

Text Mining – A Brief Summary

Amanda Hepler

What is text mining?

A common technique to extract useful information from large data documents is known as **data mining**. SAS has developed software, [Enterprise Miner](#), that performs this type of task. Recent upgrades have been made in 2002 and now this software is able to apply this mining technique to text based files. This process is known as **text mining**.

What is the purpose of text mining?

- To discover and use knowledge that is contained in a document collection as a whole, extracting essential information from document collections and from a variety of different sources (- James Cox, manager of text mining development at SAS)
- Text mining lets executives ask questions of their text-based resources, quickly extract information and find answers they never imagined.

Three Steps to Text Mining:

1. "Preprocessing" the text to distill the documents into a structured format.
2. Reducing the results into a more practical size.
3. Mining the reduced data with traditional data mining techniques.

Technical Details:

Text preprocessing transforms text into an information-rich, term-by-document matrix. This large grid indicates the frequency of every term within the document collection. During this stage, feature extraction is also used to locate specific bits of information, such as customer names, organizations and addresses.

Next, a mathematical technique called singular value decomposition (SVD) is used to replace the original term-by-document matrix with a much smaller matrix. As part of this process, unimportant words get discarded or ignored, and more important or highly relevant words are singled out. The new matrix can be used to place associated terms and documents into categories.

Lastly, clustering, classification and predictive methods are applied to the reduced data, using traditional data mining techniques. Conventional structured data sources can also be included in the analysis to enrich the discovery of underlying trends and patterns within the data.

Current Uses / Examples of Text Mining:

- Sales and marketing executives at *Compaq Computer Corp.* count on text mining tools to analyze company descriptions in their prospect database. The results help executives target customers for new sales and marketing campaigns.
- Linguists at the *Université Catholique de Louvain* in Belgium use text mining to analyze summaries of ancient and modern texts. Researchers mine textual information in several languages and use the results to address philological and psychological questions.
- A new text mining project at the *University of Louisville Medical Center* will let doctors make better use of medical databases such as Medline, PsychInfo and Toxline for evidence-based medicine. Search results of these medical databases can often yield 2,000 matches, but advanced modeling with Enterprise Miner can reduce the results to 100 highly relevant documents and sort those 100 documents into smaller subgroups or categories.