Table 1.2 *Other Keyboard Controls*

Keyboard Shortcut	Action
F6	selects the Properties panel
Tab	navigates the Properties panel
F2+Spacebar	simulates the click action for ellipsis in the Properties panel
Ctrl + Shift + n	select a node in a process flow diagram
Ctrl + Shift + c	connect selected nodes in a process flow diagram

Exceptions to Accessibility Standards

Exceptions to the accessibility standards described in Section 508 of the U.S. Rehabilitation Act of 1973 include the following:

- On-screen indication of the current focus is not well-defined in some dialog boxes, in some menus, and in tables.
- High-contrast color schemes are not universally inherited.
- Many controls are not read by JAWS, and the accessible properties of many controls are not surfaced to the Java Accessibility API.

About SAS Text Miner

SAS Text Miner provides tools that enable you to extract information from a collection of text documents and uncover the themes and concepts that are revealed therein. In addition, because you can embed SAS Text Miner nodes in a SAS Enterprise Miner process flow diagram, you can combine quantitative variables with unstructured text in the mining process. This means that you are incorporating text mining with other traditional data mining techniques.

These languages are supported in SAS Text Miner: Arabic, Chinese (simplified and traditional), Czech, Danish, Dutch, English, Finnish, French, German, Greek, Hebrew, Hungarian, Indonesian, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Spanish, Swedish, Thai, Turkish, and Vietnamese. Each language must be licensed individually.

Note: Collections of text in some unsupported languages can still be processed in SAS Text Miner by choosing a supported language that uses the same or similar text encoding as the unsupported language.

SAS Text Miner includes the following SAS Enterprise Miner nodes:

- **Text Import Node** — enables you to create data sets that contain links to documents obtained with file crawl, web crawl, or web search. For more information, see "Overview of the Text Import Node" on page 13.

- **Text Parsing Node** — enables you to parse a document collection in order to quantify information about the terms that are contained therein. For more information, see "Overview of the Text Parsing Node" on page 22.
- **Text Filter Node** — enables you to reduce the total number of parsed terms or documents that are analyzed. For more information, see "Overview of the Text Filter Node" on page 44.
- **Text Topic Node** — enables you explore the document collection by clustering documents and summarizing the collection into a set of "topics." For more information, see "Overview of the Text Topic Node" on page 115.
- **Text Cluster Node** — enables you to cluster documents from the term-document frequency matrix that is created by the **Text Parsing** node and possibly refined by the **Text Filter** node. For more information, see "Overview of the Text Cluster Node" on page 65.
- **Text Rule Builder Node** — enables you to generate rules that are useful in describing and predicting a target variable. For more information, see "Overview of the Text Rule Builder Node" on page 95.
- **Text Profile Node** — enables you see how terms change over time. For more information, see "Overview of the Text Profile Node" on page 83.

You can use SMP mode in SAS 9.4 on a properly enabled SAS Server to deploy the **HP Text Miner** node in a process flow diagram, and use the HPTMINE procedure and the HPTMSCORE procedure. Using the **HP Text Miner** node in a process flow diagram can lead to multithreaded processing gains in many cases. For more information about the **HP Text Miner** node, see the HP Text Miner Node help page in the HPDM nodes help folder, or the HP Text Miner Node chapter of the *SAS Enterprise Miner High-Performance Data Mining Node Reference*.

In SAS Text Miner, the text mining process consists generally of the steps that are listed in the following table.

Step	Action	Description	Tools
1	File Preprocessing	Create a SAS data set from a document collection that is used as input for the **Text Parsing** node.	**Text Import** node, %TMFILTER macro, or SAS DATA step.

Step	Action	Description	Tools
2	Text Parsing	Decompose textual data, and generate a quantitative representation that is suitable for data mining purposes. Parsing might include the following: • stemming • automatic recognition of multi-word terms • normalization of various entities such as dates, currency, percent, and year • part-of-speech tagging • extraction of entities such as organization names, product names, and addresses • support for synonyms • language-specific analyses	**Text Parsing** node
3	Text Filtering	Transform the quantitative representation into a compact and informative format; reduce dimensions.	**Text Filter** node
4	Document Analysis	Cluster, classify, predict, or link concepts.	**Text Topic** node, **Text Cluster** node, **Text Rule Builder** node, **Text Profile** node, and SAS Enterprise Miner predictive modeling nodes

TIP A number of data sets are provided that might be useful for learning how to use SAS Text Miner.

For more information about each action and sample data sets, see the following.

- "File Preprocessing" on page 8
- "Text Parsing" on page 8

- "Text Filtering" on page 9
- "Document Analysis" on page 9
- "SAS Text Miner Sample Data Sets" on page 10

Note: SAS Text Miner 14.1 is not included with the base version of SAS Enterprise Miner 14.1. If your site has not licensed SAS Text Miner 14.1, then the SAS Text Miner nodes will not appear in your SAS Enterprise Miner 14.1 software.

File Preprocessing

The **Text Parsing** node requires an Input Data Source node to precede it in a process flow diagram. Input data for the **Text Parsing** node must be imported into a data source. Furthermore, because the **Text Parsing** node expects input data in a particular format, in most cases you will need to preprocess data before you can import it into a data source.

The **Text Import** node and the SAS %TMFILTER macro can be used to preprocess data.

The **Text Import** node can be used to extract text from many document formats or to retrieve text from websites by crawling the web.

The SAS **%TMFILTER** macro can be used in file preprocessing to extract text from many document formats. You can use this macro to create a SAS data set that can be used to create a data source to use as input for the **Text Parsing** node. The SAS **%TMFILTER** macro does not extract data from individual XML fields. However, you can still accomplish this task. The XML LIBNAME engine and the XML mapper in Base SAS enable you to read an XML file into a SAS data set where the fields of the XML document are the data set variables. This data set can then be used in SAS Text Miner for further preprocessing.

Documents are represented internally in SAS Text Miner by a vector that contains the frequency of how many times each term occurs in each document. This approach is very effective for short, paragraph-sized documents but can cause a harmful loss of information with longer documents. Consider preprocessing long documents in order to isolate content that is really of use in the model that you intend to build. For example, if you are analyzing journal papers, you might find that analyzing only the abstracts gives the best results.

For more information about the **Text Import** node, the **%TMFILTER** macro, or the **Text Parsing** node, see the following:

- "Overview of the Text Import Node" on page 13
- "%TMFILTER Macro" on page 143
- "Overview of the Text Parsing Node" on page 22

Text Parsing

In SAS Text Miner, text parsing is done with the **Text Parsing** node. Advanced techniques enable you to break documents into terms such as words, phrases, multi-word terms, entities, punctuation marks, and terms that are in foreign languages.

- You can process multi-word groups (for example, "off the shelf" or "because of") as single terms.

- You can identify each term's part of speech, based on its context.
- You can extract entities such as addresses, dates, phone numbers, and company names.
- You can choose to ignore all terms that are a particular part of speech.
- You can use a stop list to ignore a specific set of terms, such as a group of low-information words. Conversely, you can use a start list to restrict parsing to only a specific set of terms.
- You can return the root forms (called stems) of terms and treat all terms that have the same stem as equivalent. For example, “grinds”, “grinding”, and “ground” could all be viewed as the term “grind”.
- You can specify synonyms (such as “teach”, “instruct”, “educate”, “train”), and treat them as equivalent.

For more information about the **Text Parsing** node, see the following:

- “Overview of the Text Parsing Node” on page 22

Text Filtering

In SAS Text Miner, text filtering is done with the **Text Filter** node.

- You can explore a parsed document collection with the **Interactive Filter Viewer** of the **Text Filter** node. For more information about the **Interactive Filter Viewer**, see “Interactive Filter Viewer” on page 60.
- You can subset collections of documents based on the attributes of a document or the content of the document.
- You can interactively adjust the stop list and synonyms to focus on the aspects of the collection that are of interest to you.
- For more information about the **Text Filter** node, see “Overview of the Text Filter Node” on page 44.

Document Analysis

Exploration

SAS Text Miner offers visualization diagrams, topic creation, and clustering techniques that enable you to explore a parsed document collection. Applications include content discovery of large knowledge bases such as those that contain email, customer comments, abstracts, or survey data; unsupervised learning of categories; and taxonomy creation.

- You can generate data-driven topics and supply topics that you have defined in the **Text Topic** node for use in scoring new data.
- You can perform hierarchical clustering in the **Text Cluster** node. The node uses a Ward's minimum-variance method to generate hierarchical clusters, and results are presented in a tree diagram.

- You can perform expectation-maximization (EM) clustering in the **Text Cluster** node. EM clustering identifies primary clusters, which are the densest regions of data points, and secondary clusters, which are less dense groups of data points not included in the primary clusters. This is a spatial clustering technique that allows flexibility in the size and shape of clusters.
- You can use other SAS Enterprise Miner nodes for clustering, such as the Clustering and SOM/Kohonen nodes. For more information about these nodes, see the SAS Enterprise Miner Help.

Prediction

You can use SAS Enterprise Miner modeling capabilities to predict target variables, with applications that include the following:

- automatic email routing
- filtering spam
- matching resumes with open positions
- predicting the change in a stock price from contents of news announcements about companies
- predicting the cost of a service call based on the textual description of the problem
- predicting customer satisfaction from customer comments
- identifying authorship from a predetermined set of candidate authors

The **Text Rule Builder** node can be used for prediction.

See the SAS Enterprise Miner Help for information about how to use modeling nodes for target variable prediction.

SAS Text Miner Sample Data Sets

Sample Data

The following sample data sets are provided in the SAMPSIO library for use with SAS Text Miner 14.1.

Table 1.3 *Sample Data Sets*

Data Set	Description	Used In
Abstract	document collection of abstracts of conference papers	**Input Data** node

Data Set	Description	Used In
Afinn_sentiment	This data set is adapted from the AFINN sentiment publicly available English sentiment lexicon. It contains two topics, “Positive Tone” and “Negative Tone,” that can be used as User Topics in the **Text Topic** node. The Afinn_sentiment data set contains information from the AFINN sentiment database, which is made available under the Open Database License.	**Text Topic** node
News	document collection of brief news articles	Input Data node
Tm_abstract_topic	user-defined topics	**Text Topic** node

Default Data Sets

The following data sets are used in the SAS Text Miner 14.1 nodes as default inputs for node properties (for example, stop lists or multi-term lists). They are all in the SASHELP library.

Table 1.4 *Default Data Sets*

Data Set	Description	Used In
<language>_multi	(where <language> is: Eng, Frnch, Germ, Ital, Port, or Span) multi-term lists for various languages	**Text Parsing** node
<language>stop	(where <language> is: Eng, Frch, or Grmn) stop lists for various languages	**Text Parsing** node
Engsynms	synonym list for the English language	**Text Parsing** node

Chapter 2

The Text Import Node

Overview of the Text Import Node

The **Text Import** node serves as a replacement for an Input Data node by enabling you to create data sets dynamically from files that are contained in a directory or from the web. The **Text Import** node takes an import directory that contains text files in potentially proprietary formats such as MS Word and PDF files as input. The tool traverses this directory and filters or extracts the text from the files, places a copy of the text in a plain text file, and a snippet (or possibly even all) of the text in a SAS data set. If a URL is specified, the node will crawl websites and retrieve files from the web and move them to the import directory before doing this filtering process. The output of a **Text Import** node is a data set that can be imported into the **Text Parsing** node.

In addition to filtering the text, the **Text Import** node can also identify the language that the document is in and take care of transcoding documents to the session encoding. For

more on encoding and transcoding, see "SAS Text Miner and SAS Session Encoding" on page 180.

The **Text Import** node relies on the SAS Document Conversion server to extract plain text from various file formats so that the text can be analyzed by SAS Text Miner. The machine must be accessible from the SAS Enterprise Miner server via the host name and port number that were specified at install time.

Note:

- If you run the **Text Import** node in a UTF-8 SAS session, then the node attempts to transcode all filtered text to UTF-8 encoding so that the result data set can be used in a UTF-8 SAS session. In all other SAS session encodings, the **Text Import** node does not transcode the data. Instead, it assumes that the input data is in the same encoding as the SAS session. For more information, see "SAS Text Miner and SAS Session Encoding" on page 180.
- The **Text Import** node is not supported for use in group processing (Start Groups and End Groups nodes).

Text Import Node Input Data

The **Text Import** node does not require a predecessor node in a process flow diagram.

Text Import Node Properties

Contents

Text Import Node General Properties

These are the general properties that are available on the **Text Import** node:

- **Node ID** — displays the ID that is assigned to the node. Node IDs are especially useful for distinguishing between two or more nodes of the same type in a process

flow diagram. For example, the first **Text Import** node that is added to a diagram will have the Node ID **`TextImport`**, and the second Text Import node that is added will have the Node ID **`TextImport2`**.

- **Imported Data** — accesses a list of the data sets that are imported by the node and the ports that provide them. Click the ellipsis button to open the Imported Data window, which displays this list. If data exists for an imported data set, then you can select a row in the list and do any of the following:
  - browse the data set
  - explore (sample and plot) the data in a data set
  - view the table and variable properties of a data set
- **Exported Data** — accesses a list of the data sets exported by the node and the ports to which they are provided. Click the ellipsis button to open the Exported Data window, which displays this list. If data exists for an exported data set, then you can select a row in the list and do any of the following:
  - browse the data set
  - explore (sample and plot) the data in a data set
  - view the table and variable properties of a data set
- **Notes** — accesses a window that you can use to store notes of interest, such as data or configuration information. Click the ellipsis button to open the Notes window.

Text Import Node Train Properties

General Train Properties

These are the training properties that are available on the **Text Import** node:

- **Import File Directory** — specifies the path to the directory that contains files to be processed. Click the ellipsis for this property to select a directory accessible by the server for import.
- **Destination Directory** — specifies the path to the directory that will contain plain text files after processing. Click the ellipsis for this property to specify a destination directory that is accessible by the server.
- **Language** — Specifies the possible choices that the language identifier might choose from when assigning a language to each document. Click the ellipsis for this property to open the Language dialog box to specify one or more languages. Only languages that are licensed can be used.
- **Extensions** — restricts the **Text Import** node to filtering only files that satisfy the provided file type. All file types that the SAS Document Converter supports are filtered when the setting is not specified. See SAS Document Conversion for more information.
- **Text Size** — specifies the number of characters to use in the TEXT variable of the output data set. This variable can serve as a snippet when the size is small, or you can set the value to as large as 32000, so that as much text as possible is placed in the data set.
- "Web Crawl Properties" on page 16