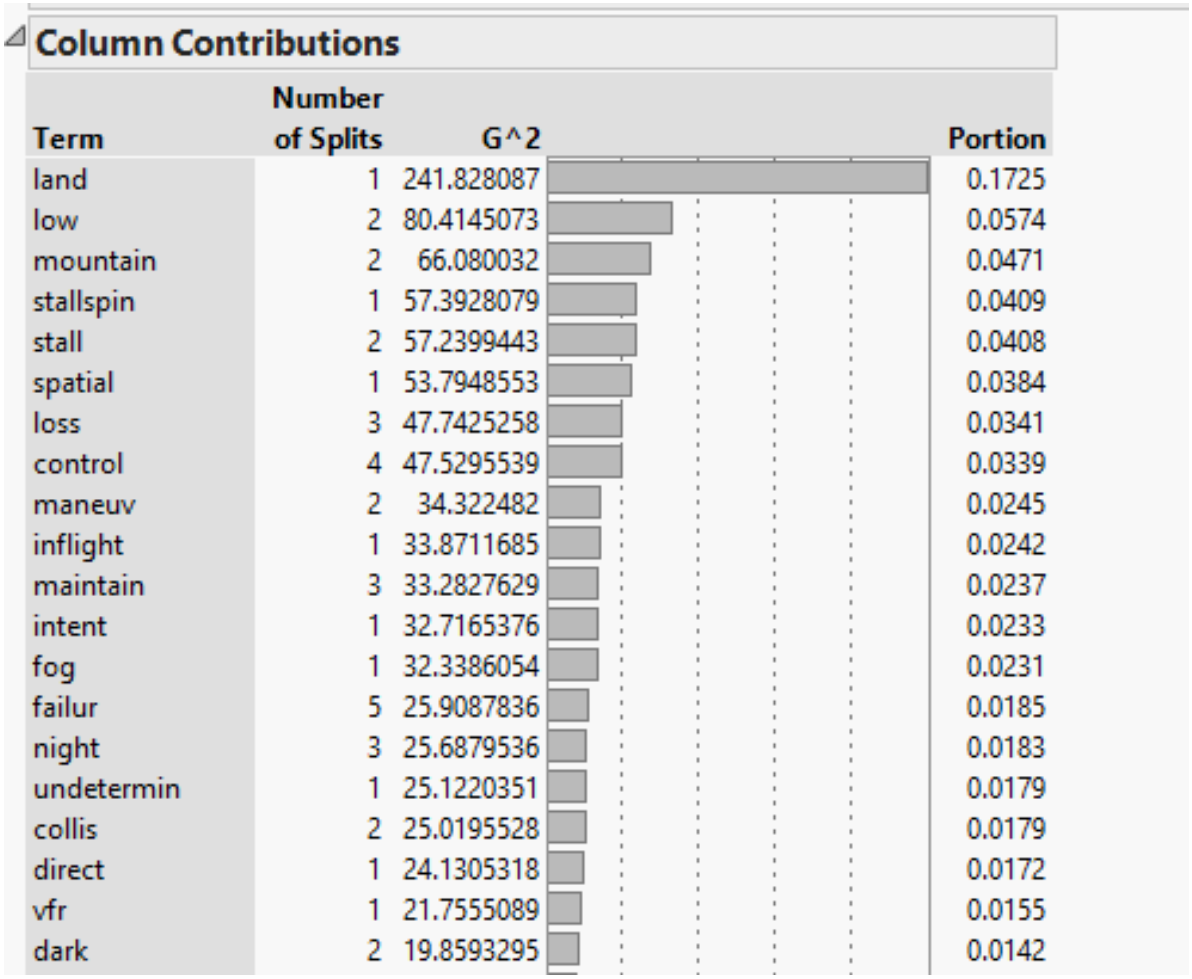
CART for *Fatal* on DTM



CART for *Fatal* on DTM



CART for *Fatal* on DTM



Column Contributions

Term	Number of Splits	G^2	Portion
land	1	241.828087	0.1725
low	2	80.4145073	0.0574
mountain	2	66.080032	0.0471
stallspin	1	57.3928079	0.0409
stall	2	57.2399443	0.0408
spatial	1	53.7948553	0.0384
loss	3	47.7425258	0.0341
control	4	47.5295539	0.0339
maneuv	2	34.322482	0.0245
inflight	1	33.8711685	0.0242
maintain	3	33.2827629	0.0237
intent	1	32.7165376	0.0233
fog	1	32.3386054	0.0231
failur	5	25.9087836	0.0185
night	3	25.6879536	0.0183
undetermin	1	25.1220351	0.0179
collis	2	25.0195528	0.0179
direct	1	24.1305318	0.0172
vfr	1	21.7555089	0.0155
dark	2	19.8593295	0.0142

APPLICATIONS OF TEXT MINING

Web Crawler – FDA Recalls

- Data: FDA Recall data.
- Objective: Crawl fda.gov to develop a corpus of recalls for a specified year.
- Software used: SAS/JMP script with R.

Text Mining Example – Recall Data

- Data: Medical device recall data from fda.gov.
- Objective: Use text mining to summarize issues in medical device recalls for a specified year.
- Software used: SAS/JMP script with R.

Text Mining Example – Recall Data

- Data: Medical device recall data from fda.gov.
- Objective: Develop a corpus of recalls from a folder of text documents (for a specified company and year).
- Software used: SAS/JMP script with R.

Text Mining Example – Recall Data

- Data: Medical device recall data from fda.gov.
- Objective: Use text mining to summarize issues in medical device recalls for a specific company in a specific year.
- Software used: SAS/JMP script with R.

Text Mining Example – Inspection Observations

- Data: Inspection observations from fda.gov.
- Objective: Determine the most frequent themes in inspection observations for a particular industry (medical device, drugs, biologics).
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

Text Mining Example – Inspection Citations

- Data: Inspection citations from fda.gov.
- Objective: Use inspection citations to determine if certain compliance themes are associated with certain companies.
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

OPTIONAL

Survey Analysis Example – Open-ended Questions

- Data: 315 respondents to a survey by a company.
- Objective: Use text mining, word clouds/frequency in a set of comments from open-ended survey responses to find general themes.
  - Why does the respondent feel that a company has the best loyalty program?
  - Why does the respondent feel that a company has the worst loyalty program?
  - Why does the respondent feel that a store is his/her favorite stop to shop?
  - Why does the respondent feel that a store is his/her least favorite stop to shop?
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Concept Extraction Example – Unabomber Manifesto

- Data: 21 Op Eds from the NY Times and the Unabomber Manifesto.
- Objective: Use text mining techniques to provide information into the author of the Unabomber Manifesto.
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Sentiment Analysis – Cars

- Data: 638 car reviews written by owners.
- Objective: Determine general sentiment about automobiles. Determine if general sentiments can be 'attached' to certain types of automobiles.
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

Concept Extraction Example – Movies

- Data: 1,527 randomly selected movie synopses.
- Objective: Use text mining techniques (using the DTM) to determine if specific text strings can be used to predict box office success.
- Software used: SAS/JMP script with R.

Text Analytics Using SAS® Text Miner. SAS Institute: Cary, 2011.

Social Media - Twitter

- Data: Live Twitter data
- Objective: Determine social media reaction to a certain current event.
- Software used: SAS/JMP script with R.

Text Mining Example – NTSB Aircraft Accident Reports

- Data: NTSB Aircraft Accident Reports.
- Objective: Determine what factors contributed to fatal accidents.
- Software used: SAS/JMP script with R.

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

APPENDIX

REFERENCES

References

- Textbooks:

Gary Miner, *et al. Statistical Analysis and Data Mining.* Academic Press: Amsterdam, 2009.

Gary Miner, *et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press: Oxford, 2012.

Text Analytics Using SAS® Text Miner. SAS Institute: Cary, 2011.

Weiss, S., et al. *Text Mining: Predictive Methods for Analyzing Unstructured Information.* Springer Publishing Company, Incorporated: New York, 2009.

- Websites:

http://nlp.stanford.edu/IR-book/pdf/17hier.pdf

http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

http://www.cs.uoi.gr/~tsap/teaching/2012f-cs059/slides-en.html