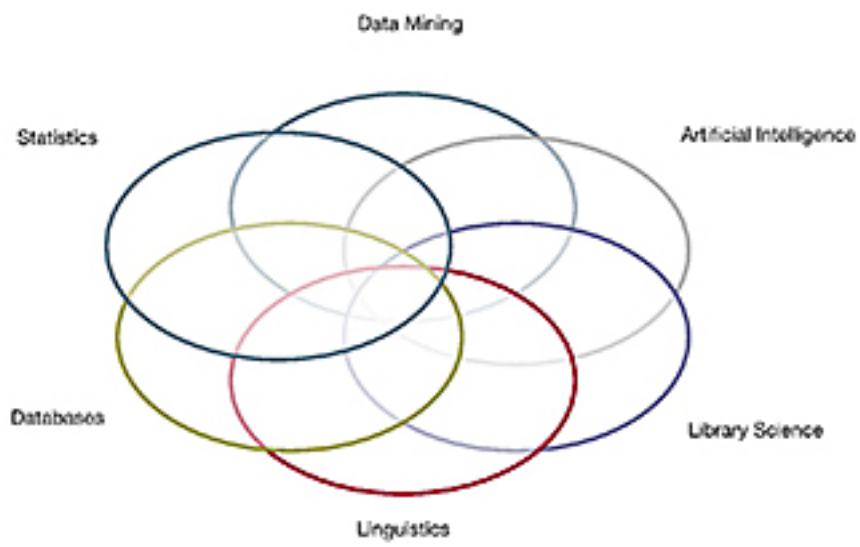
Text Mining: The Next Data Frontier

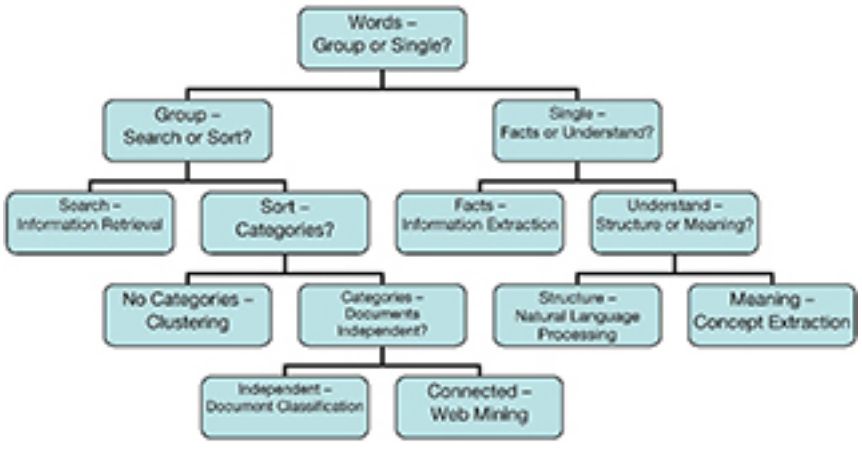
Mark A. Anawis

By some estimates, 80 percent of available information occurs as free-form text



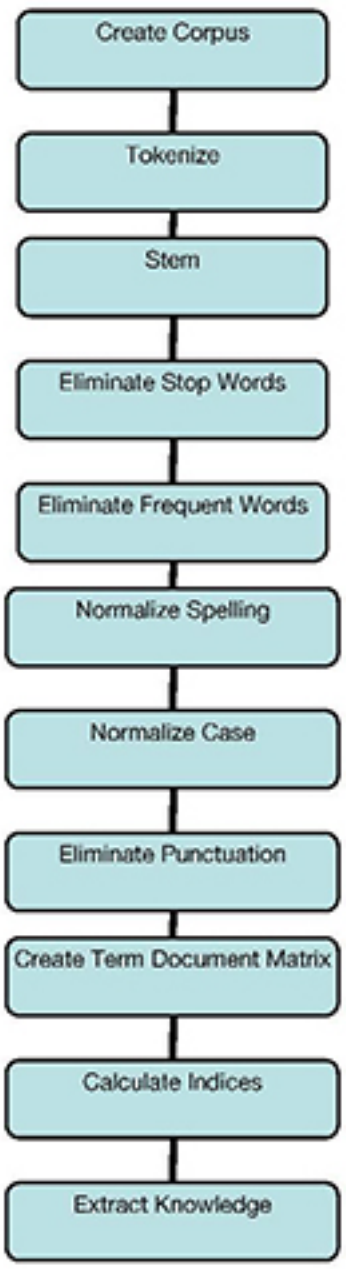
Josiah Stamp said: “The individual source of the statistics may easily be the weakest link.” Nowhere is this more true than in the new field of text mining, given the wide variety of textual information. By some estimates, 80 percent of the information available occurs as free-form text which, prior to the development of text mining, needed to be read in its entirety in order for information to be obtained from it. It has been applied to spam filters, fraud detection, sentiment analysis, identification of trends and authorship.

Text mining can be defined as the analysis of semi-structured or unstructured text data. The goal is to turn text information into numbers so that data mining algorithms can be applied. It arose from the related fields of data mining, artificial intelligence, statistics, databases, library science, and linguistics (Figure 1).



There are seven specialties within text mining that have different objectives. These can be decided by answers to the questions shown in the decision tree in Figure 2. These specialties are:

1. Information retrieval: storage and retrieval of documents
2. Document Clustering: group and categorize documents using data mining clustering algorithms
3. Document Classification: group and categorize documents based on labeled examples
4. Web mining: understand relationships of hyper linkages of documents on the web
5. Information Extraction: identify specific facts and relationships of unstructured text
6. Natural language processing: understanding language structure, such as parts of speech
7. Concept extraction: group words into similar semantic groups



Once the objective (specialty) has been determined, the basic methodology can be applied (Figure 3). It can be summarized in the following scheme:
1. Create the corpus by collecting the documents. This can be manual or automatic, such as via a Web crawler or database query. These are organized into a similar format.
2. Clean up the text according to the following steps:

a. Tokenization: Fragment the text into items that can be counted.
b. Stemming: Identify common core word fragments (e.g. “result,” “results,” “resulting,” “resulted” all become “result.”
c. Eliminate stop words: Create a dictionary of low predictive value words (e.g. a, an, the).
d. Eliminate frequent words: Create a dictionary of common, repetitive words often found as a group (e.g. “For information only”).
e. Normalize spelling: Use or create a dictionary to correct misspellings using fuzzy matching.
f. Case normalization: Convert all text to lower case to prevent counting words with different capitalization separately.
g. Eliminate punctuation: Remove punctuation to prevent counting words with and without punctuation being counted separately.

3. Reduce dimensionality and select features.

a. Term Document Matrix (TDM): Create a two dimensional matrix where each document is one row and each column is a term from the abbreviated list generated by cleaning up the text. The relationship between the row and column is represented by indices. Singular Value Decomposition (SVD) is used to expose the underlying meaning and structure by reducing the dimensionality. It is related to principal components analysis.
b. Indices: At the simplest level, this can be the count, or number of times a term appears in a document. Log or binary frequencies can be used to dampen large number of occurrences. The most commonly used index is the inverse document frequency (ITF). It represents the relative importance of a term and reflects the relative frequency of occurrence of terms and their document frequencies.

4. Extract knowledge (examples of methods)

a. Classification: Assign terms into a predetermined set of categories. A training data set is used with documents and categories. This is used in genre detection, spam filtering and Web page categorization.
b. Clustering: Terms are placed into natural occurring but meaningful groups. This is used in document retrieval to enable improved Web searches and analysis of large text collections.
c. Association: Identify terms that are frequently found together. This is known as market basket analysis in the retail industry where items are bought together.
d. Trend analysis: Identify time dependent changes in a term. This is used in identifying rising popularity of technologies.

There are numerous vendors supplying sophisticated text mining packages, such as SAS Text Miner, IBM SPSS Modeler and Statistica Text Miner. There also are free and open-source tools available, such as R and RapidMiner.

Due to the rapidly expanding amount of text-based information, text mining has already been applied regularly in the areas of spam filters, fraud detection, sentiment analysis, identification of trends and authorship. Future challenges will be applications to large corpora, domain knowledge, personalization, multi-lingual capabilities, and maintenance of dictionaries as area-specific language evolves.

References
1. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, by John Elder, Gary Miner, and Bob Nisbet,

Academic Press, 2012
2. Text Mining Handbook, Louise Francis, FCAS, MAAA, and Matt Flynn, PhD, Casualty Actuarial Society, E-Forum, Spring, 2010
3. An Introduction to Text Mining in R, Ingo Feirer, R News, Vol. 8/2, p. 19–22, Oct. 2008

Mark Anawis is a Principal Scientist and ASQ Six Sigma Black Belt at Abbott. He may be reached at editor@SceintificComputing.com.

Source URL (retrieved on *11/28/2015 - 11:36am*):
http://www.scientificcomputing.com/blogs/2014/01/text-mining-next-data-frontier