Text Mining and Analysis in Railroad Safety

BRIDGETTE FINDLEY

Text Mining is…

- The derivation of information from unstructured data by automated systems, specifically text documents
- Determines patterns within the text in order to discover a word or phrase's relevance to the topic.

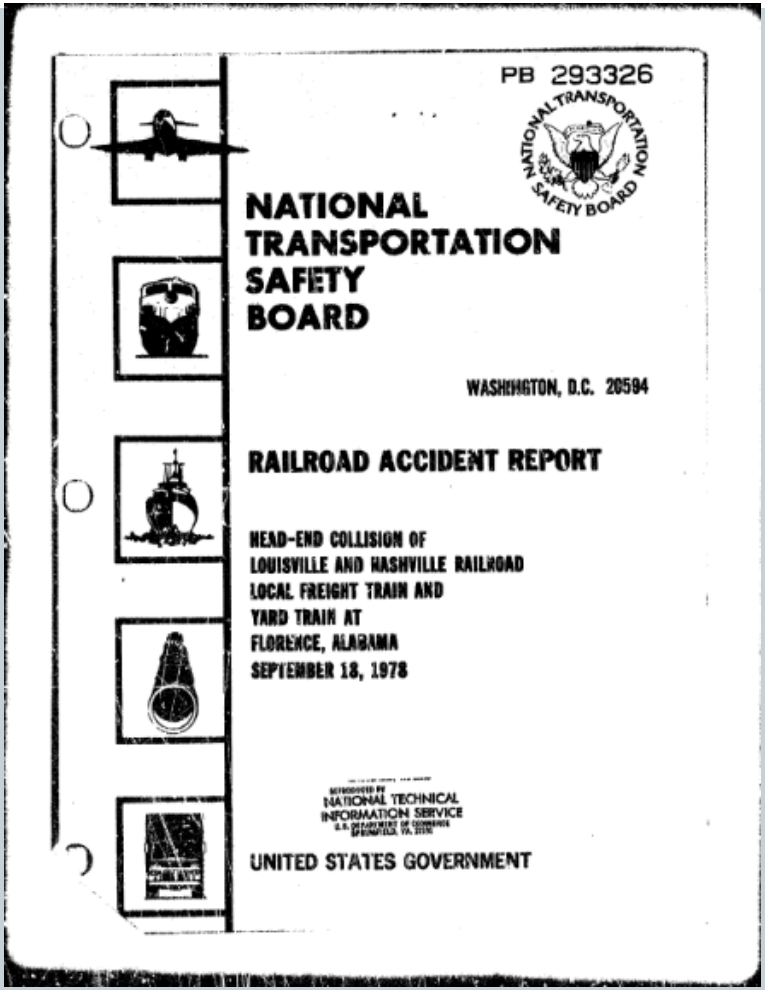
Applications

- Criminal Network Analysis
- Identifying Potential Causes of Medical Conditions
- Cost Analysis
- Identifying Common Causes of Accidents at Grade Crossings

What trends can be found in factors of major railroad accidents?

The Data

- National Transportation Safety Board has made all accident reports available since 1996
- PDFs → Raw Text
- Processing Could Yield
  - Common Causes of Deaths
  - Common System Failures
  - Differences in Accidents based on Region


PB 293326

NATIONAL TRANSPORTATION SAFETY BOARD

WASHINGTON, D.C. 20594

RAILROAD ACCIDENT REPORT

HEAD-END COLLISION OF
LOUISVILLE AND NASHVILLE RAILROAD
LOCAL FREIGHT TRAIN AND
YARD TRAIN AT
FLORENCE, ALABAMA
SEPTEMBER 18, 1978

NATIONAL TECHNICAL INFORMATION SERVICE

UNITED STATES GOVERNMENT

Techniques

- Latent Dirichlet Analysis
- Latent Semantic Indexing
- Clustering Analysis

Latent Dirichlet Analysis

- Returns the permeation of various words within the data collection, suggesting the most important topics

Table 2
Word stems associated with projects with large cost overruns.

Word	In documents	Total	High overrun project	Near low bid project	Under run
excav_asphalt	18	18	14	3	1
drainag	19	19	14	3	2
mobil_time	14	14	11	2	1
replac_bridg	14	14	10	4	0
cast_steel	7	10	8	2	0
steel_concret	7	10	8	2	0
overhead_structur	8	8	6	2	0
realign_roadwai	8	8	6	2	0
roadwai_bridg	8	8	6	2	0
temporari_signal	8	8	6	2	0
epoxi_coat	5	5	5	0	0
geosynthet	5	5	5	0	0
geosynthet_reinforc	5	5	5	0	0
reinforc_embank	4	5	5	0	0
temporari_fenc	4	5	5	0	0

Latent Semantic Indexing

- Patterns and association between words gives value to certain phrases more than others
- Used to derive ‘meaning’
- Comparable to human cognition

Clustering Analysis

- A numeric set of values is assigned to a subject based on defining factors as determined by text mining
- Subjects are grouped based on these values, denoting similarity or dissimilarity

References

- Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
- Chen, H., Chung, W., Wang, G., Qin, Y., & Chau, M. (2004, April 1). Crime Data Mining: A General Framework and Some Examples. Computer, 50-56.
- Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. Communications of the ACM, 49(9), 77-82. Retrieved June 3, 2015.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). Berkeley, California.
- Kanungo, T. (2002). An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions, 24(7), 881-892. doi:10.1109
- Williams, T., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. Automation in Construction, 43, 23-29. doi:10.1016
- Williams, T., Nelson, C., & Betak, J. (2015). Applying Topic Modeling to Railroad Grade Crossing Accident Report Text (pp. 1-5). CCICADA.