



THE
POWER
TO KNOW.

SAS[®] Text Miner 14.1 Reference Help

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS® Text Miner 14.1: Reference Help*. Cary, NC: SAS Institute Inc.

SAS® Text Miner 14.1: Reference Help

Copyright © 2015, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hardcopy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software–Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

Electronic book 1, July 2015

SAS® Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/publishing or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

Chapter 1 • Introduction to SAS Text Miner and Text Mining	1
What's New in SAS Text Miner 14.1	1
Replacing the Original Text Miner Node	3
Accessibility Features of SAS Text Miner 14.1	4
About SAS Text Miner	5
File Preprocessing	8
Text Parsing	8
Text Filtering	9
Document Analysis	9
SAS Text Miner Sample Data Sets	10
Chapter 2 • The Text Import Node	13
Overview of the Text Import Node	13
Text Import Node Input Data	14
Text Import Node Properties	14
Text Import Node Results	17
Text Import Node Output Data	18
Using the Text Import Node	18
Chapter 3 • The Text Parsing Node	21
Overview of the Text Parsing Node	22
Text Parsing Node Input Data	22
Text Parsing Node Properties	23
Text Parsing Node Results	27
Text Parsing Node Output Data	29
Using the Text Parsing Node	29
Start Lists and Stop Lists	34
Term Stemming	35
Term Roles and Attributes	36
Synonym Lists	39
Multi-Term Lists	41
Chapter 4 • The Text Filter Node	43
Overview of the Text Filter Node	44
Text Filter Node Input Data	44
Text Filter Node Properties	45
Text Filter Node Results	48
Text Filter Node Output Data	51
Using the Text Filter Node	51
How to Create a Dictionary Data Set	56
Term Weighting	57
Text Filter Node Search Expressions	59
Interactive Filter Viewer	60
Strength of Association for Concept Linking	63
Chapter 5 • The Text Cluster Node	65
Overview of the Text Cluster Node	65
Text Cluster Node Input Data	66
Text Cluster Node Properties	66
Text Cluster Node Results	68
Text Cluster Node Output Data	69

Using the Text Cluster Node	70
Singular Value Decomposition	74
Clustering Techniques	79
Chapter 6 • The Text Profile Node	83
Overview of the Text Profile Node	83
Text Profile Node Input Data	84
Text Profile Node Properties	84
Text Profile Node Results	86
Text Profile Node Output Data	88
Using the Text Profile Node	88
Chapter 7 • The Text Rule Builder Node	95
Overview of the Text Rule Builder Node	95
Text Rule Builder Node Input Data	96
Text Rule Builder Node Properties	96
Text Rule Builder Node Output Data	99
Text Rule Builder Node Results	100
Using the Text Rule Builder Node with the HPBOOLRULE Procedure	103
Using the Text Rule Builder Node	104
Chapter 8 • The Text Topic Node	115
Overview of the Text Topic Node	115
Text Topic Node Input Data	116
Text Topic Node Properties	116
Text Topic Node Results	119
Text Topic Node Output Data	121
Using the Text Topic Node	121
User-Defined Topic Lists	128
Interactive Topic Viewer	129
Chapter 9 • Output Data	131
Output Data for SAS Text Miner Nodes	131
Chapter 10 • Overview of Scoring with Text Miner	133
Overview of Scoring with SAS Text Miner	133
The Scoring Process	133
Chapter 11 • Macro Variables, Macros, and Functions	135
Using Macro Variables to Set Additional Properties	135
Overview of Macros and Functions	139
%TEXTSYN Macro	139
%TMFILTER Macro	143
TGSCORE Function	153
Chapter 12 • Additional Examples	157
Classifying News Articles	157
Scoring New Documents	163
Using the Association Node with SAS Text Miner Nodes	165
Creating a Stored Process	167
Chapter 13 • Converting Diagrams and Session Encoding	179
Converting SAS Text Miner Diagrams from a Previous Version	179
SAS Text Miner and SAS Session Encoding	180
Overview of the Text Miner Node	180
Text Miner Node Input Data	181

Text Miner Node Properties	183
Text Miner Node Results	187
Text Miner Node Output Data	189
Appendix 1 • Additional Information about Text Mining	193
Index	195

Chapter 1

Introduction to SAS Text Miner and Text Mining

What's New in SAS Text Miner 14.1	1
Replacing the Original Text Miner Node	3
Accessibility Features of SAS Text Miner 14.1	4
Overview of Accessibility Features	4
Additional Keyboard Controls for SAS Text Miner	4
Exceptions to Accessibility Standards	5
About SAS Text Miner	5
File Preprocessing	8
Text Parsing	8
Text Filtering	9
Document Analysis	9
Exploration	9
Prediction	10
SAS Text Miner Sample Data Sets	10
Sample Data	10
Default Data Sets	11

What's New in SAS Text Miner 14.1

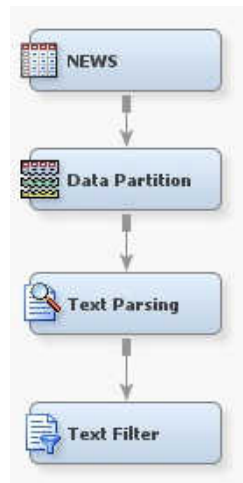
SAS Text Miner 14.1 includes the following new features and enhancements:

- A new HPBOOLRULE procedure replaces macros in the **Text Rule Builder** node. For more information about the benefits of running the **Text Rule Builder** node with the HPBOOLRULE procedure and how you can specify to run the **Text Rule Builder** node as it functioned prior to SAS Text Miner 14.1, see [“Overview of the HPBOOLRULE Procedure in the Text Rule Builder Node”](#) on page 103.
- Enhancements to the HPTMINE procedure enable you to select or ignore parts of speech, attributes, and entities, as well as to build a search index.
- The **HP Text Miner** node now uses PROC HPTMINE to perform topic rotation and to create the topic table.

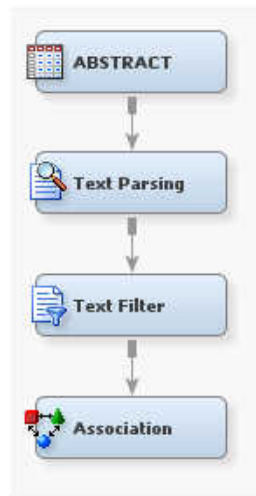
- Eleven parsing languages have been added to the **Language** property in the **HP Text Miner** node. The complete list of parsing languages includes: Chinese, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, and Turkish.
- The new macro variable EM_TERM_LOC enables users to specify a location for SAS Text Miner nodes to write output data sets. These data sets are needed as input to SAS Text Miner score code. When this macro variable is set, it is recognized by the **Text Filter**, **Text Topic**, **Text Cluster**, **Text Profile**, and **Text Rule Builder** nodes as an output location. The score code generated by these nodes will use the data sets saved in the specified location as input. For more information about using the EM_TERM_LOC macro variable, see [“Using Macro Variables to Store Prebuilt Scoring Data in Tables Instead of Prescore Code” on page 136](#). For an example that demonstrates how to create a stored process using the EM_TERM_LOC macro variable, see [“Creating a Stored Process” on page 167](#).

Note: A SAS Text Miner license is still required to score with a SAS Text Miner model even with this new macro setting.

- The TGPARSE procedure has been replaced with the HPTMINE procedure in the **Text Parsing** node when the parsing language is Chinese, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, or Turkish. For these languages, parsing is now multithreaded. The TGPARSE procedure will continue to be called when the parsing language is Arabic, Czech, Danish, Greek, Hebrew, Hungarian, Indonesian, Norwegian, Polish, Romanian, Slovak, Swedish, Thai, or Vietnamese.
- An **_item_** variable with term | role information has been added to the transaction output that is exported from the **Text Topic** node and the **Text Filter** node. This variable is added to the transaction tables **valid_trans** and **test_trans** when a **Data Partition** node is used in a process flow diagram, such as in the following:



One benefit of having term | role information exported in the transaction table is that the **Association** node will show this information in the rules that it generates if used in a process flow diagram, such as in the following:



For more information, see <http://support.sas.com/software/products/txtminer>.

Replacing the Original Text Miner Node

The functionality that was available in the original **Text Miner** node has been moved to other nodes that are available with SAS Text Miner. This restructuring of functionality conforms more to the overall philosophy of SAS Enterprise Miner components. It also improves performance because you can make changes in nodes that follow the **Text Parsing** node without having to reparse the collection. The following table might be helpful to you in replacing the functionality that you are using in the original **Text Miner** node with the new nodes.

Controls and Functionality in the Original Text Miner Node	Replacement in New SAS Text Miner Nodes
Parsing	Text Parsing node
Term weightings	Text Filter node
Concept Linking Diagram	Text Filter node. The concept linking diagram is available in the Interactive Filter Viewer.
Creating and removing synonyms interactively	Text Filter node
Dynamically keeping and dropping terms	Text Filter node
Subsetting data for reclustering	Text Filter node
Clustering (Expectation Minimization and Hierarchical)	Text Cluster node
Generation of SVD values	Text Cluster node. Use this node output as input to your predictive models, for example.

Controls and Functionality in the Original Text Miner Node	Replacement in New SAS Text Miner Nodes
Roll-up Terms	Text Topic node. Set the number of single term topics to the number of Roll-up Terms that you desire.

Accessibility Features of SAS Text Miner 14.1

Overview of Accessibility Features

SAS Text Miner 14.1 includes accessibility and compatibility features that improve the usability of the product for users with disabilities, with exceptions noted below. These features are related to accessibility standards for electronic information technology that were adopted by the U.S. Government under Section 508 of the U.S. Rehabilitation Act of 1973, as amended. SAS Text Miner 14.1 conforms to accessibility standards for the Windows platform.

For specific information about Windows accessibility features, refer to your operating system's help. If you have questions or concerns about the accessibility of SAS products, send email to accessibility@sas.com.

For information about the accessibility features of SAS Enterprise Miner 14.1, see the Accessibility topic in the SAS Enterprise Miner 14.1 Help.

- [“Additional Keyboard Controls for SAS Text Miner” on page 4](#)
- [“Exceptions to Accessibility Standards” on page 5](#)

Additional Keyboard Controls for SAS Text Miner

In addition to standard keyboard controls, SAS Text Miner supports the following additional keyboard controls.

Table 1.1 Keyboard Controls for Tables

Keyboard Shortcut	Action
Shift + Page Up	selects all rows in the table from the first row that is visible in the scroll pane to the selected row
Shift + Page Down	selects all rows in the table from the selected row to the last row that is visible in the scroll pane
Ctrl + Shift + End	selects all rows in the table from the selected row to the last row
Ctrl + Shift + Home	selects all rows in the table from the first row to the selected row