ADSURGO
ascend with analytics

Speed of Need

Applied Text Analytics for Test and Evaluation

33rd ITEA Symposium
5 October 2016

Jim Wisnowski
james.wisnowski@adsurgo.com
www.adsurgo.com

What is Text Mining?

- Text mining: semi-automated process of detecting patterns (useful information and knowledge) from large amounts of *unstructured* data sources
- Text analysis: an examination of structure, composition, and meaning that provides insight to advance some purpose…that characterize and describe a text itself. Analysis may be heuristic, informal, and/or qualitative.
- Text analytics: methods used for intelligent analyses of textual data; a larger set of activities around inference steps of discovering information, grouping documents, summarizing information, finding themes, and transforming into structured data for predictive analyses.
  - Systematic application of numerical and statistical methods that derive and deliver quantitative information, whether in the form of indicators, tables, or visualizations. Analytics is formal and repeatable.

Essential tool in the T&E "war chest" of capabilities to exploit unstructured data

Text Analytics: Trending Upward

- In 2015, SAS text mined 7,000 data scientist job descriptions for top skills required
- Text analytics is fastest growing area
- **~80% of data is unstructured and is waiting to be analyzed**
- **"Enhanced use of data to improve acquisition program outcomes" SASC Bill for FY17 NDAA**

Top 10 Types of Analysis Mentioned	
ML_NEURAL_NETS_SVM	52%
OPTIMIZATION	23%
TEXT_ANALYTICS	17%
TIME_SERIES	17%
DATA_WRANGLING	16%
CLUSTERING	16%
LINEAR_REGRESSION	15%
DATA_VISUALIZATION	13%
MATLAB	12%
DESIGN_OF_EXPERIMENTS_AB	10%

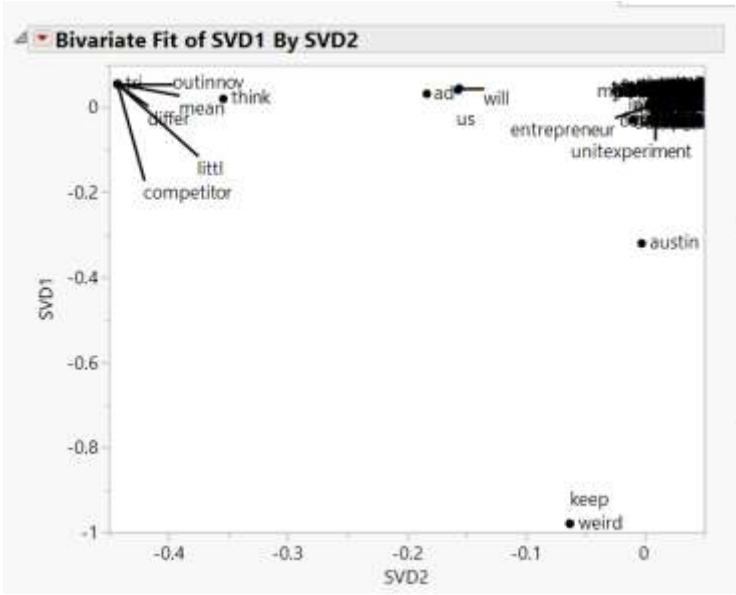
n=7027

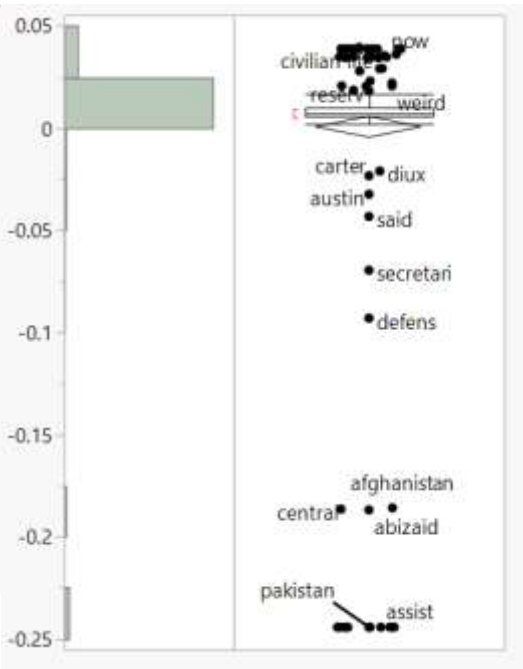
http://blogs.sas.com/content/text-mining/page/2/

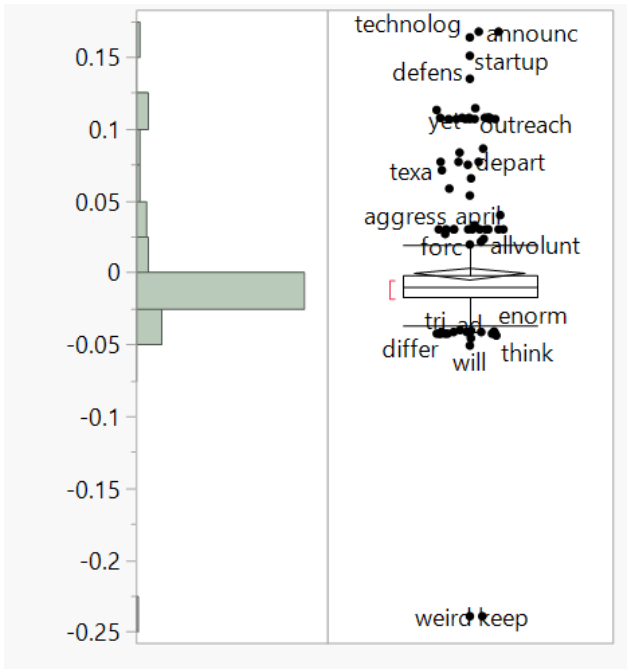
Warm Up…WordCloud

http://www.defense.gov/News/Article/Article/944767/carter-opens-diux-outpost-in-texas

Warm Up…More Structure







- Knowing the word frequencies is helpful
- Understanding what words go together for themes is essential

Some T&E Applications of Text Mining

- Finding patterns in voice-to-text translations of communications during test and evaluation events
- Chat logs and mission reports
- Learning about the new challenging environments and technologies
- Analysis of text fields in system maintenance reports, software trouble reports, FRACAS databases, and other suitability data products
- Using mission report text fields to find and group reports by common themes
- Boosting survey insights with deeper analysis of free text responses
- Mining historical contractor and industry reports and field data
- Summarizing volumes of MIL-STDS; searching for specific related topics
- Sifting through T&E data output from instrumentation
- Identifying patterns through automated software test output analyses

There is a world of unstructured & owned data waiting to inform the evaluation

Social Media Analytics—Twitter Sentiment Analysis

- Real-time feed of social media provides intelligence opportunity
- Example: LeBron James after announcement he is going home
  - All Tweets for 5 mins the day LeBron made his statement
- Sentiment analysis/opinion with text mining tabulates the number of positive terms and number of negative terms (Harvard IV dictionary) from all Tweets

	Negative	Positive
1132	liable	pleasantry
1133	liar	please
1134	lie	pleased
1135	lifeless	pleasurable
1136	limit	pleasure
1137	limitation	pledge
1138	limp	plentiful
1139	liquidate	plenty
1140	liquidation	poetic
1141	litter	poignant
1142	load	poise
1143	lone	polish
1144	loneliness	polite
1145	lonely	politeness
1146	loner	pomp
1147	lonesome	popular
1148	loom	popularity
1149	lose	populous
1150	loser	portable

Positive	Sum	3533
Negative	Sum	1626

Social Media Analytics—Twitter Sentiment Analysis Part II

- Real-time feed of social media provides intelligence opportunity
- Example: LeBron James day after he won championship
- Sentiment analysis/opinion with text mining tabulates the number of positive terms and number of negative terms (Harvard IV dictionary) from all Tweets

	Negative	Positive
1132	liable	pleasantry
1133	liar	please
1134	lie	pleased
1135	lifeless	pleasurable
1136	limit	pleasure
1137	limitation	pledge
1138	limp	plentiful
1139	liquidate	plenty
1140	liquidation	poetic
1141	litter	poignant
1142	load	poise
1143	lone	polish
1144	loneliness	polite
1145	lonely	politeness
1146	loner	pomp
1147	lonesome	popular
1148	loom	popularity
1149	lose	populous
1150	loser	portable

Positive	Negative
Sum	Sum
2722	1257

Text Analytics Flow

Define Problem Statement

- Determine clear study objectives and end-state
- Identify relevant data sources to answer research questions

Collect and Extract Data

- Scrape internet with web crawling and social media tools
- Extract text from disparate file types (pptx, doc, txt, pdf, html)
- Strip off code, figures, extraneous characters

Process and Filter Text

- Clean manually with character functions, queries, filters, R&R
- Remove punctuation, numbers, stop words
- Stem and tokenize text, change to lowercase, identify multiwords.

Transform Text

- Create document term matrix
- Weight matrix based on analysis objectives
- Use Singular Value Decomposition to get structured data

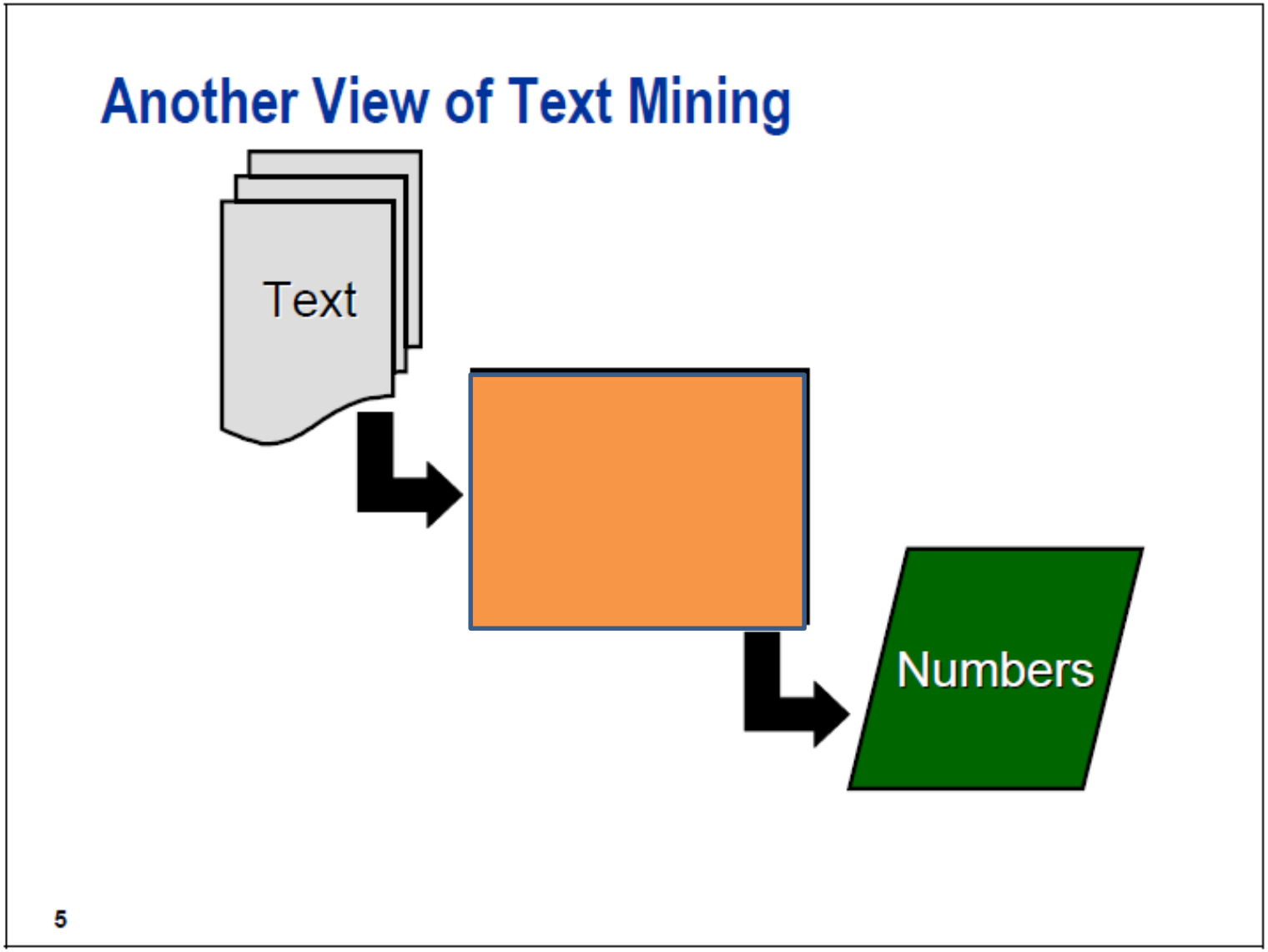
Text Mining Exploration

- Discover topics and common themes
- Group like documents and words
- Subset documents and link concepts

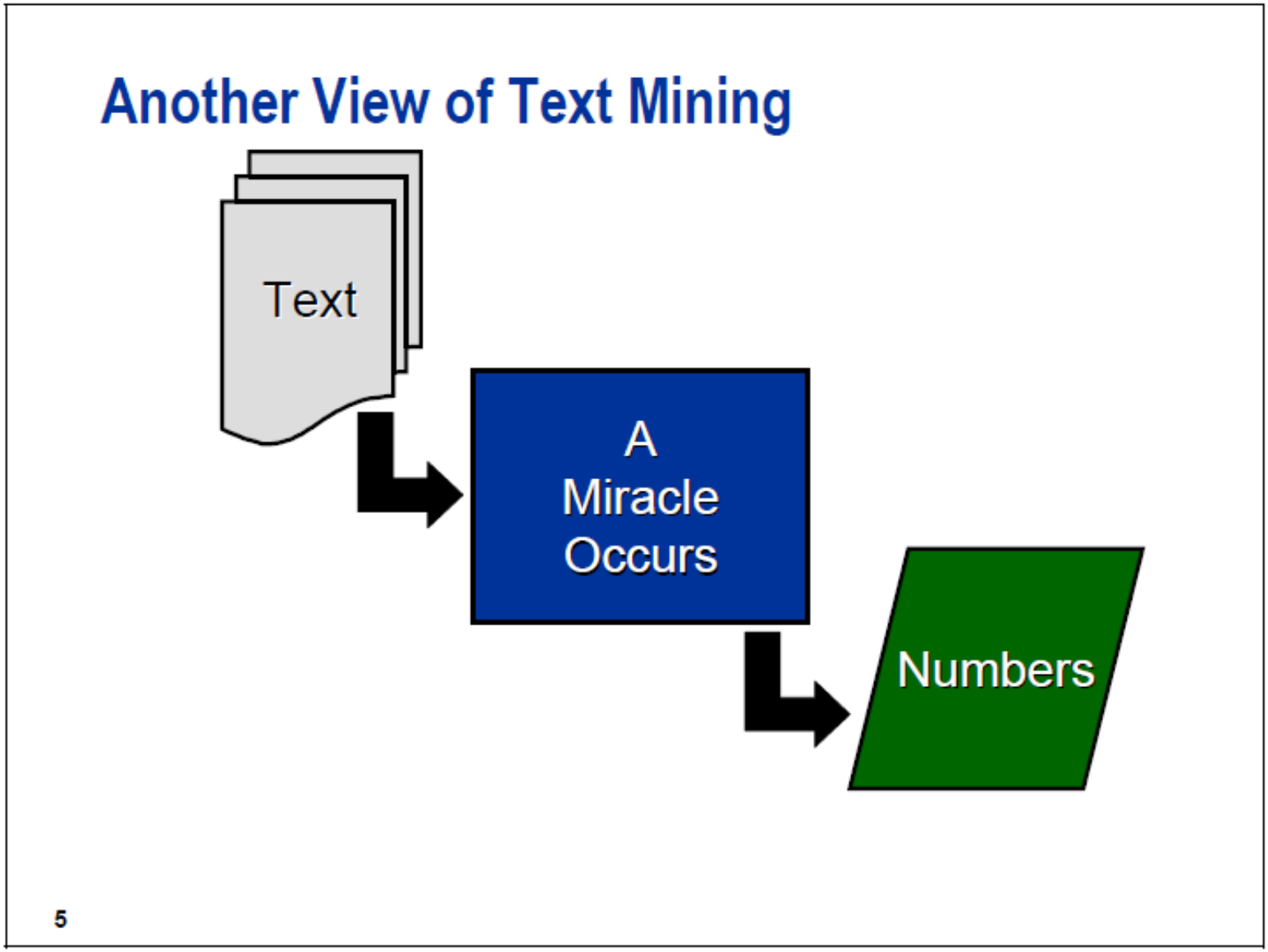
Predictive Analytics

- Combine with structured data
- Visualize exploitable patterns
- Understand sentiments and trends

How Exactly Does TM Work?

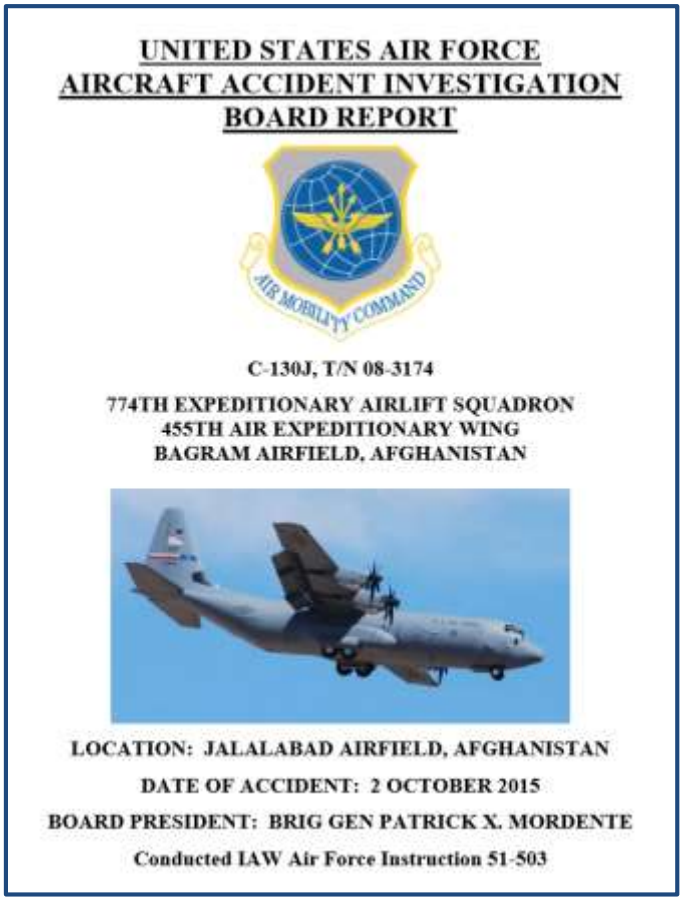


Answer



Unclassified Text Mining Example
Aircraft Accident Reports

- Data: Air Force Accident Investigation Board Reports
- Objectives:
  - what are common themes?
  - what factors contribute to fatal accidents?
  - what words group together?
  - what reports group together?
  - how can we link structured and unstructured data fields?



UNITED STATES AIR FORCE
AIRCRAFT ACCIDENT INVESTIGATION BOARD REPORT

C-130J, T/N 08-3174

774TH EXPEDITIONARY AIRLIFT SQUADRON
455TH AIR EXPEDITIONARY WING
BAGRAM AIRFIELD, AFGHANISTAN

LOCATION: JALALABAD AIRFIELD, AFGHANISTAN

DATE OF ACCIDENT: 2 OCTOBER 2015

BOARD PRESIDENT: BRIG GEN PATRICK X. MORDENTE

Conducted IAW Air Force Instruction 51-503

Word Frequency

- We know word counts alone are highly useful to mine data
  - Origin of abstracts for academic articles base on frequencies.
  - Many companies basing large decisions from word counts in enterprise unstructured data fields like survey responses, voice to text translations.
  - Think of frequency as the ***magnitude*** of a vector.

Key Quantity: Document Term Matrix

- DTM is a sparse matrix with documents as rows and terms as columns=> 3,200 rows (accident reports) by 800 cols (words)
- Tallies the word counts for each document
  - Mostly 0's
  - Can use binary, term frequency, inverse document frequency, and other weighting schemes
- DTM itself is helpful (i.e could do correlation analysis on columns), but need to take it a step further with Latent Semantic Analysis

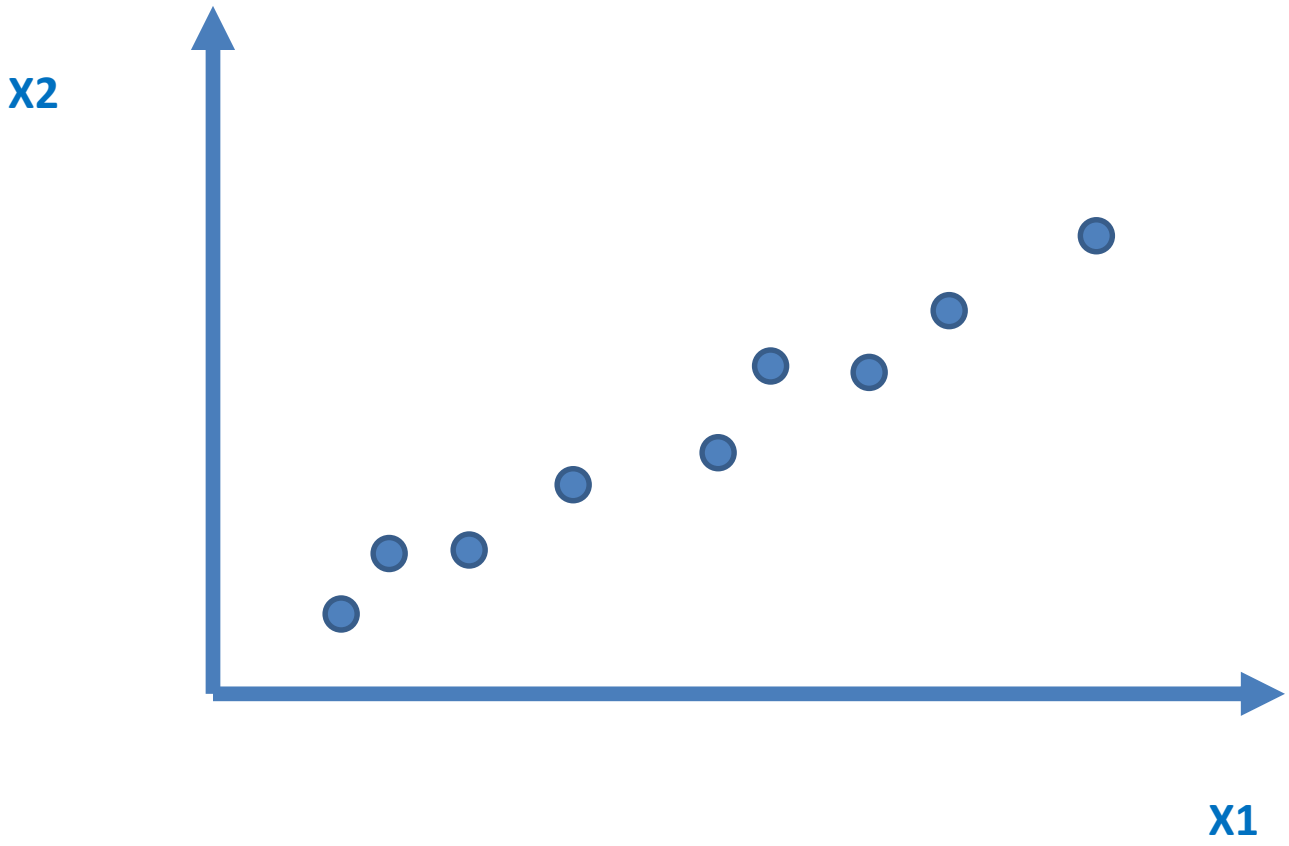
narr_cause	adequ	adjac	adjust	advers	advisori
the pilot's failure to maintain directional control. Factors were the crosswind, th...	0	1	0	0	0
The flight instructor's failure to ensure (supervision) the student had an adequat...	1	0	0	0	0
The pilot's inflight decision to continued visual flight into instrument meteorolo...	0	0	0	0	0
the loss of power to both engines for undetermined reasons during approach. ...	0	0	0	0	0
The pilots decision not to fly to the alternate airport, his decision to continue th...	0	0	0	1	0

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.

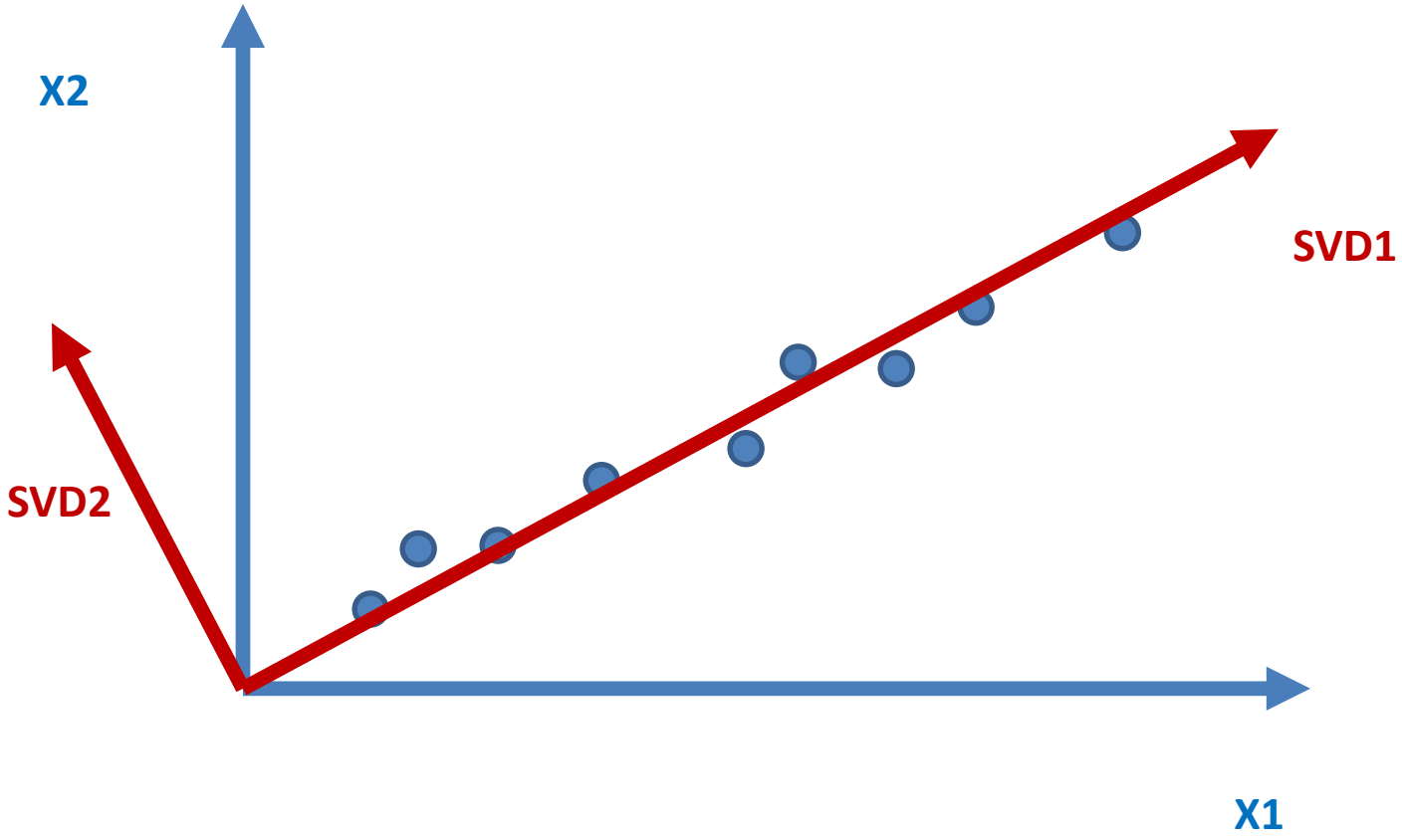
SVD Example

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.



Singular Value Decomposition

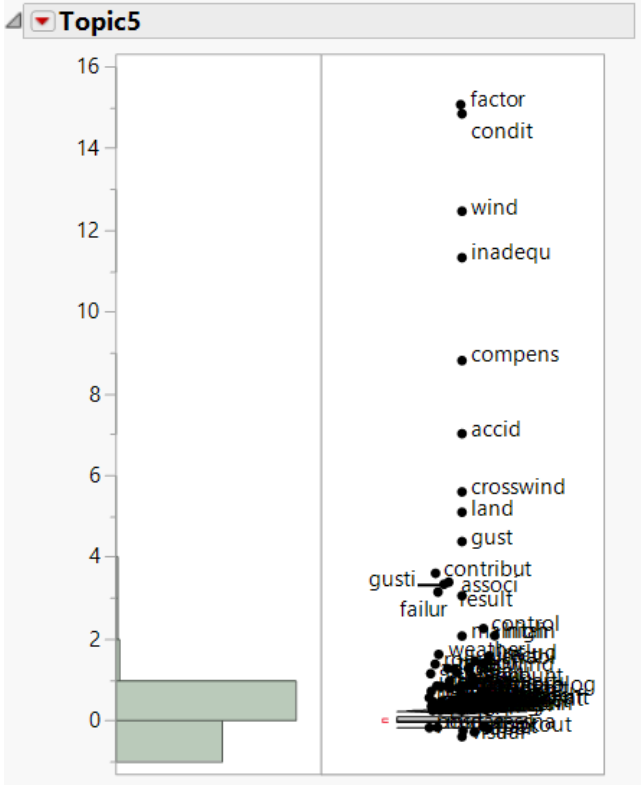
- For a DTM *X,* the SVD factorization is

$$X \approx UDV^t,$$

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)*).

Topic Extraction With SVD V Matrix



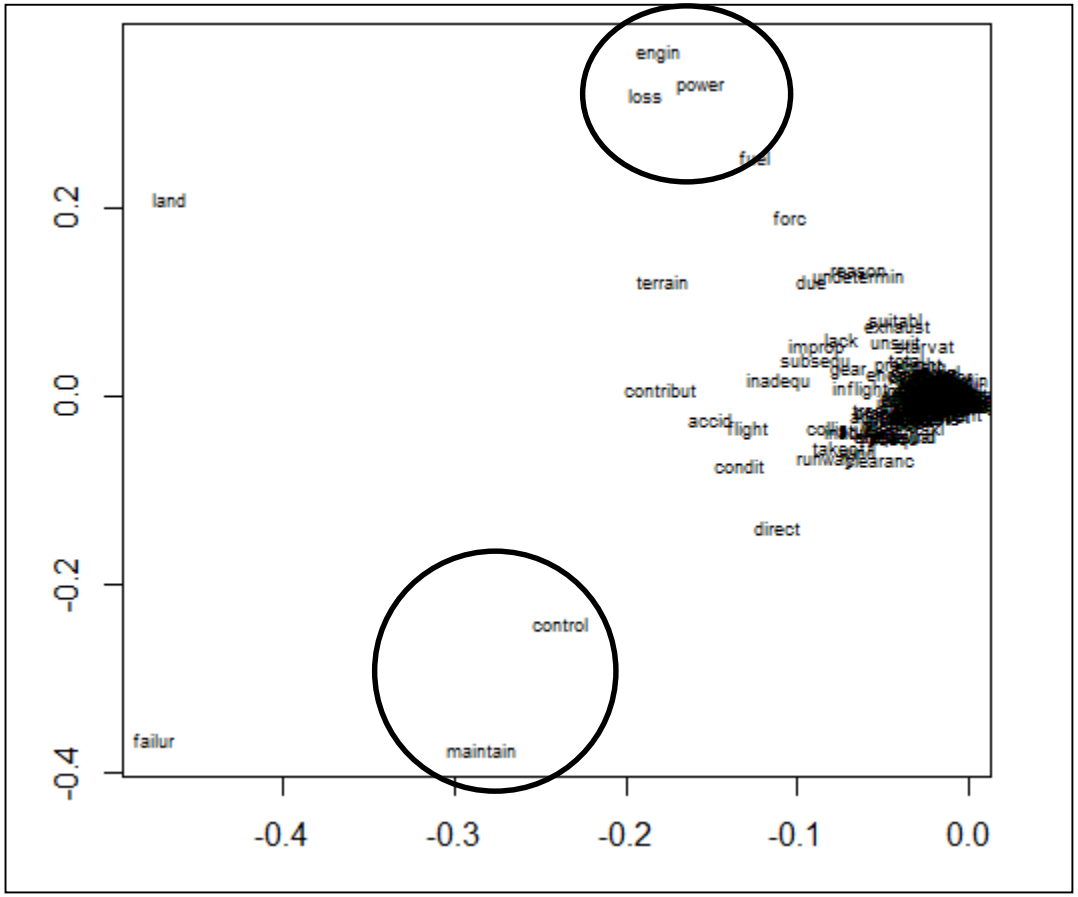
- The 5th eigenvector from the V matrix loads on inadequate compensation for crosswinds

Finding Documents Related to Topic 5: SVD U Matrix

	narr_cause	SVD5
1	The pilot's inadequate compensation for gusty wind conditions. Factors associated with the accident were the pilot's inadequate ...	2.5480873522
2	The pilot's inadequate compensation for wind conditions during the landing roll, which resulted in an inadvertent ground loop/sw...	2.5178671713
3	The pilot's inadequate compensation for the winds conditions which resulted in the failure to maintain directional control of the ai...	2.4973844764
4	The pilot's inadequate compensation for the wind conditions that resulted in directional control not being maintained during the l...	2.4872052169
5	The pilot's inadequate compensation for wind and his failure to maintain directional control of the airplane which resulted in a gro...	2.482312475
6	The pilot's inadequate compensation for wind conditions during the landing roll, resulting in a nose over. A factor associated wit...	2.4742882397
7	The pilot's inadequate compensation for the wind conditions which resulted in the failure to maintain directional control of the air...	2.4740669529
8	The pilot's inadequate compensation for wind conditions during the takeoff roll. A factor associated with the accident was a cross...	2.4421888586
9	The pilot's inadequate compensation for wind conditions while on approach, and the inadvertent stall of the airplane while attem...	2.4388091067
10	The pilot failed to maintain directional control of the airplane due to inadequate compensation for the wind conditions. Factors a...	2.4363088679
11	The pilot's inadequate compensation for wind conditions. Factors associated in the accident were a crosswind, and a worn tailwhe...	2.4288790043
12	The pilot's inadequate compensation for wind conditions. A factor associated with the accident was a crosswind.	2.4158701461
13	the inadequate rotation speed and compensation for wind conditions by the pilot. Contributing factors were the crosswind and g...	2.3758763124
14	The pilot's inadequate compensation for wind conditions during the takeoff run, which resulted in a loss of control and subsequen...	2.3750758348
15	the pilot's inadequate compensation for wind conditions. Factors were the crosswind and the gusts.	2.3677401171
16	The pilot displayed inadequate compensation for the wind conditions that existed at the time of the accident and directional contr...	2.3549840887
17	the pilot's inadequate compensation for the wind conditions which resulted in a loss of control during landing. A contributing fac...	2.3503264913
18	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control during an aborted landing....	2.3333072329
19	The pilot's inadequate compensation for wind conditions which resulted in an in-flight collision with trees. A factor related to the ...	2.3188632769
20	Inadequate preflight planning and inadequate compensation for the wind conditions which resulted in a failure to maintain directi...	2.3182120659
21	The pilot's inadequate compensation for wind conditions while landing. Factors associated with the accident were the pilot's inad...	2.3039591872
22	The pilot's inadequate compensation for wind conditions during initial climb, which resulted in an in-flight collision with trees. A f...	2.2971297018
23	The pilot's inadequate compensation for wind conditions during takeoff, which resulted in an in-flight collision with trees. A facto...	2.2900786845
24	The student pilots inadequate compensation for wind conditions which resulted in an off-field landing, and the CFI's improper dec...	2.2833687328
25	The pilot's inadequate compensation for wind conditions during takeoff. Factors associated with the accident were trees and a va...	2.2826089858
26	The pilot's inadequate compensation for wind conditions during takeoff. Factors associated with the accident were variable winds...	2.2826089858
27	The pilot's inadequate compensation for wind conditions, resulting in an inadvertent ground loop. Factors include wind gusts duri...	2.2729439036
28	The pilot's failure to adequately compensate for wind conditions after encountering a crosswind gust during the landing roll. Fact...	2.2675327857
29	The pilot's inadequate compensation for wind conditions, and the excessive use of the airplane's brakes, which resulted in a nose ...	2.2661909279
30	The pilot's inadequate compensation for wind conditions. Factors associated with the accident are variable winds, and a downdraft.	2.2660439108
31	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control. Contributing factors were...	2.2648282967
32	The pilot's inadequate compensation for wind conditions during takeoff. A factor associated with the accident was a variable wind.	2.2579465664
33	The pilot's inadequate compensation for wind conditions during the landing roll. A crosswind was a factor.	2.2515447851
34	The pilot's inadvertent stall of the airplane during takeoff. Factors associated with the accident were the pilot's inadequate weath...	2.250859567
35	The pilot's inadequate compensation for wind conditions and his failure to maintain directional control during the landing roll. Fac...	2.2497623998
36	the student pilot's inadequate compensation for the crosswind during landing roll. A contributing factor was the cross wind weat...	2.2386981287
37	The pilot's inadequate compensation for wind conditions. A factor associated with the accident was a sudden wind shift.	2.2340296545
38	the pilot's inadequate compensation for the gusting and shifting wind conditions, which resulted in a failure to maintain direction...	2.2275070311

- If we sort the corresponding U matrix on SVD 5 descending, we see the reports with inadequate wind compensation rise to the top

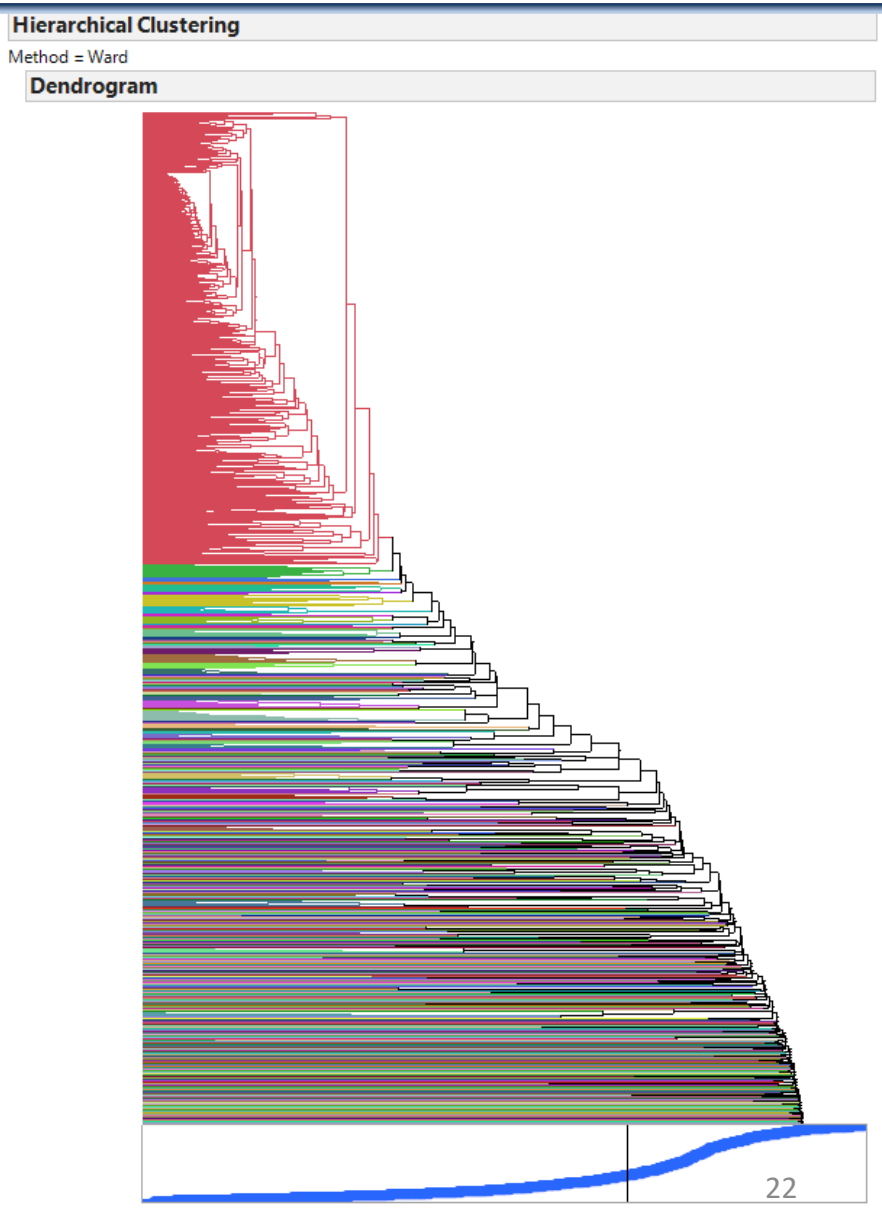
SVD1 vs. SVD2



- Plotting first two eigenvectors is often helpful and a recommended first step

Clustering Terms

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



Clustering Terms

- Here we can see with a large number of clusters, words typically associated with one another are in the same cluster
- The distance matrix allows us to see the closest terms to any specified word

	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

Clustering Accident Reports

- It is possible to cluster the eigenvectors of the U matrix to group accident reports that have common themes
- We can cross-tabulate the clusters with the proportion that were fatal
- These clusters were had about 35 records each and all were non-fatal

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%

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

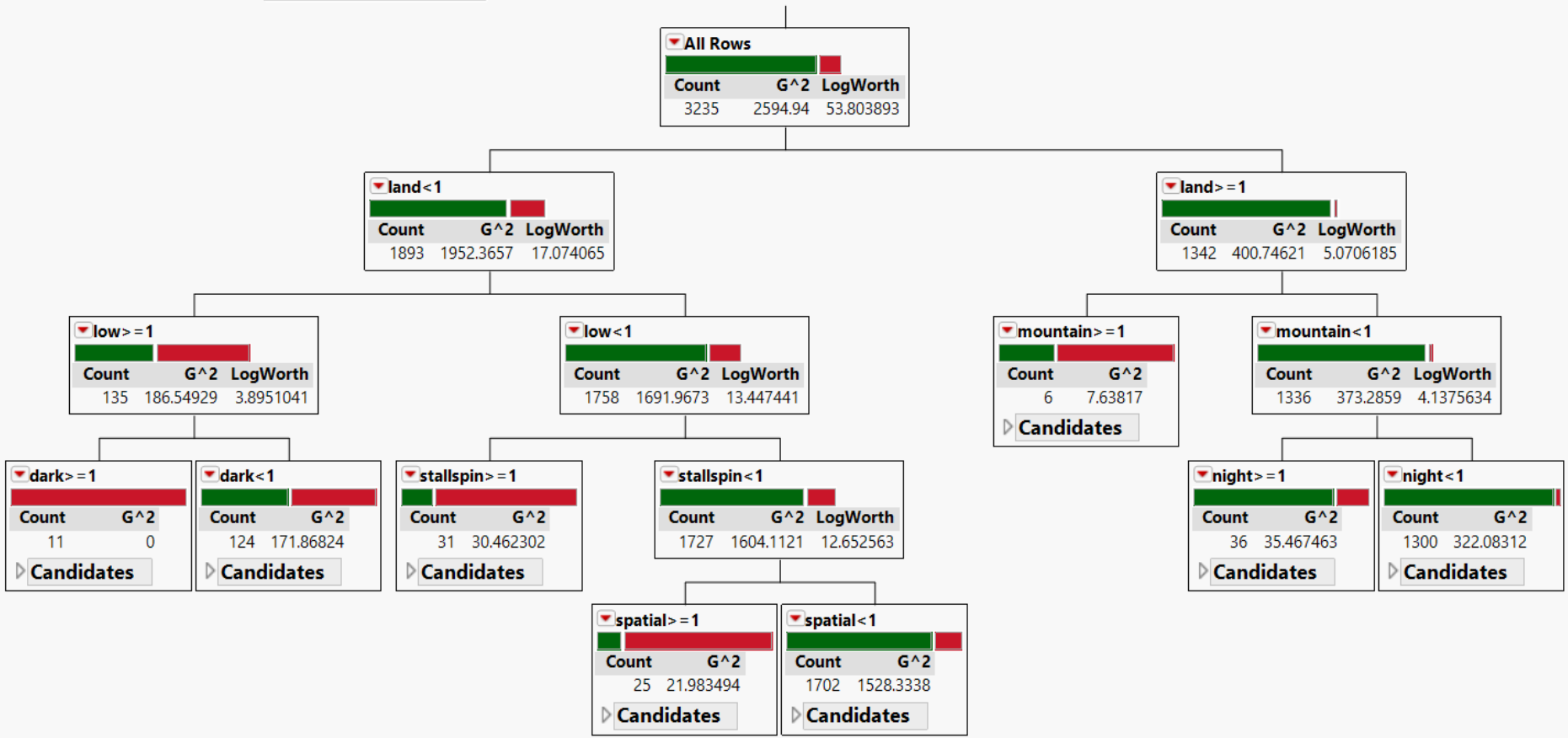
Cluster	fatal	narr_cause
53	YES	The airplane flightcrew's failure to maintain adequate distance/altitude from mountainous terrain during a departure climb to cruise flight, and the captain's impairment from drugs. Factors ir
53	YES	The pilot's inadequate altitude clearance above water while conducting 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/model, and h
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 lack of e
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 stall/sp
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 factors
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 ground
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, the repo
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 in an in
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 factor was th
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 descen
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 subsequent
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 power for ur
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 night light
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 deploy t

Decision Tree for *Fatal* Using DTM

- We can use all 800 columns (each a word) in the Document Term Matrix as input variables to predict whether or not an accident will be fatal
- If land is in the narrative, it will almost surely not be fatal



Most Useful Words to Predict *Fatal*

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

Summary

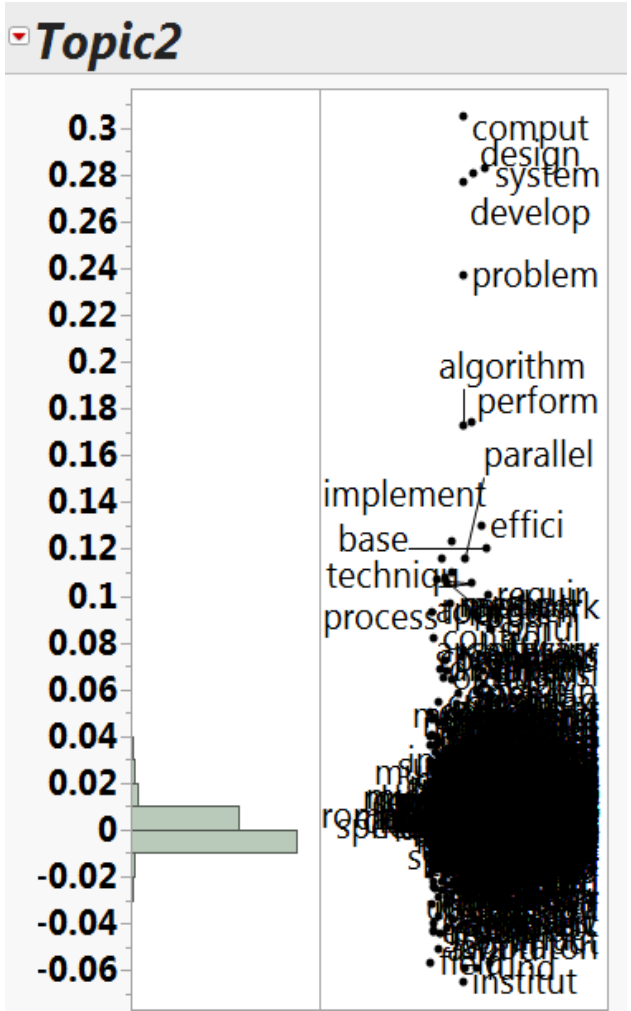
- Data is growing exponentially across DOD
- Much of this is unstructured text data
- Text mining takes statistical tools and converts the text into meaningful mathematical expressions
- Example uses of text mining
  - Concept extraction
  - Grouping like documents or records together
  - Grouping terms together
  - Creating structured variables that represent the text fields to use in predictive analytics
- Relatively short learning curve to perform powerful text analytics using open source software and commercial solutions

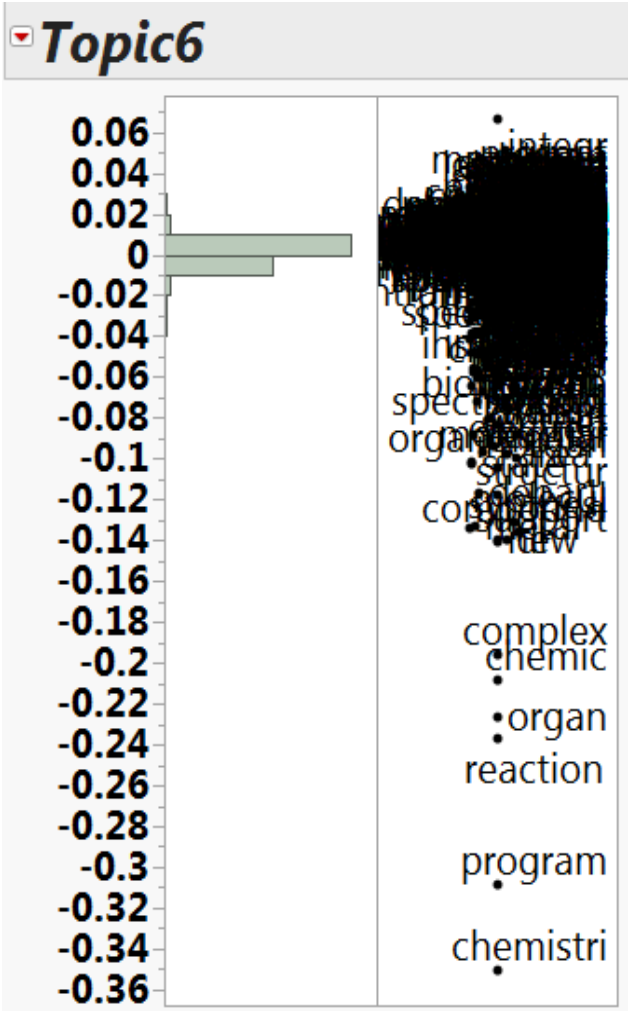
Backup Example

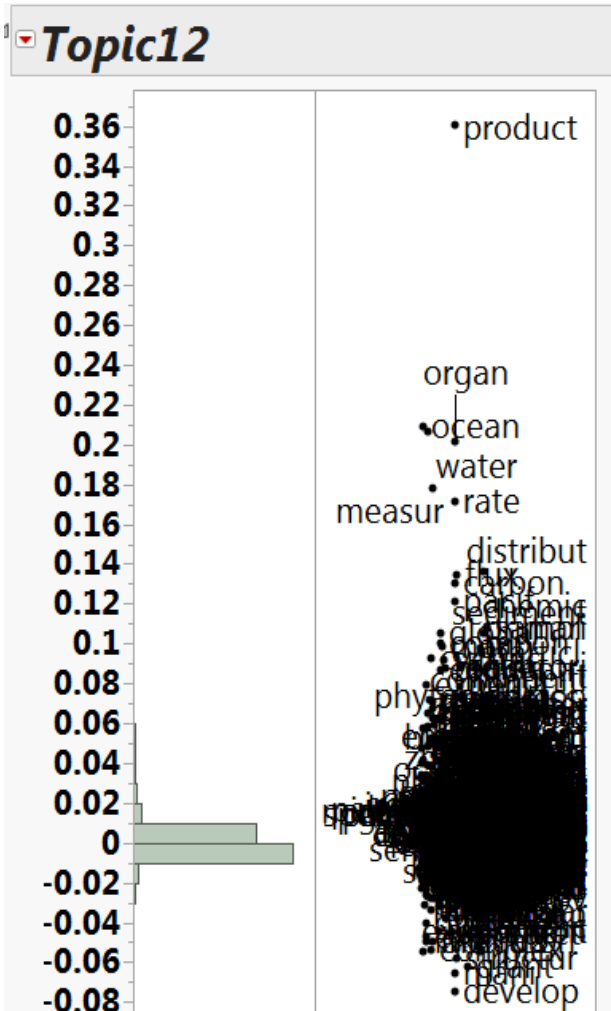
10,000 NSF Abstracts Demonstration

1. What are some common themes?
2. Can we quickly find abstracts within a common theme?
3. Can we place each abstracts in a specific group?
4. Are there certain words that often appear together?

1. Common Themes







2. Finding MISREPS with a Common Theme

e	text
5	The stateof theart control system design methodology for complex nonlinear systems is best described as a black art The traditior
7	The fundamental goal of this research is the development of the science base necessary to reduce the time and effort required to
5	This award provides support to develop computing infrastructure in the areas of parallel languages and compilers automatic paralle
4	High Speed Signal Processing is an important application of special purpose numerical processor designs These digital processing
7	The growing demand for high speed realtime signal and image processing has led to many new parallel architectures Among these
3	This award will support collaborative research between US and French computer scientists The US investigators are Dr CL Liu Dr Ec
0	This infrastructure award is for the construction of a distributed computing facility The network consists of desktop workstations cc
7	This is a renewal of grant IRI9045939 This project aims at the development of neural network architectures to explain how behavin
7	The objective of this investigation is to address software architectural issues which will facilitate the implementation of advanced
5	The PIs currently interested in research in massively parallel disturbed and collective computations artificial neural systems learning
5	This is an infrastructure award to support the acquisition of a high speed network of data servers computation servers parallel proc
3	A generalpurpose tool for computing and visualizing solutions to systems of algebraic equations shall be built Diverse algorithms e
2	This award is to provide research support to Dr Zafiriou under the National Science Foundations Presidential Young Investigator Av
9	Due to their potential for high reliability and throughput via the multiplicity of components distributed computing systems are bein
5	Pattern recognition based control methods can produce stable and robust performance in a range of applications and these meth
.	Research in science is shifting from its traditional focus on the discovery of new information toward the computationally intensive t
3	The selection of combinations of sites at which to locate facilities has long been an important issue within the theoretical domain
3	A broad range of chemical process engineering problems including systematic process design retrofit design batch scheduling pro
4	This award provides support for a two year research project between Professor Edmund Clarke School of Computer Science Carne
.	This research focuses on performance issues in rulebased knowledge bases that have to manage large amounts of rules and data

3. Clustering MISREPs into Mutually Exclusive Bins

text	Stop	Cluster
Agarwal The goal of the Alewife experiment is to demonstrate that a parallel computer system can be made both sc		43
The research project combines work on graph embeddings and parallel algorithms to study a variety of emulation pr		43
The purpose of this research project is to attack the issues of programming and performance evaluation of multipro		43
The question of how processor scheduling should be done in a parallel timesharing environment will be investigate		43
Two areas which need further enhancement in the parallel computing effort of the National Center for Supercompu		43
Although standard Unix permits applications to be piped together the standard syntax has functional and practical		43
As the speed and memory size of supercomputers increase the need for file storage space also grows Simple scalin		43
IPS2 is a performance measurement system for parallel and distributed programs It allows application programmer		43
One important aim of Computational Neural Science which studies information processing brains and adopts neura		43
This award supports continued research into programming language support for parallel scientific computation on		43
This award provides support to develop computing infrastructure in the areas of parallel languages and compilers au		43
This project uses tools and methods of statistical mechanics to introduce a unified computational approach to analy		43
This award provides support for a two year research project between Professor Edmund Clarke School of Computer		43
The long term goal of this project is to investigate the relationship between the structure of parallel algorithms and		43
This award will support the purchase of the 570node Intel Delta Touchstone System This will be the worlds fastest cc		43
This project continues research on shared virtual memory for multicomputers and further research on fault tolerant		43
With the increasing availability of message based distributed parallel computing systems the difficulties associated v		43
The award to SUNY Stony Brook is for the acquisition of a MIMD computer to support research in a number of diffe		43
This research revolves around the acquisition of a Single Instruction Multiple Data SIMD computer and its applicatio		43

4. Words That Group Together

	Label	Cluster
8	confer	247
9	abstract	246
10	hous	246
11	presid	246
12	white	246

	Label	Cluster
36	parallel	235
37	comput	234
38	problem	233
39	algorithm	232

	Label	Cluster
138	protein	189
139	heat	188
140	transfer	188
141	grant	187
142	codata	186
143	columbus	186
144	juli	186
145	ohio	186

	Label	Cluster
326	composit	114
327	molecular	113
328	biolog	112
329	genet	111
330	popul	111
331	dna	110
332	sequenc	110
333	speci	109
334	among	108
335	evolut	108
336	evolutionari	108
337	famili	108
338	genus	108
339	morpholog	108
340	phylogenet	108
341	relationship	108

Text Mining Tasks

- Collect text to form corpus
  - Text fields from existing xlsx or jmp columns
  - Assemble from directory of disparate file types (pptx, pdf, doc, txt)
  - Scrape from websites and social media (Twitter, FB)
- Clean text
  - Correct misspellings
  - Synonyms
- Parse text and string processing
  - Change to lower case, remove punctuation
  - Select tokenization method
  - Stem words
  - Parts of speech tagging
- Filter words
  - Create stop word list or start word list
  - Identify multi-words (operations research) and multiple word phrases of interest (pilot failure to properly)

Text Mining Tasks

- Investigate word frequencies
  - Word count Pareto diagram and word clouds
  - Quickly identify documents containing specific words
- Form Document Term Matrix
  - Sparse matrix with documents as rows and columns are words
  - Identify proper weighting scheme (binary, term frequency, inverse document frequency…)
- Determine correlations between words
- Reduce dimensionality of DTM
  - Perform Singular Value Decomposition or Principal Components
  - Create two matrices U-document and V-term
- Find topics and themes
  - Plot and interpret rotated latent vectors of V matrix
  - Form clusters of like terms from the V matrix
  - Use Latent Class Analysis methods
- Discover words that often appear together and link concepts
  - Form clusters from the V matrix
  - Find terms closest to specific words with distance matrix from clustering

Text Mining Tasks

- Identify documents with common theme
  - Sort U matrix based on corresponding latent dimension
  - Perform cluster analysis on the U matrix
- Find and subset documents containing specific words, phrases, or themes
- Perform sentiment analysis
  - Tally word counts in each document against dictionary of positive/negative terms
- Score a response variable by word occurrence
- Conduct association analysis (market basket analysis) to frequent word sets
- Combine text data with structured data for predictive analytics
- Classify new documents into existing groups or clusters