as single terms and entries in the index. These lists can be edited and saved for future and repetitive use, so the system can be customized to specific terminologies for different domains.

**Stemming and Support for Different Languages**

Stemming refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word. For example, stemming will ensure that both *travel* and *traveled* will be recognized by the program as the same word. The software includes stemming algorithms for most European languages including English, French, German, Italian, and Spanish.

**Indexing of Input Documents: Scalability of STATISTICA Text Mining and Document Retrieval**

The indexing of the input documents is extremely fast and efficient, and based on relational database components built into the program. The contents of this database can be saved for further updating in future sessions, or for “deployment,” i.e., to score input documents using only previously selected key terms.

**Results, Summaries, and Transformations**

The Text Mining Results dialog contains numerous options for summarizing the frequency counts of different words and terms. You can also combine terms or phrases (to count them as a single term or phrase), or clear only some of the terms in the analyses.

Options are available for reviewing word/term frequencies or document frequencies, as well as transformations of those frequencies better suited for subsequent analyses (e.g., inverse document frequencies). The Results dialog also contains options for performing singular value decomposition on the documents-by-terms frequency matrix (or transformations of frequencies) to extract dominant “dimensions” into which terms and documents can be mapped.

The scores and coefficients for the extracted dimensions can also be saved for subsequent processing of new documents to map those documents into the same space. Because of the integrated architecture of the STATISTICA system, all results spreadsheets can be used as input data for subsequent analyses or graphs. Hence, it is easy to apply any of the large number of analytic algorithms available in the software to the outputs generated by the Text Mining and Document Retrieval module, for example, to apply cluster analysis methods or any of the methods for predictive data mining to include textual information in those projects.

Let’s look at some examples to see how interesting information can be extracted from unstructured data (or text corpus) using the text mining approach. In the process of the following illustrations, the different features available within the STATISTICA Text Mining and Document Retrieval module will be demonstrated. At the end of this tutorial, you will also understand how the outputs from the Text Mining module can be integrated into the Data Mining module to analyze the numerated text data.