ADSURGO
ascend with analytics

Harness the Power of Text Mining: Unstructured Data Analysis for Quality and Productivity Applications

June 10, 2015

W. Heath Rushing
Heath.Rushing@adsurgo.com

Outline

- Introduction
- Demonstration: Recent article on cnn.com
- Text Mining using product recall data
- Demonstrations:
  - Customer Survey with SAS TM
  - NTSB with R
  - FDA Citations with JMP/R script
- References

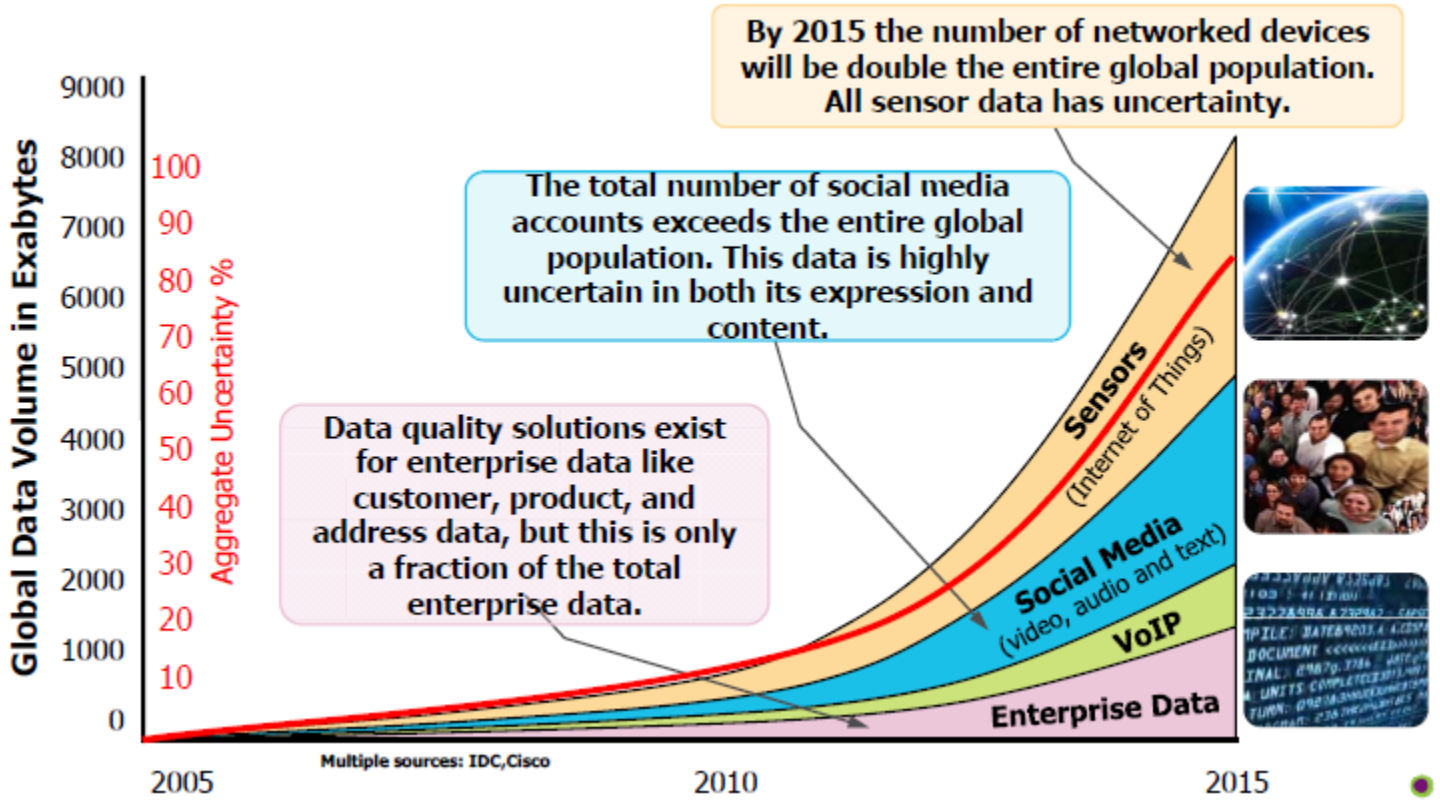
What is Text Mining?

- Text mining: semi-automated process of detecting patterns (useful information and knowledge) from large amounts of *unstructured* data sources
- 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, etc.

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

The Nature of Unstructured Data

- As much as 80% of all data is unstructured but still has exploitable information available



IBM Research India, presented at Text Mining Workshop Jan 2014

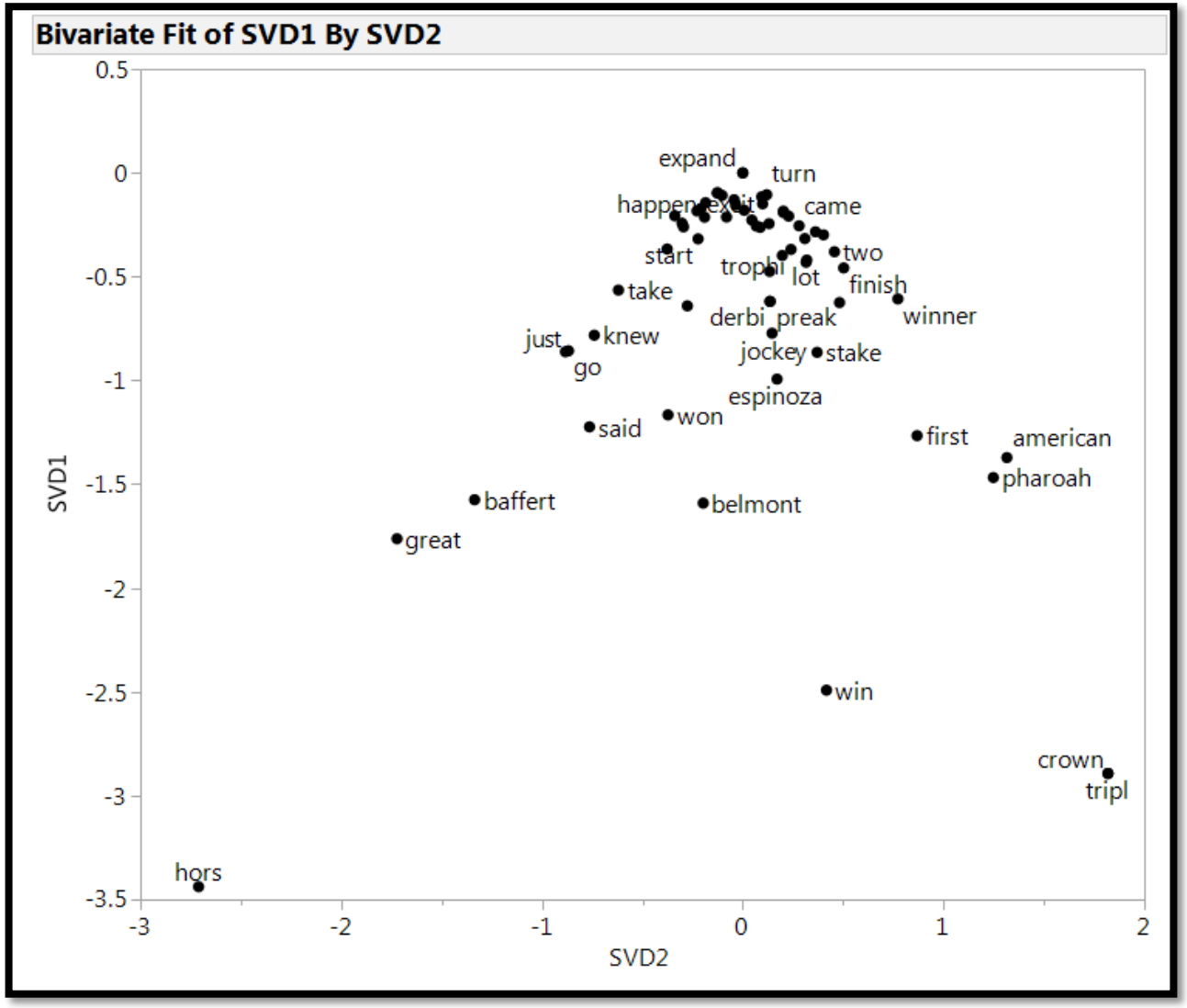
Warm Up

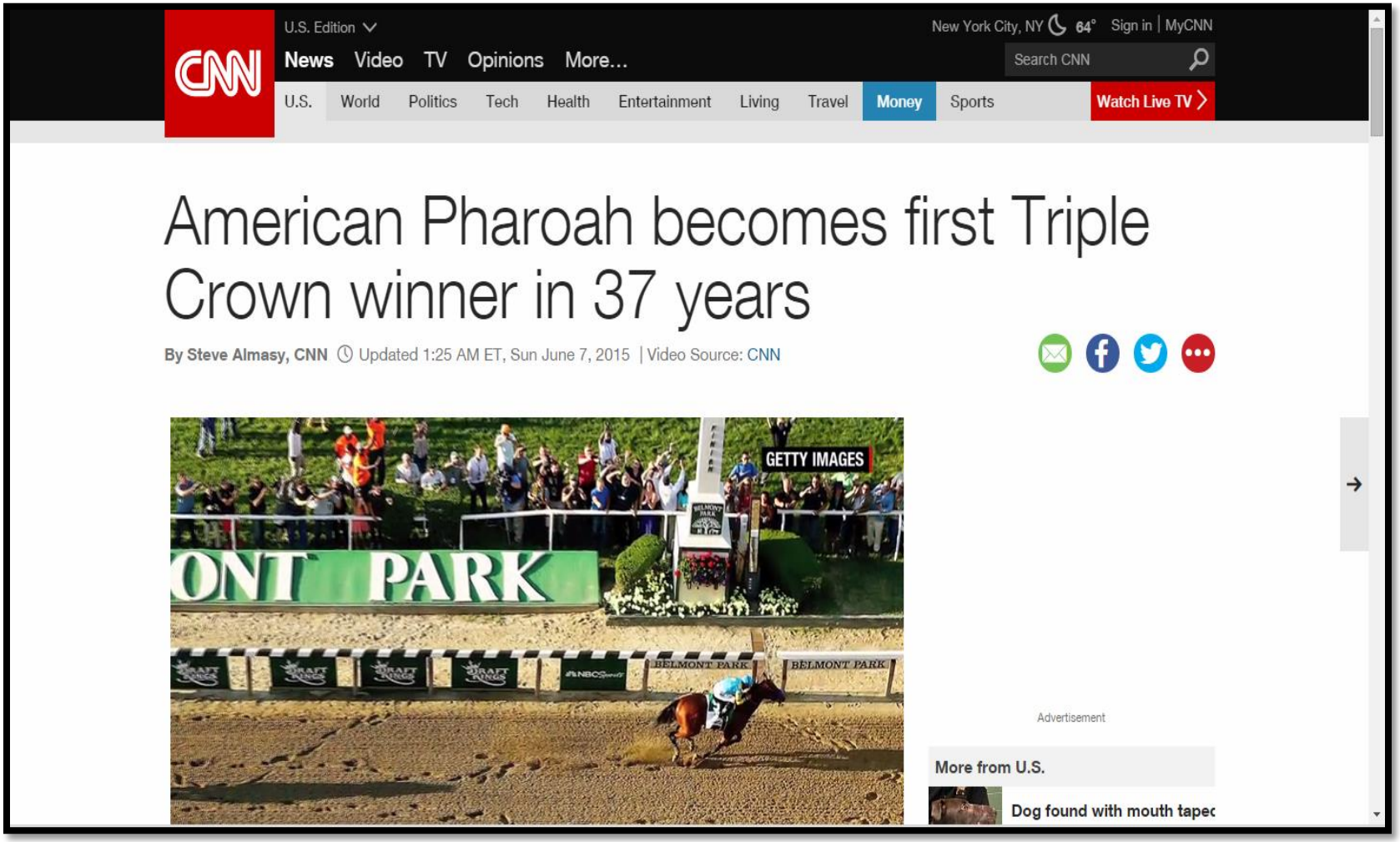
- Consider the word counts of a recent article on cnn.com
- Can you get the general idea of what might have happened by these frequencies alone?

Level	Count
hors	16
crown	14
tripl	14
pharoah	10
win	10
american	9
belmont	8
baffert	7
first	7
said	7
won	6
espinoza	5
just	5
race	5

Warm Up…WordCloud

Warm Up… More Insight





http://www.cnn.com/2015/06/06/us/belmont-stakes-american-pharoah/

TEXT MINING

Extracting Numerical Representations of Text

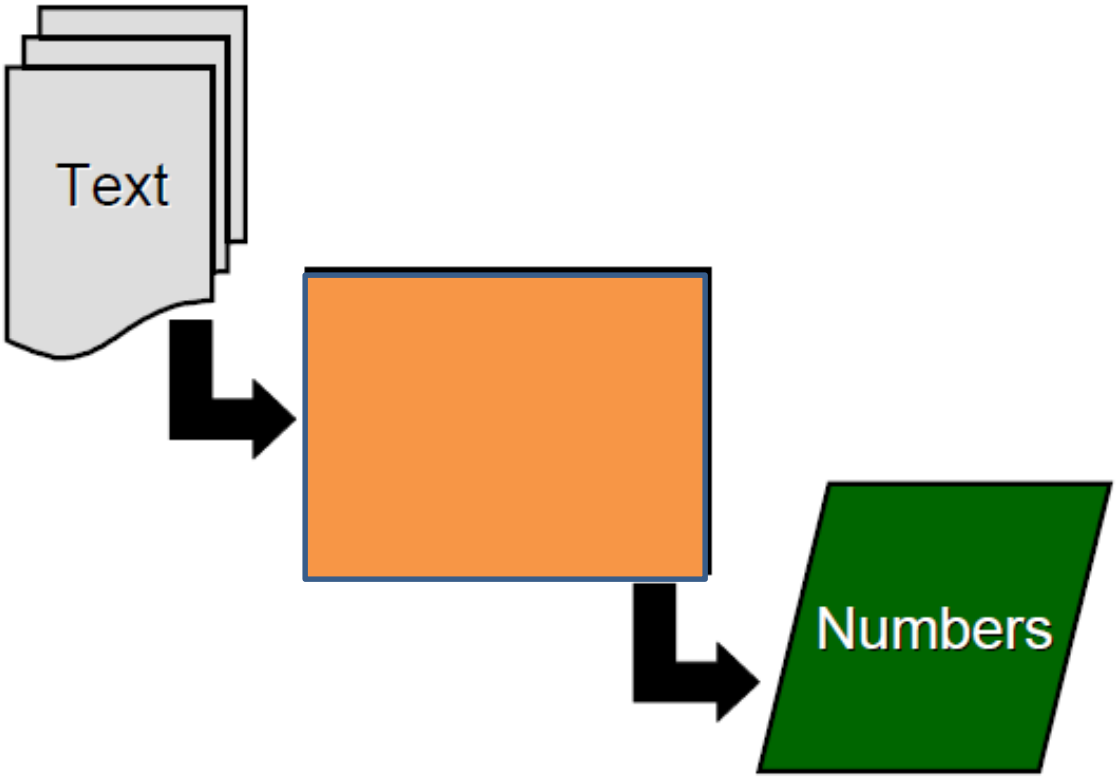
- In order to analyze text in a systematic and structured way, we first need to develop a numerical representation of the text.
- Obviously, there is not a unique solution to this problem. The appropriate mapping of text->numbers depends on the goal of the study.

Possibilities

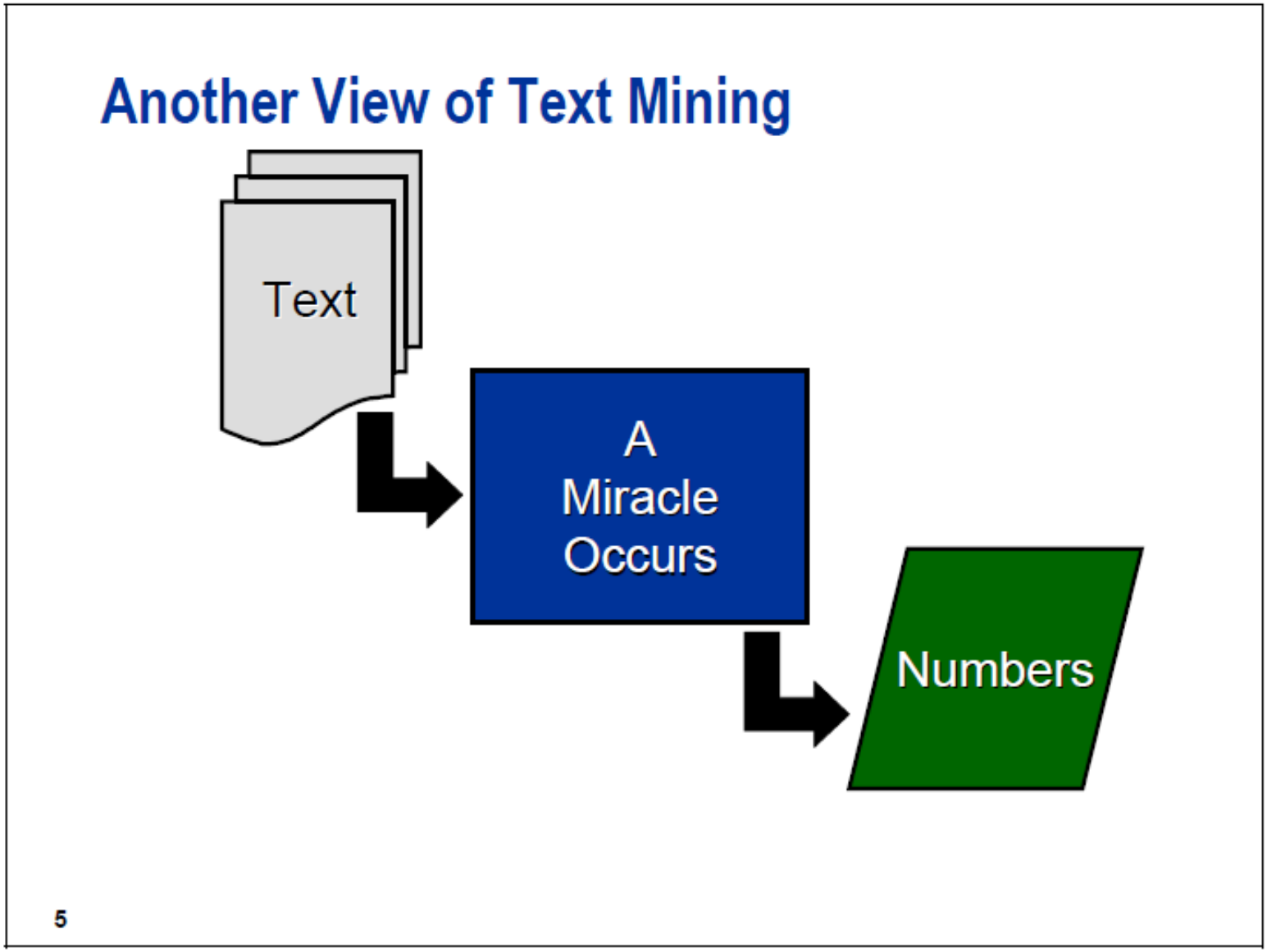
- If we could represent text with numerical indices, we could use those indices as input to
  - Supervised learning methods (target variable)
    - Linear and logistic regression
    - Classification and Regression Trees (CART)
  - Unsupervised methods (no target variable)
    - Hierarchical Clustering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

Numerical Representations of Text

Another View of Text Mining



Numerical Representations of Text



DSURGO
ascend with analytics

STRING PROCESSING

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.

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.

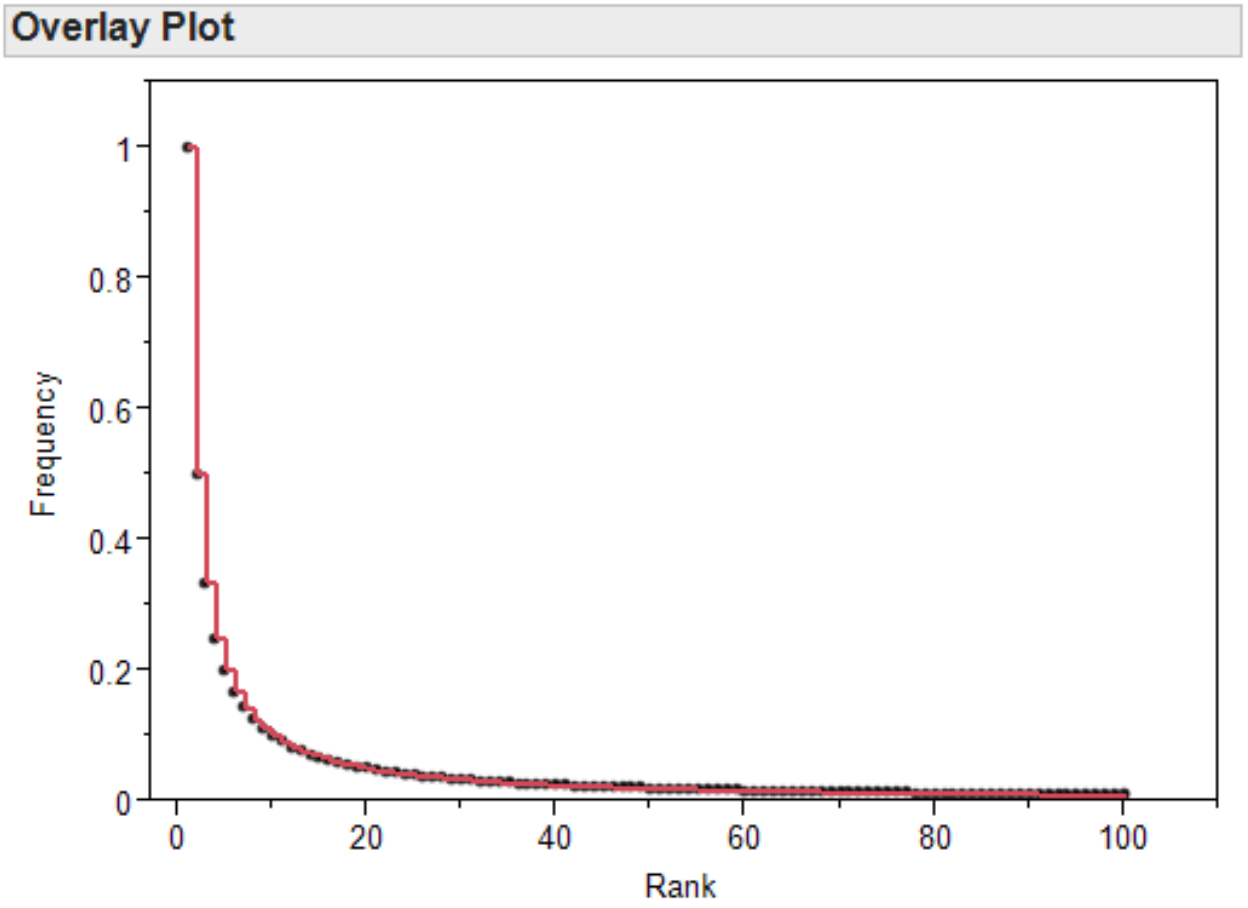
Processing Text

- 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.

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.



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).

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

	Year	title	advers	fda	patient	product
1	2001-2005 FDA Recalls	Baxter MeridianÂ® Hemodialyis (HD) Instrument	0	2	1	2
2	2001-2005 FDA Recalls	Baxter Healthcare Corporation COLLEAGUEÂ® ...	0	0	2	2
3	2001-2005 FDA Recalls	bioMÃ©rieux, Inc. VeriCalÂ® Calibrator Sets	0	1	2	3
4	2001-2005 FDA Recalls	Baxter Healthcare Corp. COLLEAGUEÂ® Volum...	0	0	3	1
5	2001-2005 FDA Recalls	Guidant Corp. Pacemakers	0	2	0	2
6	2001-2005 FDA Recalls	CONTAK RENEWAL and CONTAK RENEWAL 2 ...	0	1	1	1
7	2001-2005 FDA Recalls	Ventak PRIZM 2 DR ICDs	0	1	1	1
8	2001-2005 FDA Recalls	Boston Scientific HemashieldÂ® VANTAGEâ „	0	0	3	1
9	2001-2005 FDA Recalls	Vail Products, Inc. Enclosed Bed Systems	0	2	12	2

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!

tf-idf

6/0 Cols	Year	title	advers	fda	patient	product
1	2001-2005 FDA Recalls	Baxter MeridianÂ® Hemodialyis (HD) Instrument	0	0.0017430038	0.0024847562	0.0004702839
2	2001-2005 FDA Recalls	Baxter Healthcare Corporation COLLEAGUEÂ® ...	0	0	0.0057527507	0.0005444047
3	2001-2005 FDA Recalls	bioMÃ©rieux, Inc. VeriCalÂ® Calibrator Sets	0	0.0010984018	0.0062633499	0.0008890869
4	2001-2005 FDA Recalls	Baxter Healthcare Corp. COLLEAGUEÂ® Volum...	0	0	0.0100491089	0.0003169951
5	2001-2005 FDA Recalls	Guidant Corp. Pacemakers	0	0.001953999	0	0.000527213
6	2001-2005 FDA Recalls	CONTAK RENEWAL and CONTAK RENEWAL 2 ...	0	0.0014850392	0.0042340245	0.0004006819
7	2001-2005 FDA Recalls	Ventak PRIZM 2 DR ICDs	0	0.0016002578	0.0045625264	0.0004317692
8	2001-2005 FDA Recalls	Boston Scientific HemashieldÂ® VANTAGEâ „	0	0	0.0128045097	0.0004039132
9	2001-2005 FDA Recalls	Vail Products, Inc. Enclosed Bed Systems	0	0.0015933897	0.0272576687	0.0004299161
10	2001-2005 FDA Recalls	Abbott Blood Glucose Meters	0	0.000655936	0.0018701522	0.0001769796
11	2001-2005 FDA Recalls	Laerdal Medical Corp. CM 100 HeartstartÂ® Au...	0	0.0010034049	0.0057216548	0.0008121929
12	2001-2005 FDA Recalls	Welch Allyn Co., AED 20Â®, Automatic External...	0	0.0011529808	0.0098618584	0.0006221768
13	2001-2005 FDA Recalls	LifeScan, Inc. Blood Glucose Meters	0.001...	0.0007639091	0.0087119847	0.0008244483
14	2001-2005 FDA Recalls	BioMerieux Simplastin HTF Tissue Reagent	0	0.0009470913	0.010801083	0.0015332214
15	2001-2005 FDA Recalls	HeartSine Technologies, Inc. Automatic Externa...	0	0.0010792436	0.0061541054	0.0011647728
16	2001-2005 FDA Recalls	MicroScanÂ® Rapid Pos Inoculum Broth	0	0.0011182524	0.0031882715	0.0015085913
17	2001-2005 FDA Recalls	Certain Medtronic LIFEPAK 500 Automated Ext...	0	0.0012802062	0.0073000423	0.0006908308
18	2001-2005 FDA Recalls	IV Flush Heparin and Saline IV Catheter Flushes	0	0.0064754617	0	0.0005823864
19	2001-2005 FDA Recalls	Becton Dickinson ProbeTec ET Urine Processing...	0	0.0026330483	0.0037535679	0.0010656432
20	2001-2005 FDA Recalls	BioMed Unomedical Airway Adapter	0	0.0032005156	0	0.0034541539

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

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

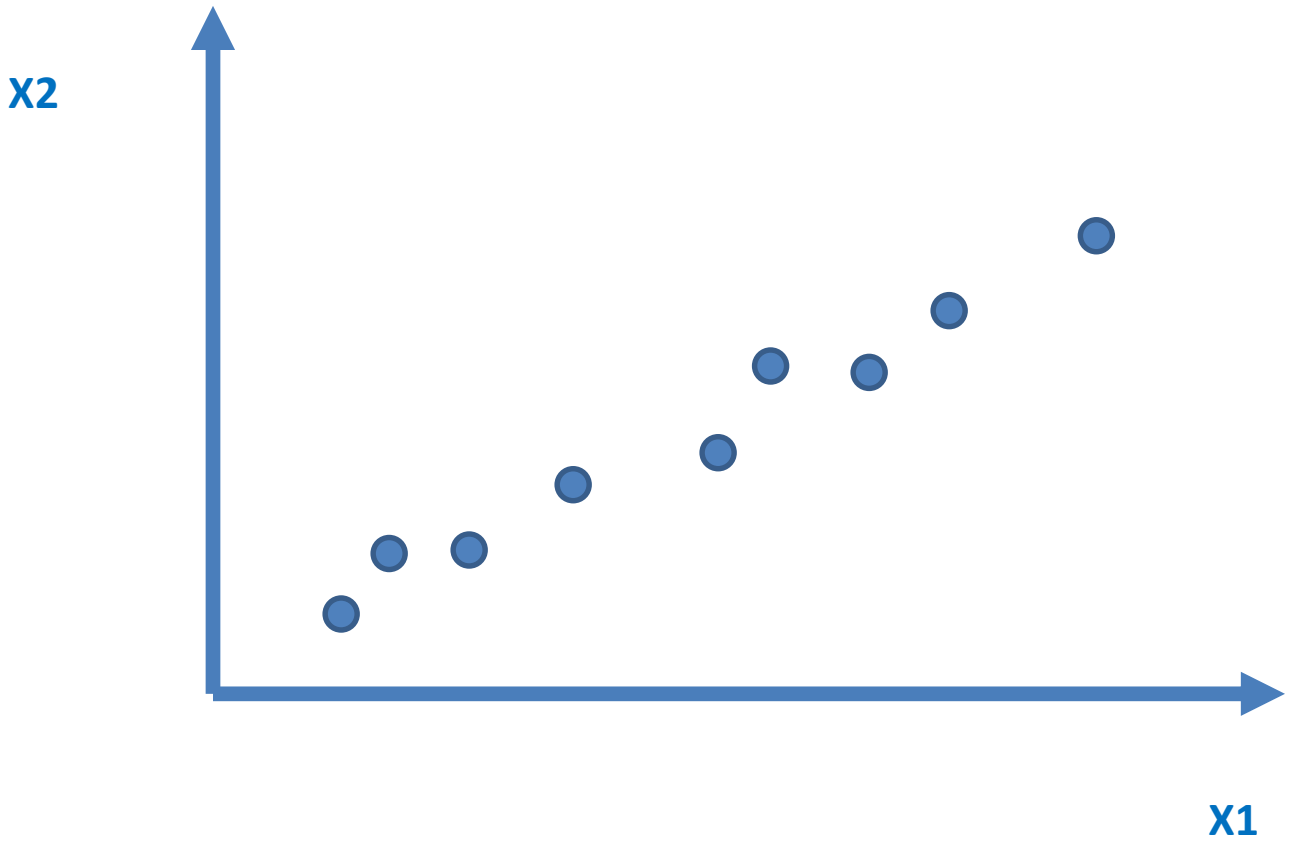
advers	advic	advis	advisori	advocaci	aedcardiacsci	aeruginosa	affair	affect	affix	age	agenc	agent	agre	aid
0	0	0	0	0	0	0	0	0.0028809636	0.0293507939	0	0	0	0	0
0	0	0.015...	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.034...	0.035...	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0096891356	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.011...	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.015...
0	0	0	0	0	0	0	0	0.0033170014	0.0337930762	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0038114612	0	0	0	0	0	0
0.001...	0	0.011...	0	0	0	0	0	0.0050505782	0	0	0	0	0	0.018...
0	0	0	0	0	0	0	0	0.0031308431	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0142708199	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0073933163	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0084640725	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.019...	0	0	0	0	0	0	0	0	0	0	0	0.031...
0	0	0	0	0	0	0	0	0.0158701359	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0039847095	0	0	0	0	0	0
0	0	0.022...	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0025892205	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0035065443	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.0447...	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0041462517	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0045455204	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0.0035065443	0	0	0	0	0	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.

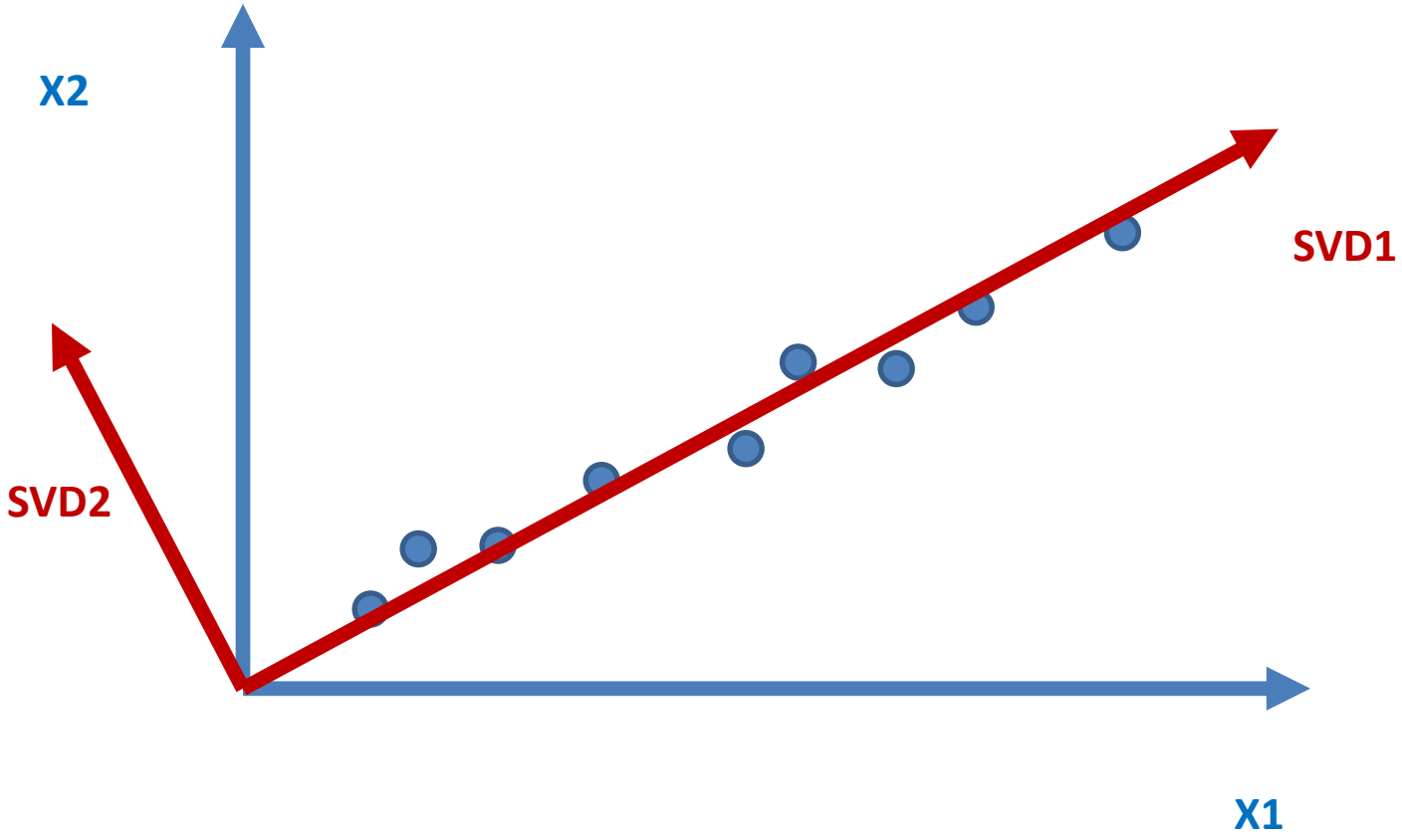
SVD Example

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

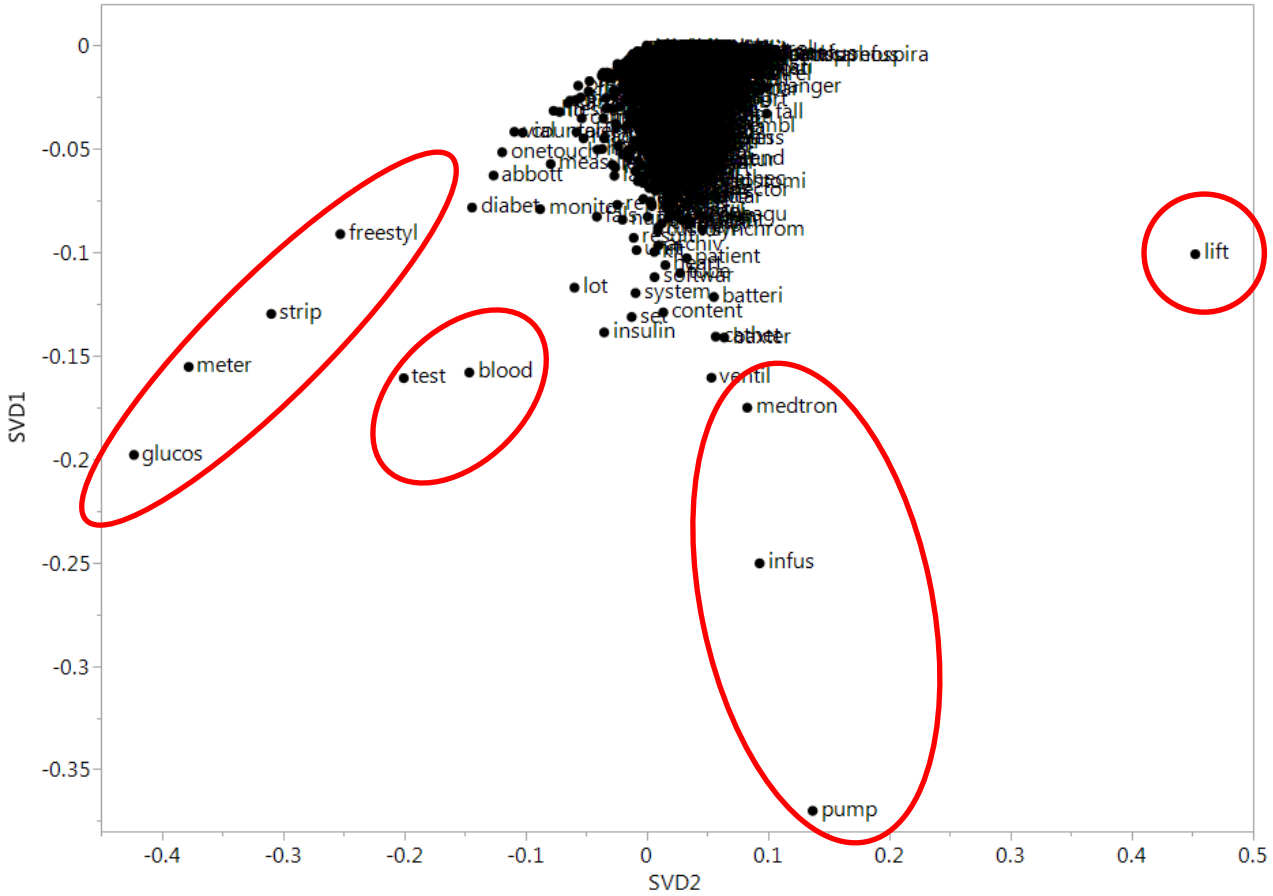


SVD Example

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.



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).

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?

APPLICATIONS OF TEXT MINING

Survey Analysis Example – Open-ended Questions

- Data: 315 respondents to a survey by a company.
- Objective: Use text mining and **word clustering** techniques to find evidence of potential fraud cases in a set of comments from open-ended survey responses.
  - 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 Text Miner

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

NTSB Aircraft Accident Reports

- Data: NTSB Aircraft Accident Reports.
- Objective: Use **document clustering** techniques and **CART** to determine what factors contributed to fatal accidents.
- Software used: R statistical software program.

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

FDA 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.

OPTIONAL

Social Media - Twitter

- Data: Live Twitter data
- Objective: Determine social media reaction to a current event.

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