SAS® Contextual Analysis 13.2: User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2014. *SAS® Contextual Analysis 13.2: User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Contextual Analysis 13.2: User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

August 2014

SAS provides a complete selection of books and electronic products to help customers use SAS® software to its fullest potential. For more information about our offerings, visit **support.sas.com/bookstore** or call 1-800-727-3228.

Contents

Using This Book

Audience

This book is designed for new users of SAS Contextual Analysis. It describes the terminology used in SAS Contextual Analysis and provides instructions for tasks.

What's New

What's New in SAS Contextual Analysis 13.2

Overview

SAS Contextual Analysis 13.2 is a major product release that offers an enhanced user interface and includes the following new and enhanced features:

- wizard for creating and editing projects
- import capability for SAS Enterprise Content Categorization projects
- enhanced Properties page with messages
- document-level sentiment scoring
- score code downloading
- concept creation
- generated category rules that follow the format of SAS Enterprise Content Categorization category rules (MCAT)
- enhanced document viewing
- online Help topics

New Wizards and Import Feature

The Create Project wizard provides an easy-to-use interface that enables you to quickly create projects. Through the wizard, you specify options for your project; these options identify data sources, the variables you want to analyze, start lists, and stop lists. You can also import SAS Enterprise Content Categorization projects through the wizard. The Edit Project wizard provides flexibility and convenience for updating your project data and rerunning your project.

Enhanced Properties Page with Messages

The Properties page has been enhanced to include run status for each analysis task. The enhanced Properties page also includes a messages panel that provides detailed messages for each task that you run in your project. This enhancement helps you quickly troubleshoot your analysis model.

New Document-Level Sentiment Scoring and Score Code Downloading

A sentiment score is now provided for each document that contains a topic or is selected for a category. You can view score code for sentiment, concepts, and categories, and download it for your use in scoring other documents.

Concept Creation

You can now create concepts by supplying your own concept rules, which you can validate before you apply. You can enable and disable concepts for inclusion in your project run. Predefined concepts are provided for commonly used concepts such as **ADDRESS**, **COMPANY**, and **DATE**.

Enhanced Category Rule Generation and Document Viewing

You can now generate, edit, and validate category rules that follow the format of SAS Enterprise Content Categorization category rules (MCAT). The documents in the categories and topics tasks can be displayed in several views to meet your purposes.

Accessibility

For information about the accessibility of this product, see Accessibility Features of SAS Contextual Analysis 13.2 at support.sas.com.

Recommended Reading

Here is the recommended reading list for this title:

- The online Help for SAS Contextual Analysis 13.2
- *SAS Contextual Analysis: Administrator's Guide*
- *SAS Content Categorization Single User Servers 12.1: Administrator's Guide*
- *SAS Text Miner: Reference Help*
- *SAS Enterprise Content Categorization Studio 12.1: User's Guide*

The recommended reading list from SAS Press includes the following title:

- *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*

For a complete list of SAS books, go to support.sas.com/bookstore. If you have questions about which titles you need, please contact a SAS Book Sales Representative:

SAS Books
SAS Campus Drive
Cary, NC 27513-2414
Phone: 1-800-727-3228
Fax: 1-919-677-8166
E-mail: sasbook@sas.com
Web address: support.sas.com/bookstore

1

Introduction to SAS Contextual Analysis

What Is SAS Contextual Analysis?

SAS Contextual Analysis is a web-based text analytics application that uses contextual analysis to provide a comprehensive solution to the challenge of identifying and categorizing key textual data. Using this application, you can build models (based on training documents) that automatically analyze and categorize a set of documents. You can then customize your models in order to realize the value of your text-based data.

SAS Contextual Analysis combines the machine-learning capabilities of SAS Text Miner with the rules-based linguistic methods of categorization and extraction in SAS Enterprise Content Categorization. These capabilities, along with document-level sentiment scoring, are combined in a single user interface.

Using SAS Contextual Analysis, you can identify key textual data in your document collections, categorize those data, build concept models, and remove meaningless textual data.

By default, words that provide little or no value are excluded from analysis. Examples of these words include the articles *a*, *an*, and *the* and conjunctions such as *and*, *or*, and

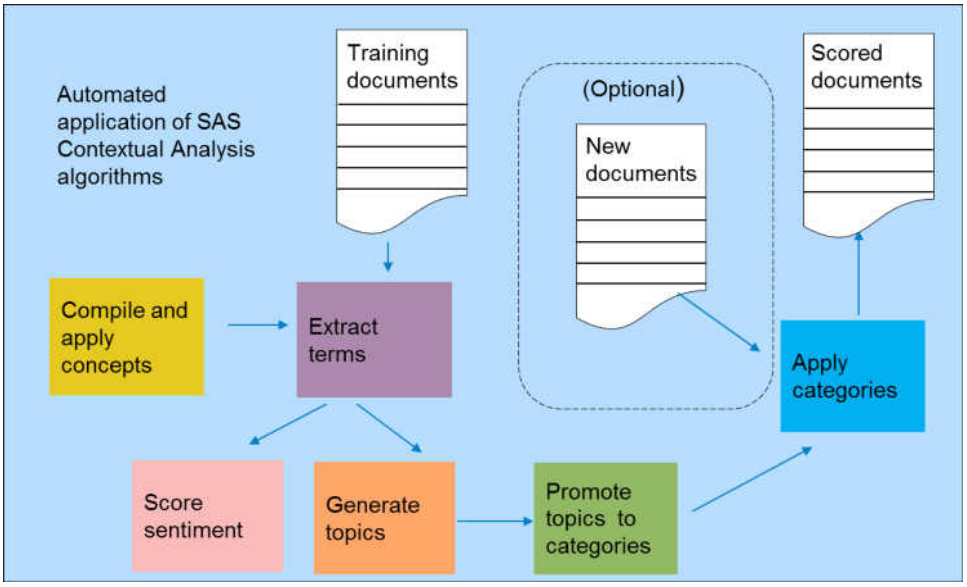
but. Other terms that are specific to your document collection but provide little or no value are also identified and excluded.

SAS Contextual Analysis is designed for users who have no SAS programming or SAS macro language experience.

How Does SAS Contextual Analysis Work?

Figure 1.1 provides an overview of the SAS Contextual Analysis processes.

Figure 1.1 *Process Overview*



SAS Contextual Analysis enables you to select concepts or create additional custom concepts that you want to discover in a document or set of documents. For more information about concepts, see the section "Concepts" on page 18.

The SAS Contextual Analysis algorithms group similar documents in a collection into topics. The documents in each topic often contain similar subject matter, such as motorcycle accidents, computer graphics, or weather patterns. Automatic topic identification enables you to easily categorize each document in your collection. After

you determine the topics that you want to use, you promote those topics into categories. Preliminary rules are generated when you promote a topic to a category.

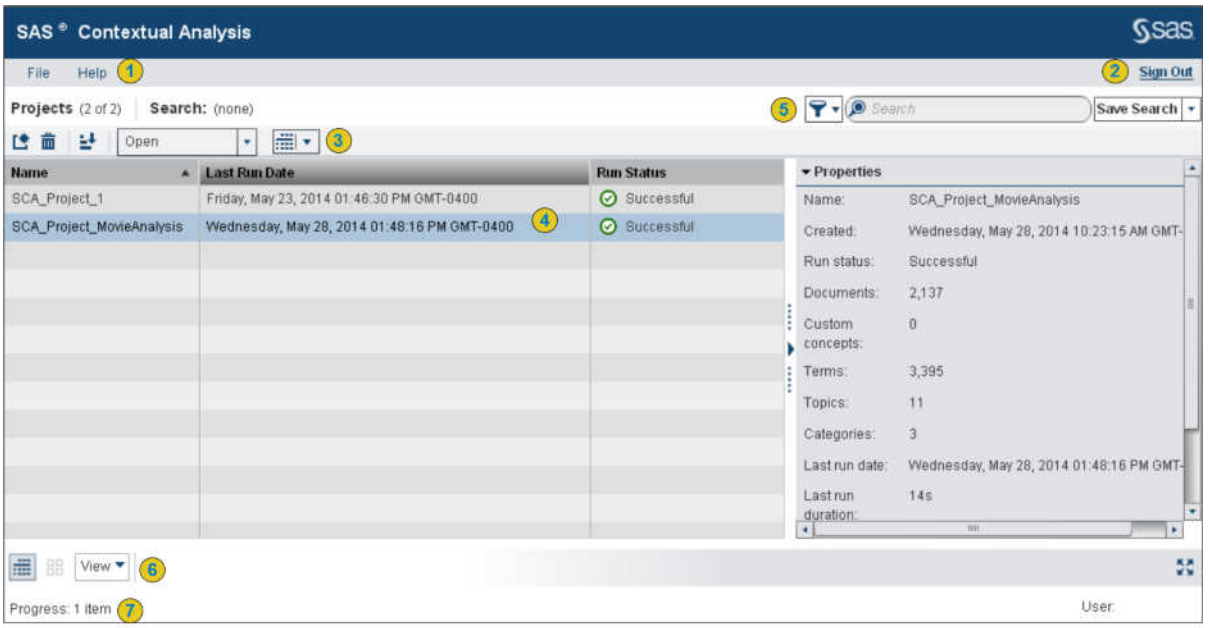
Whether you use the automated processes and rules that are available for extracting terms and subject matter or you customize the processes and rules, context sensitivity is an essential component of your model. To enhance context sensitivity, you can modify the preliminary rules. You can add or modify Boolean operators, characters, and other selections to make the rule matching more context-sensitive.

Finally, you deploy your model to automate the process of classifying a set of input documents.

Using the Interface

The main components of the user interface are shown in Figure 1.2.

Figure 1.2 *SAS Contextual Analysis Interface*



1 Application menu

2 Sign out

3 Application toolbar

4 Project list

5 Search options

6 View options

7 Progress panel (click to open)

2

Projects in SAS Contextual Analysis

Overview of a Project

In SAS Contextual Analysis, you create projects, which are basically containers for your data and analysis. A project contains the input data, text mining options, and analysis tasks (working with concepts, terms, topics, and categories). SAS Contextual Analysis is designed so that you can create and run multiple projects simultaneously. Text

analysis is performed in the background so that you can open one project while performing analysis on a different project.

When you build a model, you choose input data that contain the document collection that you want to use as a training data set. It is important to ensure that your training data are representative of the data to which this model will be applied. Concepts, topics, and categories are built based on the terms in this document collection.

Before you can run your project on a SAS data set, you must specify the text field that you want to analyze. You can also specify one or more category variables for the analysis. Next, you can choose to specify either a start list or a stop list. Finally, you can specify whether to use a synonym list.

After the project runs, you can view the terms and automatically discovered topics that were created during the initial text mining. Then, you use the topics to create categories, which are groups of documents that contain similar terms. SAS Contextual Analysis builds a set of rules for each category.

Preparing the Document Collection

Before you create a project in SAS Contextual Analysis, you need to prepare your document collection for analysis. SAS Contextual Analysis enables you to analyze a document collection that is stored as a SAS data set or in text-based file formats such as MS Office, OpenDocument (OpenOffice), PDF, XML, HTML, and others. You can select a SAS data set and then identify the text variables and category variables to be analyzed. Or you can specify a directory that contains the files that you want to use as training data.

When preparing the input document collection, you should select a set of documents that is representative of the documents that you want to categorize later. The terms that exist in the input document collection are used to build the topics and categories.

There are no standard rules for creating an input document collection. However, the following guidelines can help you prepare your input document collection:

- Include at least 15 to 20 documents for each category that you want to discover.

- Be familiar with the contents of the documents in order to anticipate term discovery and rule creation.
- Do not store SAS data sets in the same directory where you store a collection of Microsoft Word or Adobe PDF documents.

Note: When using a SAS data set, you must register that data set with the SAS Metadata Server before it is available in SAS Contextual Analysis. You can use SAS Management Console and SAS Enterprise Guide to register data sets. When using a collection of documents in a folder (rather than a SAS data set), you must locate the folder on the server where the SAS Contextual Analysis workspace server is installed. For more information, see *SAS Contextual Analysis 13.2: Administrator's Guide*.

Importing an Existing SAS Enterprise Content Categorization Project

During project creation, you can import an existing SAS Enterprise Content Categorization project for analysis. Concepts that were defined by using the LITI syntax in an imported project can be used in the project that you are creating. Rules that were created in the format that is used in SAS Enterprise Content Categorization (MCAT) are imported as categories in SAS Contextual Analysis.

Note: In order for the LITI concepts to be parsed correctly in SAS Contextual Analysis, the parsing priority for disabled concepts must be honored. To ensure this, open your existing project in SAS Enterprise Content Categorization. For any child concept that was disabled, modify its parent concept so that the parent has a higher priority than the child. Save the project before you import it into SAS Contextual Analysis.

Creating a New Project

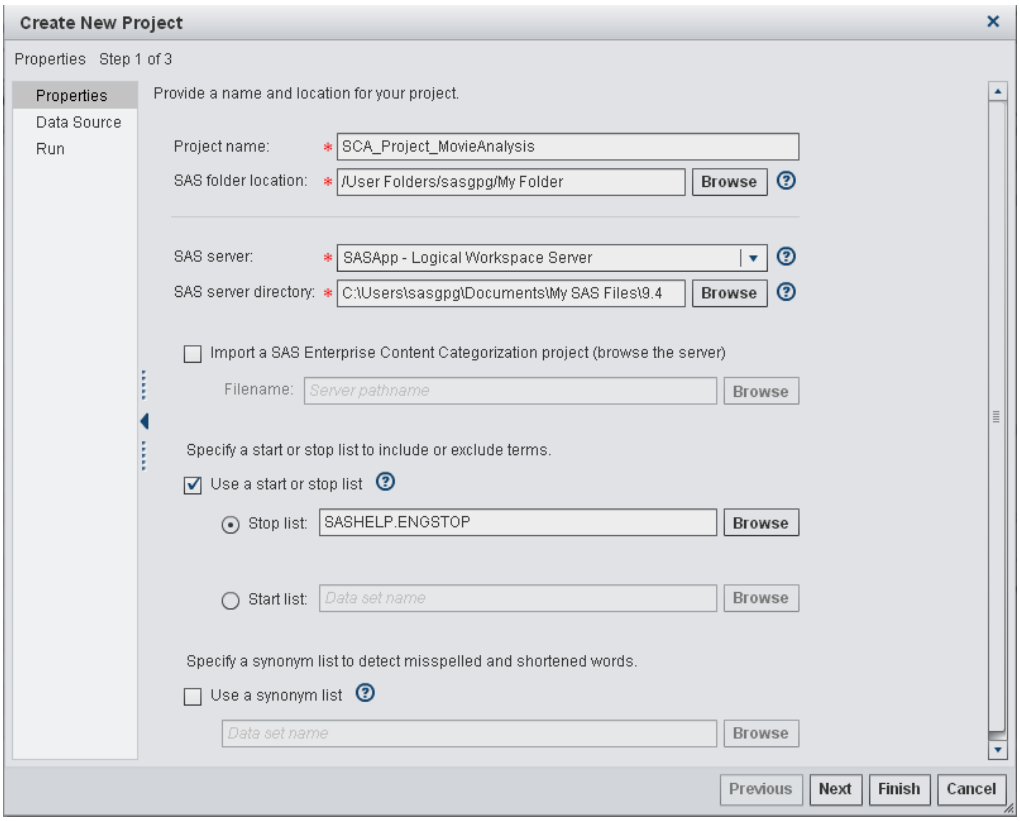
Using the Create Project Wizard

The first time you log on to SAS Contextual Analysis, you must create a project before you can do anything else. To create a new project, click the  icon near the upper left corner of the main window. The Create New Project wizard appears, where you can enter all the specifics for your project.

TIP Click the Help icon  in the Create New Project wizard for information about a specific field or pane.

Step 1: Identify Project Files, Servers, and Lists

1. Enter the project name and specify where the project folder can be accessed in SAS metadata. Indicate the SAS server and SAS server directory.
2. (Optional) Import a SAS Enterprise Content Categorization project. For more information, see the section “Importing an Existing SAS Enterprise Content Categorization Project” on page 7.
3. (Optional) Identify a start list or a stop list (but not both) to control which terms to include or exclude during text mining analysis. For more information, see the section “Start Lists and Stop Lists” on page 19. A default stop list is selected by default.
4. (Optional) Specify a synonym list to identify pairs of words that should be treated as single terms for analysis. For more information, see the section “Terms and Synonyms” on page 18.



Step 2: Identify a Data Source

Here are the options for selecting a data source:

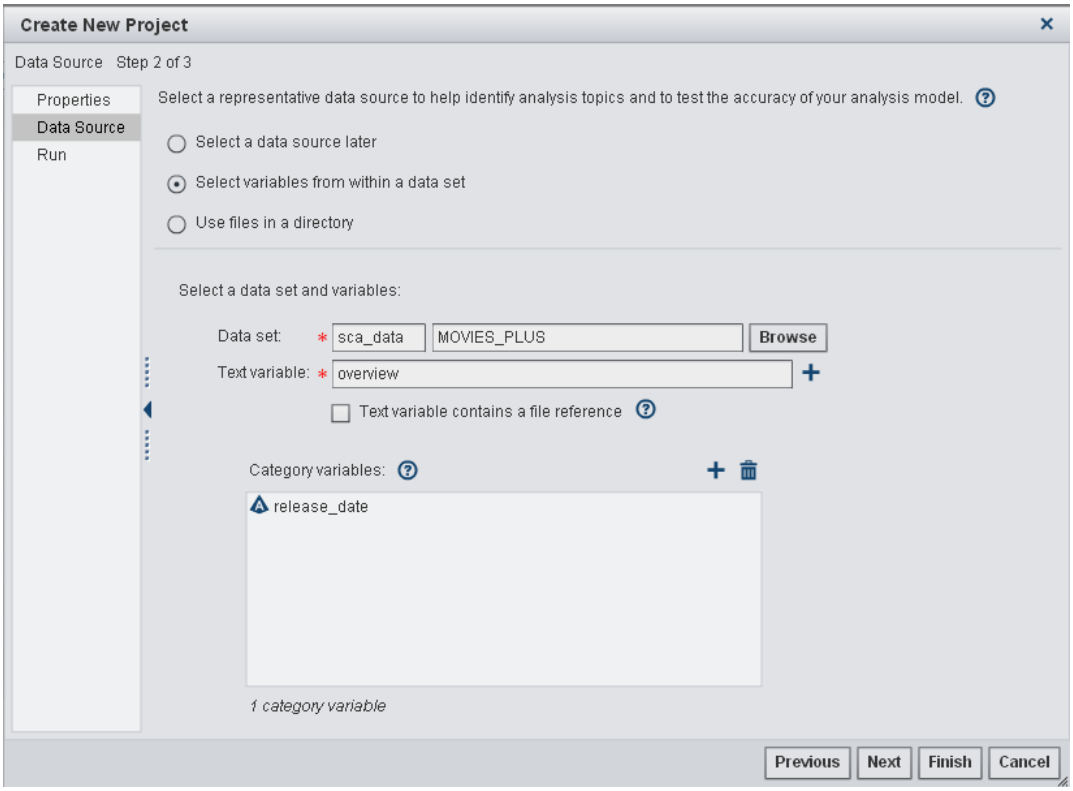
- You can choose a data source now or later. If you choose a data source later, you can still enter more information for your project in the Edit Project wizard.
- Specify analysis variables from within a SAS data set. If you choose this option, you must specify the data set library and name and specify the text variable that you want to analyze. Rather than including the full text of a document in a SAS variable, you can enter a file reference, which identifies the location of a file. Using a file reference is the only way to analyze documents that are longer than 32,767 characters.

  In addition, you can specify one or more category variables to indicate how you want the documents to be grouped. For example, suppose you are analyzing customer

comments from hotel stays. The data column Hotels includes names of hotels where customers stayed. If you specify Hotels as a category variable, then category rules are automatically generated. Subcategory rules are also generated for each hotel that appears frequently in the data.

- Specify a document collection that is stored in text-based file formats such as MS Office, OpenDocument (OpenOffice), PDF, XML, or HTML. The files must be located in a folder. You can define categories later.

In the following example, the text variable *overview* and the category variable *release_date* are selected from the data set *MOVIES_PLUS*.



Step 3: Run the Project

You can choose to run the project now or later. See the Help for more information about when to run the project.

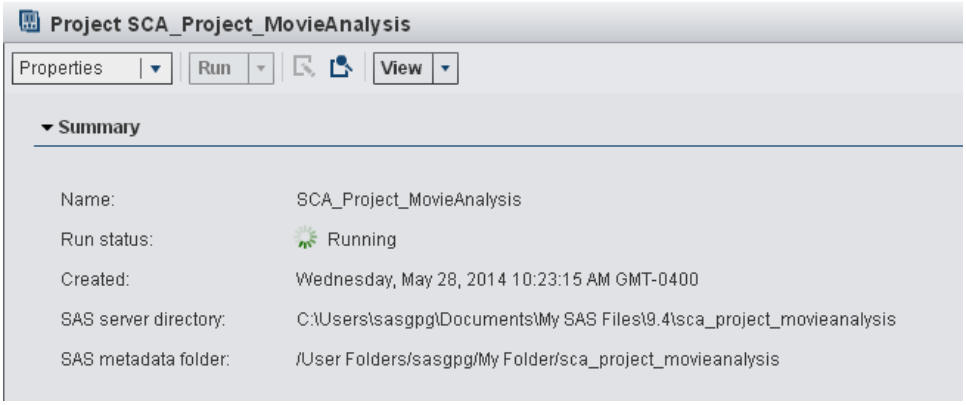
If you choose to run the project now, the following events occur for data sources that are provided:

- Parsing takes place.
- Topics are generated.
- Rules are generated and run for any category variables that you specified.

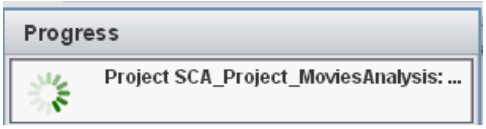
If you import a SAS Enterprise Content Categorization project and then run the project, the following events occur:

- Imported concepts are compiled and applied to the data source that is provided.
- Imported categories are compiled and applied to the data source provided.

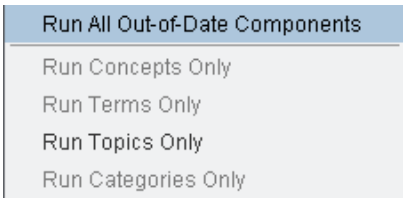
The **Run status** field on the Properties page indicates whether a project is running.



Note: You can also check the **Progress** panel, which is located in the lower left corner of the main window. Click the word **Progress** to see which projects are running and which projects have finished running.



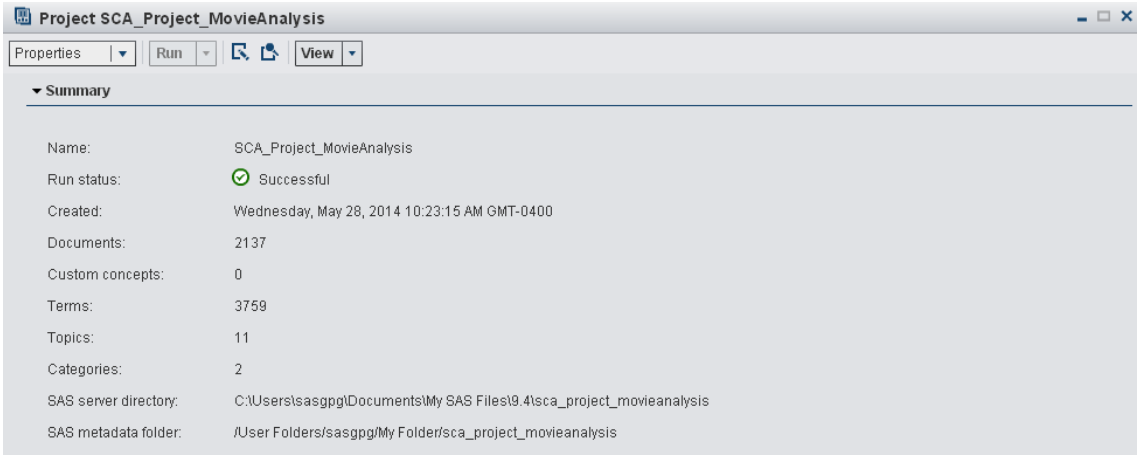
The **Run** menu enables you to run components individually or run only the components that are out of date (and their dependent tasks, if any).



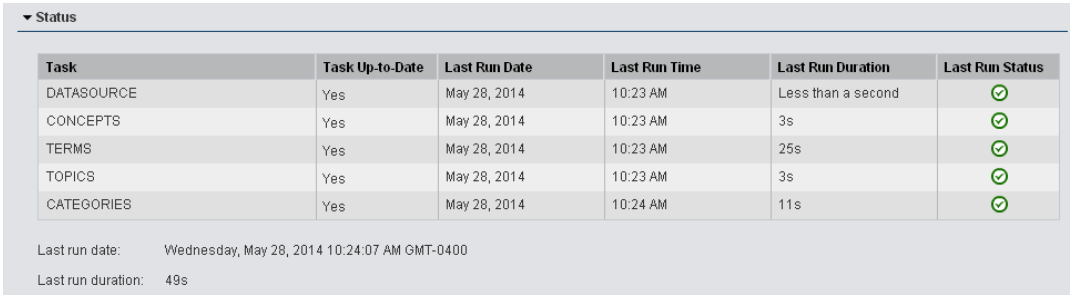
Using the Properties Page

Checking Project Status

The Properties page indicates whether the project ran successfully and provides basic information about the data that were analyzed.



The resulting status of each analysis task that was run is displayed in the following fields in the **Status** section:



▾ Status

Task	Task Up-to-Date	Last Run Date	Last Run Time	Last Run Duration	Last Run Status
DATASOURCE	Yes	May 28, 2014	10:23 AM	Less than a second	✓
CONCEPTS	Yes	May 28, 2014	10:23 AM	3s	✓
TERMS	Yes	May 28, 2014	10:23 AM	25s	✓
TOPICS	Yes	May 28, 2014	10:23 AM	3s	✓
CATEGORIES	Yes	May 28, 2014	10:24 AM	11s	✓

Last run date: Wednesday, May 28, 2014 10:24:07 AM GMT-0400

Last run duration: 49s

Task

The name of the analysis task

Task Up-to-Date

Indicates whether information in the task has changed since the last time the task was run. If no information has changed, the value is **`Yes`** and no further action is required. If information has changed in the task since the last run, then the value is **`No`** and the task should be rerun.

Last Run Date and Last Run Time

The last date and time the task was run

Last Run Duration

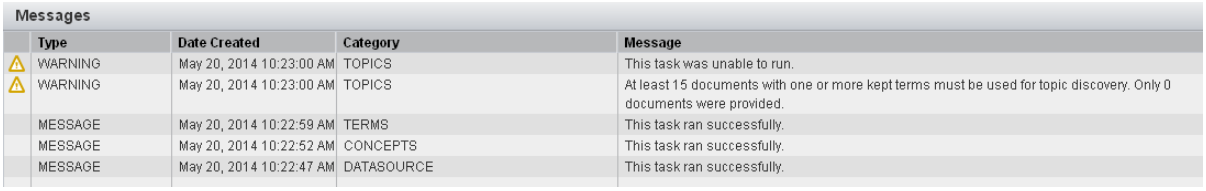
The duration of the task's last run

Last Run Status

Indicates whether the task's run was successful ✓, the run failed ⊗, the task did not run ○, or the task was unable to run (because of missing or insufficient information) ⚠. A status of **Not run** ○ is displayed if the failure of a task prevents dependent tasks from running. In the following example, the failure of the **`CONCEPTS`** task prevented the **`TERMS`** and **`TOPICS`** tasks from running. You must correct the error and rerun the project until it runs successfully.

Task	Task Up-to-Date	Last Run Date	Last Run Time	Last Run Duration	Last Run Status
DATASOURCE	Yes	May 28, 2014	10:23 AM	Less than a second	✓
CONCEPTS	No	May 28, 2014	10:54 AM	4s	⊗
TERMS	No				○
TOPICS	No				○
CATEGORIES	Yes	May 28, 2014	10:24 AM	11s	✓

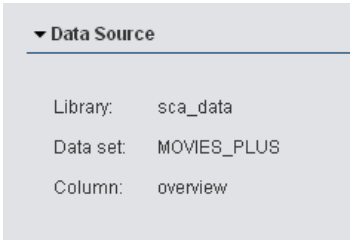
Click the Messages icon to see messages about the status of each task. Following is an example of the Messages window for a task that has warnings. The **Category** column indicates the analysis task.



Messages

	Type	Date Created	Category	Message
⚠	WARNING	May 20, 2014 10:23:00 AM	TOPICS	This task was unable to run.
⚠	WARNING	May 20, 2014 10:23:00 AM	TOPICS	At least 15 documents with one or more kept terms must be used for topic discovery. Only 0 documents were provided.
	MESSAGE	May 20, 2014 10:22:59 AM	TERMS	This task ran successfully.
	MESSAGE	May 20, 2014 10:22:52 AM	CONCEPTS	This task ran successfully.
	MESSAGE	May 20, 2014 10:22:47 AM	DATASOURCE	This task ran successfully.

The data source information is displayed at the bottom of the Properties page.



Data Source

Library: sca_data

Data set: MOVIES_PLUS

Column: overview

Editing Project Information

Click the Edit icon to edit basic information for your project (such as project name). The Edit Project wizard appears. Items that you cannot edit appear in gray.

Note: You must rerun the project to see the effects of your changes.

Viewing and Downloading Code

You can view and download SAS score code that is created. Score code enables you to apply the text analytic models in your project (concepts, categories, and sentiment) to other data.

On the Properties page, click **View**. Select **Concept code**, **Sentiment code**, or **Categories code**. Click to copy the code for use in other programs. If you want to

download the code to a file, click  and follow the prompts to specify a location for the file.

Scoring an External Data Set

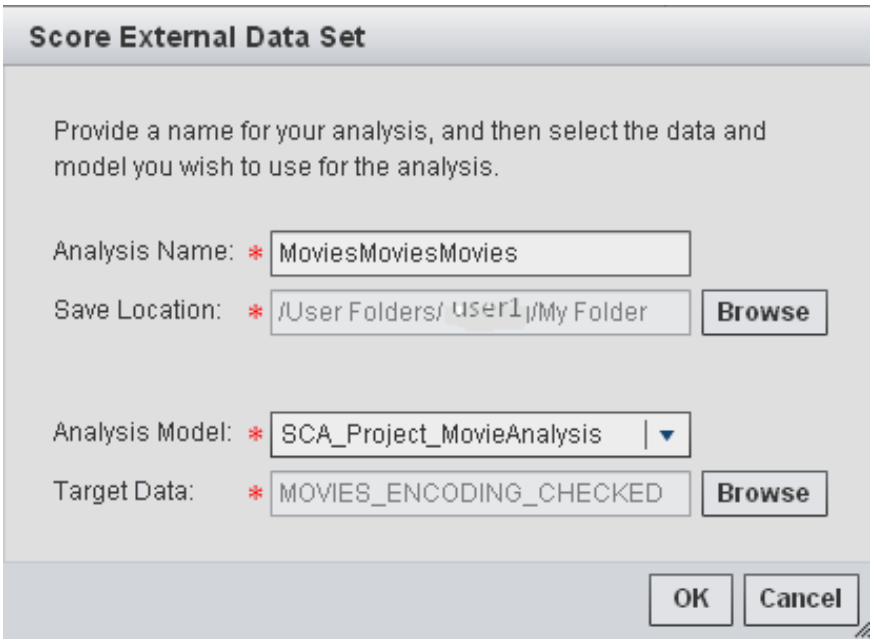
You can use the model that you built in your SAS Contextual Analysis project to score an external data set. When you score an external data set, the category model is applied to the external data set (called the target data set) and categorization information for the document collection is output into a scored data set.

Note: The data set must be stored in a file outside of a folder. If your project uses a folder as a data source, you cannot score data sets within the folder.

To score an external data set:

1. Select **File ->Score External Data Set** from the application's main menu.
2. In the **Analysis Name** field, enter the name for the scored data set that is to be generated.
3. Enter a metadata location for saving the data.
4. In the **Analysis Model** field, enter the name of the project that contains the model that you are using.
5. Provide the target data set to score. The target data set must have the same text variable as the selected project's data source.

Note: To be eligible for scoring, a project must have a compiled category binary file. A category binary file is generated when you run a project that contains categories.



After the scoring begins, the project's run status changes to **Running**. After the scoring is complete, the scored data set is placed in the library folder where the project that you used as your analysis model is stored.

3

Performing the Analysis Tasks

Overview of the Analysis Tasks

Introduction

When you run a project, the following analysis tasks are performed (if data are present).

- concepts
- terms (including synonyms)
- topics

- categories

The following sections describe each task.

Concepts

A *concept* is a property such as a book title, last name, city, gender, and so on. Concepts are useful for analyzing information in context. You can write rules for recognizing concepts that are important to you, thereby creating custom concepts. For example, you can specify that the concept *kitchen* is identified when the terms *refrigerator*, *sink*, and *countertop* are encountered in text.

SAS Contextual Analysis provides *predefined concepts*, which are concepts whose rules are already written. Predefined concepts save time by providing you with commonly used concepts and their definitions, such as **COMPANY** or **TITLE**. You cannot edit predefined concepts or their rules, but you can append additional rules in the code editor that is provided.

A *custom concept* is a concept whose rules you must write.

Note: If you have imported a SAS Enterprise Content Categorization project, the concepts that were created using LITI rules will appear in your project as custom concepts. You can edit them further by using the rules editor.

For more information about writing concept rules, see the sections “The Rule Types,” “Contextual Extraction Concept Rule Examples,” and “Specifying Regular Expressions” in *SAS Enterprise Content Categorization 12.1: User’s Guide*. For information about writing Boolean rules, see the section “About Boolean Operators” in *SAS Content Categorization Studio 12.1: User’s Guide*.

Terms and Synonyms

A *term* is defined as a label for a group of strings or patterns that represent a single concept (an idea) as defined by underlying rules or algorithms. In SAS Contextual Analysis, a term is the basic building block for topics, term maps, and category rules. Each term has an associated role that either is blank or identifies the term’s part of speech. A term reflects one or more surface forms. A *surface form* is a variant of a term

that is located in a matched subset of text. Surface forms can include stems, synonyms, misspellings, and other ways of referring to a term.

A *synonym list* is a SAS data set that identifies pairs of words that should be treated as single terms for the purposes of analysis. You can specify a synonym list in the Create New Project wizard and in the Edit Project wizard. Synonym lists are stored in data sets and have a required format. You must include the following variables:

- TERM, which contains a term to treat as a synonym of the PARENT.
- PARENT, which contains the representative term to which the TERM should be assigned.

Optionally, you can also include the following variables:

- TERMROLE, which enables you to specify that the synonym is assigned only when the TERM occurs in the role specified in this variable. A *term role* is a function performed by a term in a given context, including part-of-speech roles, entity roles, and user-defined roles.
- PARENTROLE, which enables you to specify the role of the PARENT.

For more information about roles, see the section “Term Roles and Attributes” in *SAS Text Miner 13.1: Reference Help*.

Note: If a synonym list includes multiple entries that assign the same terms to different parents, then the parsing results will reflect only the first entry.

SAS Contextual Analysis can identify and classify misspellings of terms based on similarity and frequency. Because misspellings actually refer to another term, they are treated as synonyms during analysis.

Start Lists and Stop Lists

You use start lists and stop lists to control which terms are or are not used in a text mining analysis. A *start list* is a data set that contains a list of terms to include in the parsing results. If you use a start list, then only terms that are included in that list appear in parsing results. A *stop list* is a data set that contains a list of terms to exclude from the parsing results. You can use stop lists to exclude terms that contain little information

or that are extraneous to your text mining tasks. A default stop list is provided for English (Sashelp.EngStop).

Start lists and stop lists have the same required format. You must include the variable TERM, which contains the terms to include (start) or exclude (stop). You can also include the variable ROLE, which contains an associated role. If you specify a ROLE variable, then terms are kept (for a start list) or dropped (for a stop list) only if their role is the one that is specified in the ROLE variable.

Topics

Topics are derived from natural groupings of important terms that occur in your documents. In SAS Contextual Analysis, topics are automatically generated and assigned to documents. A single document can contain more than one topic.

The Topics page displays all the topics that SAS Contextual Analysis identified. The default name of a topic is the top five terms that appear frequently in the topic. These terms are sorted in descending order based on their weight.

Categories

A *category* identifies a group of documents that share a common characteristic.

For example, you could use categories to identify the following:

- areas of complaints for hotel stays
- themes in abstracts of published articles
- recurring problems in a warranty call center

You create categories by promoting a topic to a category, specifying a category variable in the New Project wizard, or creating a new category. You can also import categories from SAS Enterprise Content Categorization. You can edit the rules that are automatically generated for category variables and for topics that are promoted to categories.

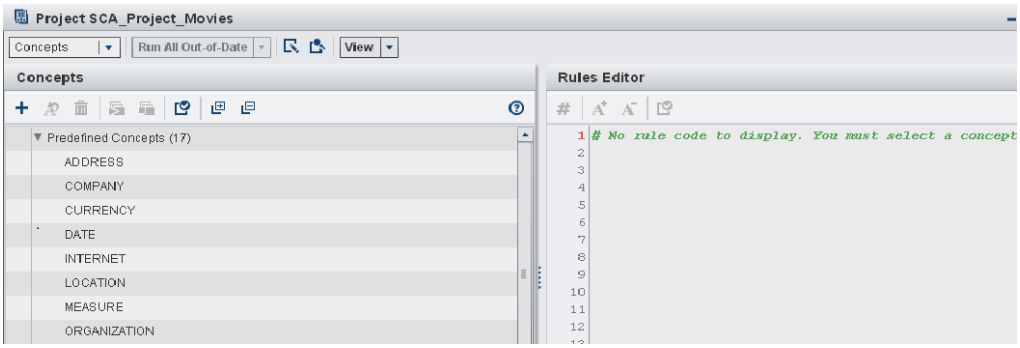
Note: The category rules are in the format that SAS Enterprise Content Categorization (MCAT) uses, rather than in LITI format.

For more information about writing concept rules, see the sections “The Rule Types,” “Contextual Extraction Concept Rule Examples,” and “Specifying Regular Expressions” in *SAS Enterprise Content Categorization 12.1: User’s Guide*. For information about writing Boolean rules, see the section “About Boolean Operators” in *SAS Content Categorization Studio 12.1: User’s Guide*.

Using the Analysis Task Pages

Concepts Page

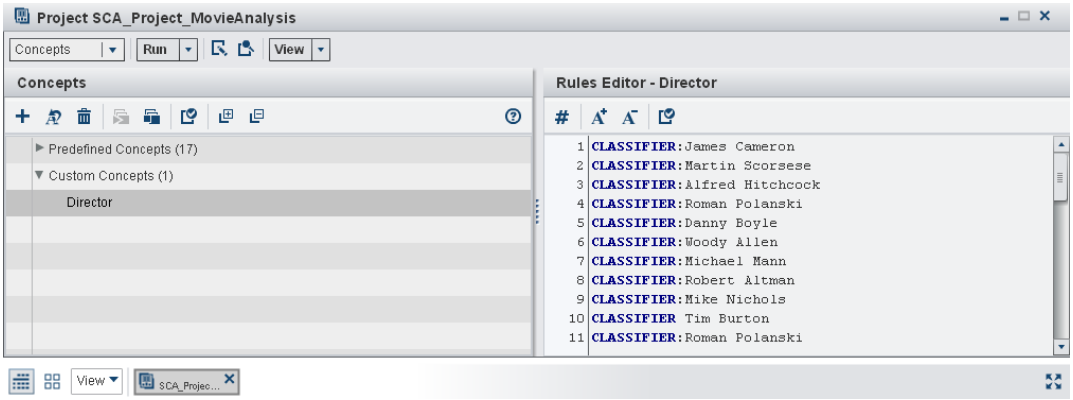
Expand the list of predefined concepts to see what is automatically included in your analysis.



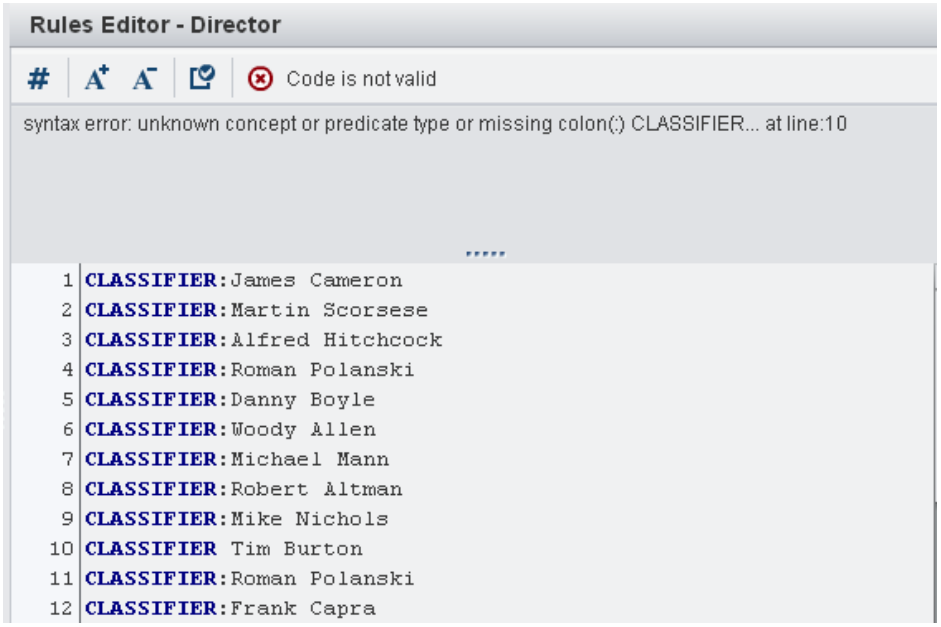
Click the toolbar buttons to disable or enable a concept.

Note: Any terms that are associated with a disabled concept are removed from the Terms panel and disregarded during parsing.

Click + to add a custom concept for which you create your own rules. In the **Rules Editor** pane, enter the LITI rules for the concept. You must validate the rules before the concept can be used in the analysis.



Click in the Rules Editor pane to validate each rule individually, or click in the **Concepts** pane to validate all the rules. Errors are displayed in the **Rules Editor** pane.



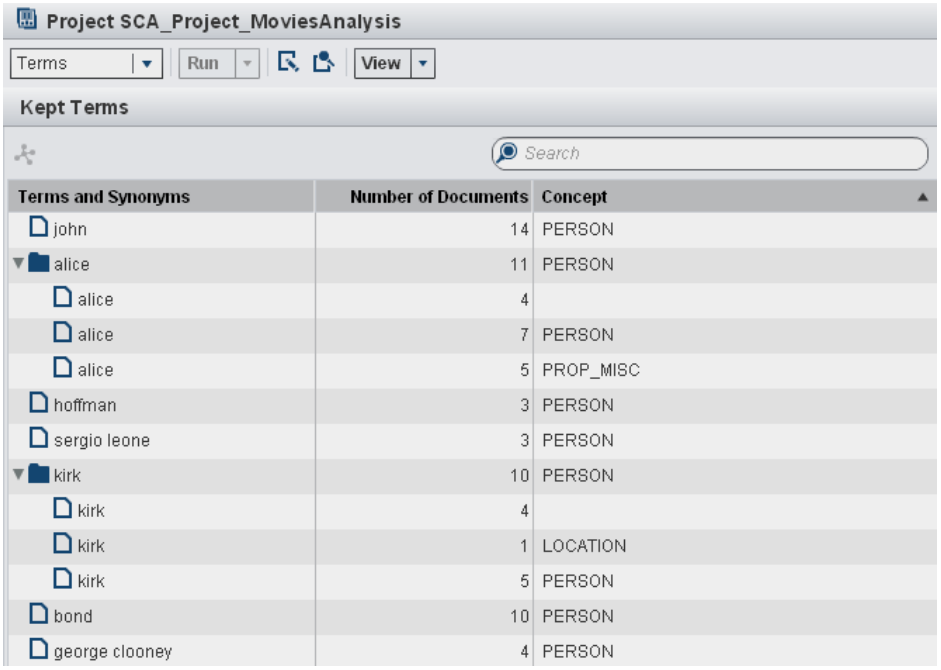
Terms Page

After a project is successfully run, open the **Terms** pane to view the terms that were discovered in your document collection. The default view shows the **Kept Terms** pane on the left and the **Dropped Terms** on the right. By default, both panes are sorted in

descending order based on the number of documents in which each term appears. You can drag parent terms from one column to another.

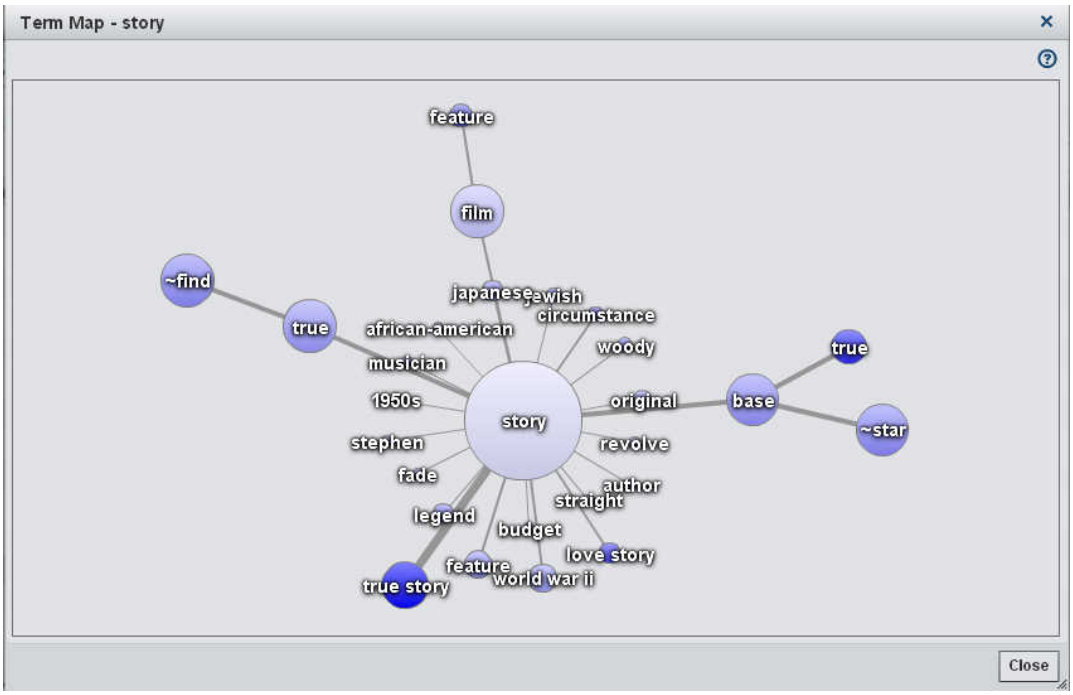
Note: If you make changes to the terms, you must click **Run** to rerun the project and see the effects of your changes.

The **Kept Terms** pane displays all the terms in the document collection that were kept. The **Concept** column displays each term's role, if one can be determined. To view the synonyms that were assigned to a term, click the triangle that appears next to that term.



Terms and Synonyms	Number of Documents	Concept
john	14	PERSON
alice	11	PERSON
alice	4	
alice	7	PERSON
alice	5	PROP_MISC
hoffman	3	PERSON
sergio leone	3	PERSON
kirk	10	PERSON
kirk	4	
kirk	1	LOCATION
kirk	5	PERSON
bond	10	PERSON
george clooney	4	PERSON

To view a **Term Map** for a term, select that term in the **Kept Terms** pane and click the icon.



The Term Map window displays a term map for the selected term. In the preceding image, the selected term is *story*, and it is represented by the largest circle in the map. For more information about reading the map, click within the term map.

Topics Page

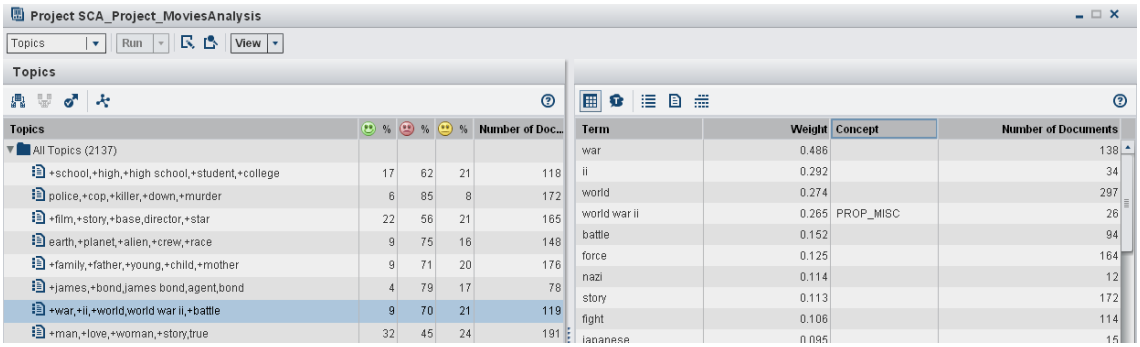
After you view the **Terms** page, open the **Topics** page.

To analyze a topic, select that topic in the **Topics** pane. In the following image, the selected topic is identified by the terms **war**, **ii**, **world**, **world war ii**, and **battle**.

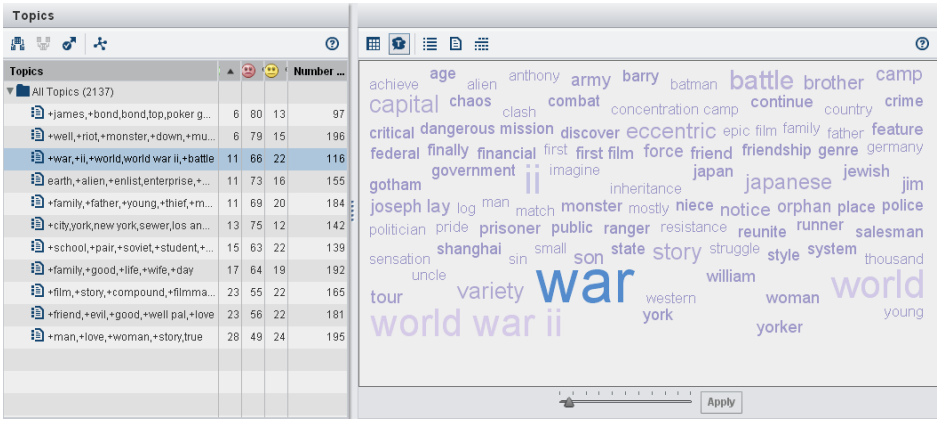
The percentage of the documents in the topic that have a sentiment score of positive, negative, and neutral appears with each topic. When you select a topic, the view in the right pane is updated. Use the icons in the right pane's toolbar to switch views. For information about each view, click in the right pane.

Click in the right pane to view the terms list in tabular form. The terms table displays every term in the topic, its calculated weight, its assigned role (concept), and the number of documents that contain that term.

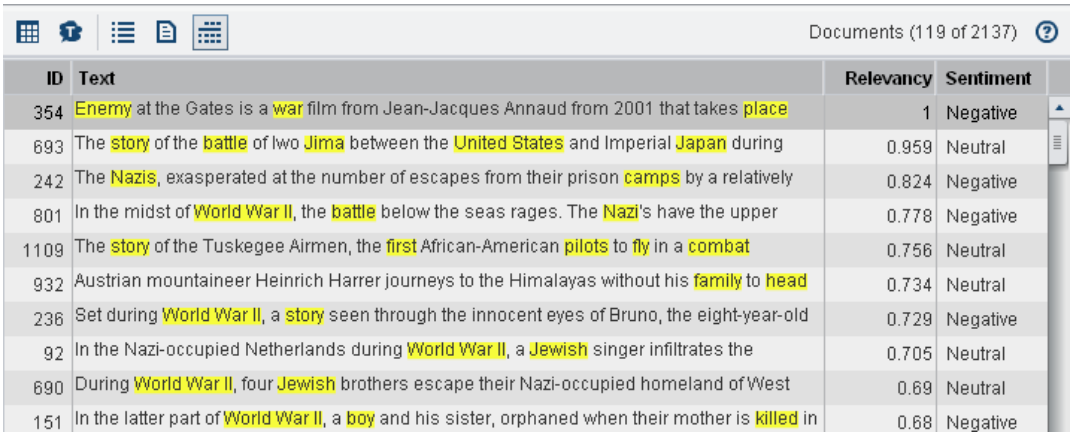
The following image shows a tabular view of the documents for the selected topic.



In the following image, the word-cloud view  was selected. A slider at the bottom of the word cloud pane enables you to adjust the minimum absolute weight necessary for a term to be included in this topic. The word cloud is updated as you move the slider to the right. Click **Apply** to finalize the changes that you make.



Select the documents view  to see the selected terms in context. In the following image, the term *story* is highlighted. The highlighting indicates that *story* is one of the terms that mark the documents as being a part of this topic.

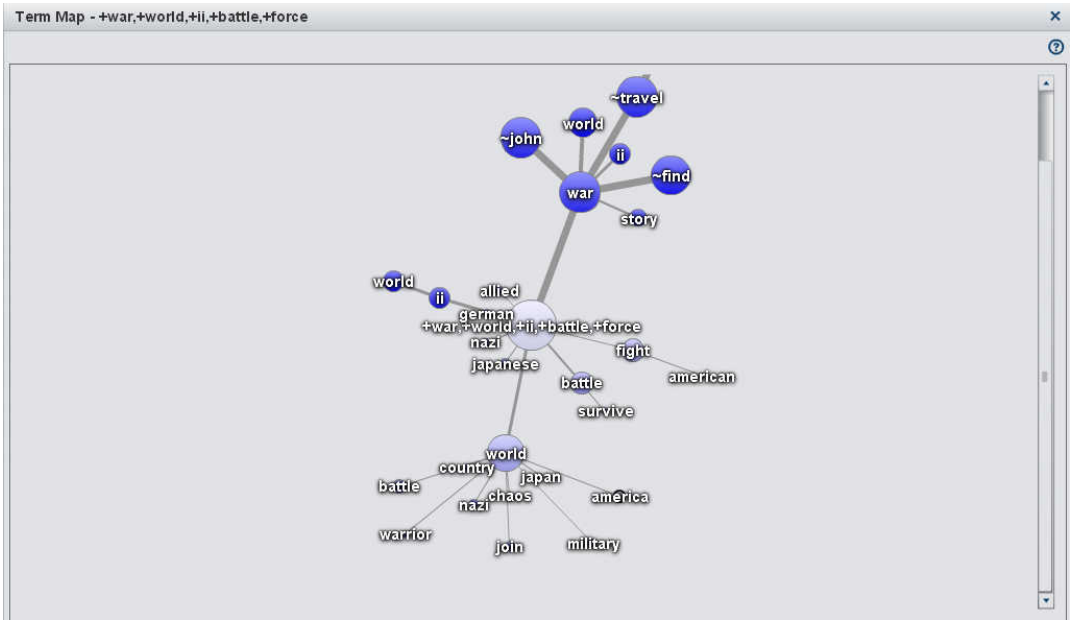


ID	Text	Relevancy	Sentiment
354	Enemy at the Gates is a war film from Jean-Jacques Annaud from 2001 that takes place	1	Negative
693	The story of the battle of Iwo Jima between the United States and Imperial Japan during	0.959	Neutral
242	The Nazis, exasperated at the number of escapes from their prison camps by a relatively	0.824	Negative
801	In the midst of World War II, the battle below the seas rages. The Nazi's have the upper	0.778	Negative
1109	The story of the Tuskegee Airmen, the first African-American pilots to fly in a combat	0.756	Neutral
932	Austrian mountaineer Heinrich Harrer journeys to the Himalayas without his family to head	0.734	Neutral
236	Set during World War II, a story seen through the innocent eyes of Bruno, the eight-year-old	0.729	Negative
92	In the Nazi-occupied Netherlands during World War II, a Jewish singer infiltrates the	0.705	Neutral
690	During World War II, four Jewish brothers escape their Nazi-occupied homeland of West	0.69	Neutral
151	In the latter part of World War II, a boy and his sister, orphaned when their mother is killed in	0.68	Negative

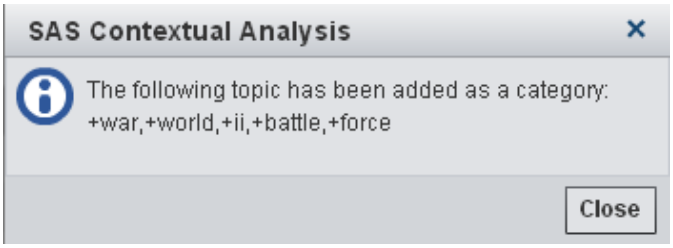
If you consider a topic too broad for your purposes, you can split it by selecting it in the **Topics** pane and then clicking the icon. This action splits the selected topic into two new topics.

If you see two topics that seem related, you can merge them by selecting them and clicking the icon. This action combines all the selected topics into the same topic.

You can view a term map from the **Topics** pane by selecting a topic and clicking the icon. In a term map, the tilde at the beginning of a term is treated as a NOT operator. For more information about reading the map, click within the topic term map.

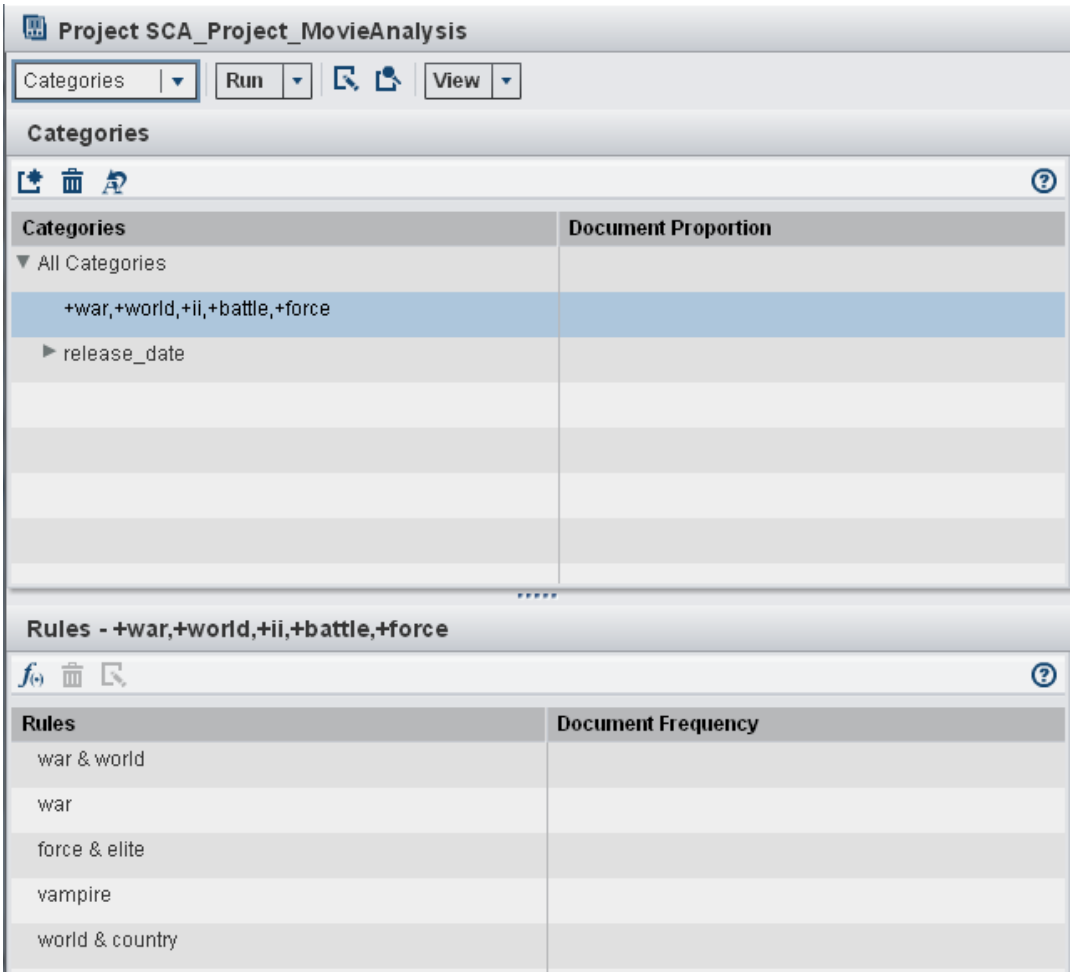


The next step in the analysis process is to identify which topics you want to promote to categories. To promote a topic to a category, select that topic in the **Topics** pane and click the  icon. When you click this icon, SAS Contextual Analysis adds the selected topic to the **Categories** page. You can promote multiple topics to categories at one time.

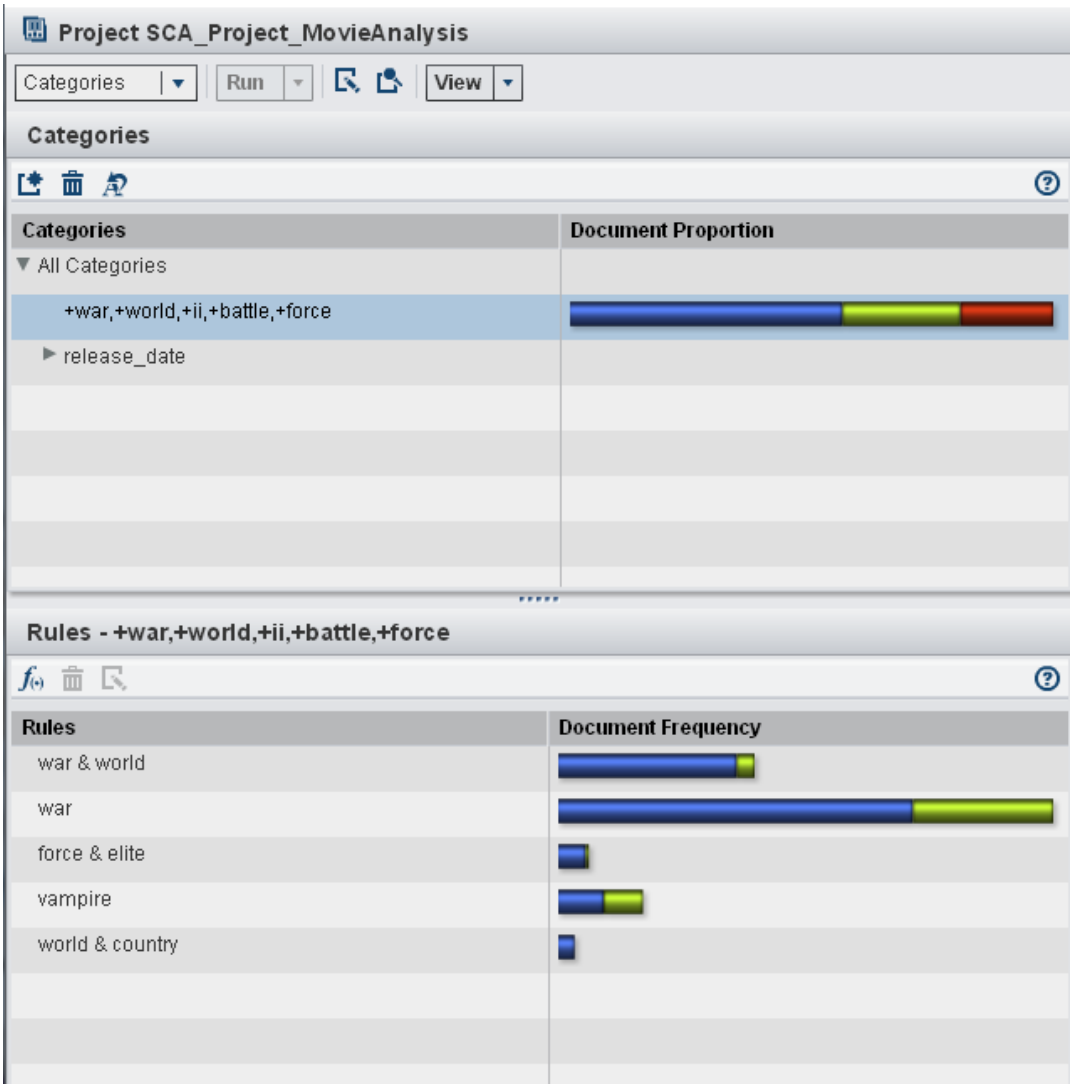


Categories Page

After you create a category from a topic in the **Topics** page, the category appears in the **Categories** page. In the **Rules** pane, you see the rules that were generated for that category. The **Documents** pane is not populated until you run the category.



Use the **Run** menu to compile your topics into categories. The rules for each category are executed against the input data set. This action updates the **Document Proportion** column in the **Categories** pane and the **Document Frequency** column in the **Rules** pane (for categories from promoted topics and for category variables).



In the **Document Proportion** and **Document Frequency** columns:

Blue - True Positive
: The category rules captured documents that you were intentionally trying to capture. You want to maximize this number.

Green - False Positive
: The category rules captured documents that you did not intend to capture because the documents do not fit the category.

Red - False Negative

The category rules missed documents that were found in the topic.

To view only the documents that are true positives, click the blue portion of the bar. To view only the false positives, click the green portion of the bar, and so on.

The **Documents** pane is updated to display only the documents that meet your selection. Use the icons in the **Documents** pane to switch between views.

The sentiment score for each document is displayed. Highlighted terms were used to determine the document’s membership in the category.

Glossary

category
a classification for documents that is based on a common characteristic. Category membership is indicated as a binary property. In order to determine when a document is likely to be a member of a category, one or more Boolean rules comprising the category text definition must be satisfied.

concept
an abstract class of meanings. In order to determine when a concept is likely to be referenced in a subset of text, the rules comprising the concept text definition must be satisfied.

model scoring
the process of applying a model to new data in order to compute outputs.

parse
to analyze text, such as a SAS statement, for the purpose of separating it into its constituent words, phrases, punctuation marks, values, or other types of information. The information can then be analyzed according to a definition or set of rules.

relevancy score
a score that indicates how well a document satisfies a rule or model. The best match has a score of 1 and reflects a perfect (100%) match.

scoring
See model scoring

sentiment
an attitude that is expressed about an item that is being analyzed, which can be a segment of text, a grouping of text segments, or a specific subject of interest.

sentiment analysis
the use of natural language processing, computational linguistics, and text analytics to determine the attitude of a speaker or writer with respect to a topic, document, or other item of analysis. Sentiment analysis results in a positive, negative, or neutral score on the target of analysis.

stemming
the process of finding and returning the root form of a word. For example, the root form of grind, grinds, grinding, and ground is grind.

stop list
a SAS data set that contains a simple collection of low-information or extraneous words that you want to remove from text mining analysis.

string
See text string

subset of text
the matched text for a concept text definition; this consists of one or more strings that are contained in a document.

surface form
a variant of a term that is contained in a matched subset of text in one or more documents. These forms include stems, synonyms, misspellings, and alternate ways of referring to the same entity.

taxonomy
a hierarchical relationship of parent and child category nodes. In a true taxonomy, whenever a category is detected, it is implied that all parents are also represented. For example, if something is identified as human, it must also be a primate, mammal, animal, and so on.

term
a representation of a single concept in one or more textual forms, as defined by rules or algorithms.

term map
a node-arc graph that centers around an "object of interest," which could be a category, concept, topic, or term. Corresponding nodes in the graph indicate rules that are predictive of the object of interest, with better rules shown as larger nodes. The arcs represent the addition or exclusion of terms that are used to build up the rules.

term role
a function that is performed by a term in a particular context. A term can function as a part of speech, entity type, or other purpose that is user-defined.

term table
a list of every term in a collection of documents including the representative text form for each term, its role, and all of its surface forms that appear within that collection.

text definition
a set of rules that determine whether a given text is likely to instantiate a concept (i.e. concept text definition), or when that document is likely to be a category member (i.e. category text definition). The rules are intended to reflect a substantial subset of the different meanings and their surface forms. These rules consist of operators with parts of speech, strings, regular expressions, and other qualifiers as primitives. In the case of category rules, the individual rules use Boolean OR operators that together form a single Boolean operation. In the case of concept rules, the exported text that matches a rule is considered an entity. For example, the PERSON concept is instantiated as the “James H. Knight” entity when a rule matches any of its different equivalent surface forms.

text string
a subset of text that consists of adjacent characters of any type. Depending on the specified options, strings can be either case-sensitive or case-insensitive.

topic
a machine-generated category, the purpose of which is to indicate what documents are about.A topic identifies groupings of important terms in a document collection. A single document can contain one or more topics, or no topics.

topic document weight
See topic-specific document weight

topic term weight
See topic-specific term weight

topic-specific document weight
an indicator of the importance of a topic to a document. A value that is above a specified cutoff value indicates that a document contains that topic.

topic-specific term weight
an indicator of the relative importance of a term in a topic as compared to other terms. A term with a value above a specified cutoff value contributes to the assignment of a document to the topic.

weight
a numeric indicator that is assigned to an item and that indicates the relative importance of the item in a frequency distribution or population.