SAS® Contextual Analysis 12.3: User's Guide

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2013. *SAS® Contextual Analysis 12.3: User's Guide*. Cary, NC: SAS Institute Inc.

SAS® Contextual Analysis 12.3: User's Guide

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

October 2013

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 e-books, e-learning products, CDs, and hard-copy books, 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. This book provides instructions for tasks and describes the language used in SAS Contextual Analysis.

1

Introduction to SAS Contextual Analysis

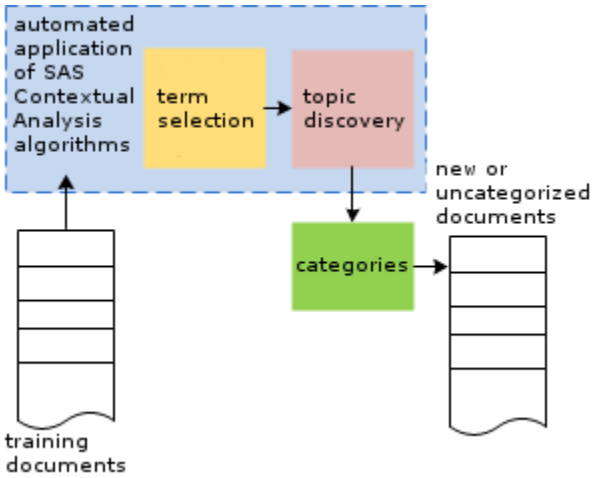
What Is SAS Contextual Analysis?

SAS Contextual Analysis is a web-based categorization application that combines the powers of SAS Text Miner and SAS Enterprise Content Categorization into a single user interface. Using SAS Contextual Analysis, you can build models that automatically categorize a set of input documents. You have the option of modifying the analytical processes to suit your context and current needs.

SAS Contextual Analysis enables you to identify key textual data in your document collections, remove meaningless textual data, categorize that data, and customize your models in order to realize the value of your text-based data.

By default, words that provide little to 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 to no value are also identified and excluded.

This process overview diagram illustrates the SAS Contextual Analysis workflow:



SAS Contextual Analysis is designed for users with no SAS programming or macro language experience. This application combines several key technologies in order to provide a comprehensive solution to the challenge of identifying and categorizing key textual data using contextual analysis. Together, these technologies are bundled into one product and accessible using a single user interface.

How Does SAS Contextual Analysis Work?

Automatic topic identification enables you to easily categorize each document in your collection. The SAS Contextual Analysis algorithms create several sets of rules to group similar documents in a collection into topics. The documents contained in each topic often focus on the same subject matter, such as motorcycle accidents, computer graphics, or weather patterns.

After you determine the topics that you want to analyze, you promote those topics into categories. Doing so enables you to modify the generated rules. Finally, you deploy your model to automate the process of classifying a set of input documents.

Whether you deploy the automated processes and rules that are available for term and subject matter extraction, 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 that are exported when you assign a topic to a category. You can add or modify Boolean operators, characters, and other selections to make this matching more context sensitive.

2

Projects in SAS Contextual Analysis

Overview of a Project

In SAS Contextual Analysis, you create a project that contains the input data, text mining options, and analysis tasks. SAS Contextual Analysis is designed so that you can create and run multiple projects simultaneously. Data mining is performed in the background, so that you can open one project while performing analysis on a different project.

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

Before you can run your project, 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 a synonym list is used.

After the project has run, you can view the terms and topics that were created during the initial data mining. Use the automatically discovered topics to create categories, which are groups of documents with similar terms. SAS Contextual Analysis builds a set of rules for each user-defined 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 as a collection of Microsoft Office or Adobe PDF documents. When using a SAS data set, your document collection must be contained in a single data set. Your document collection cannot contain both a SAS data set and Microsoft Word or Adobe PDF files. However, you can specify a combination of Microsoft Word and Adobe PDF files as your document collection.

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. Any terms that appear in fewer than five documents are automatically dropped from the analysis.

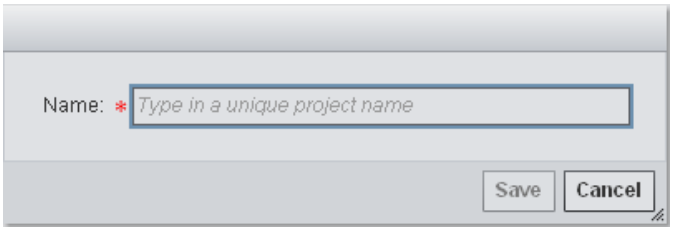
There are no standard rules for creating an input document collection. However, the following guidelines should help you prepare your input document collection:

- You should include at least 15 to 20 documents for each category that you want to discover.
- You should be familiar with the contents of the documents in order to anticipate term discovery and rule creation.
- You should not store SAS data sets with a collection of Microsoft Word or Adobe PDF documents.

When using a SAS data set, you must register that data set with the metadata server before it is available in SAS Contextual Analysis. When using a collection of Microsoft Word or Adobe PDF documents, you must locate the folder that contains these files on the SAS Contextual Analysis workspace server.

Creating a New Project

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 window. This opens a **Name** dialog box that enables you to name your new project.

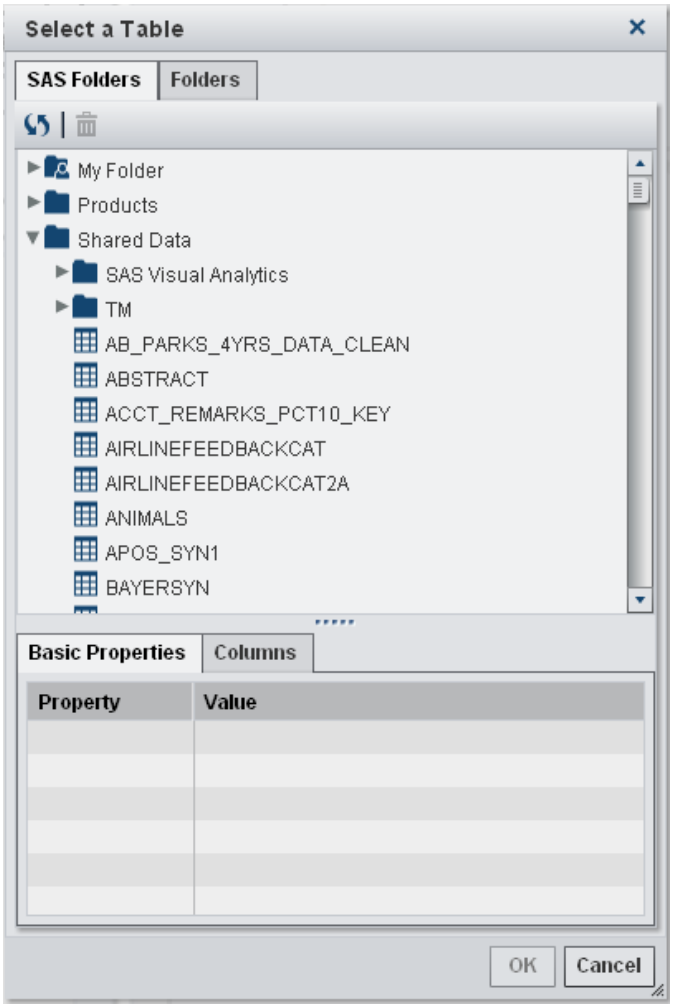


After you have named your project, the **Data** pane is automatically opened. Use the **Data Source** drop-down menu to select whether your input document collection is contained in a **Data Set** or a **Folder**. The **Options** and **Assets** panes become available after you have made a selection.

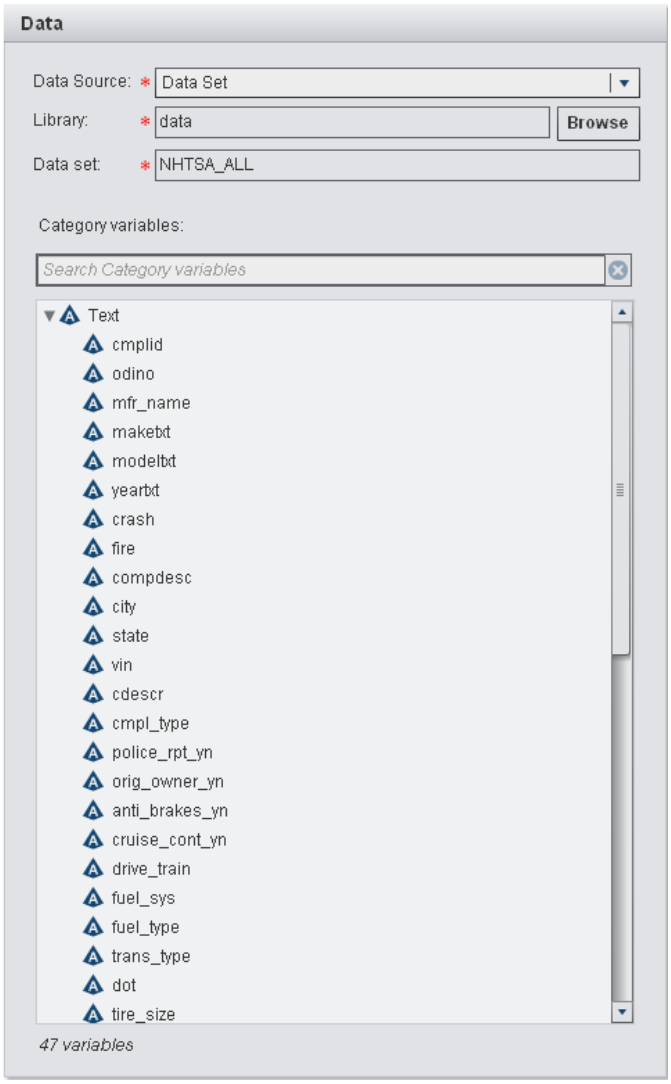
Running a Project

Using a SAS Data Set

If you select **Data Set**, click the **Browse** button to open the Select a Table window.



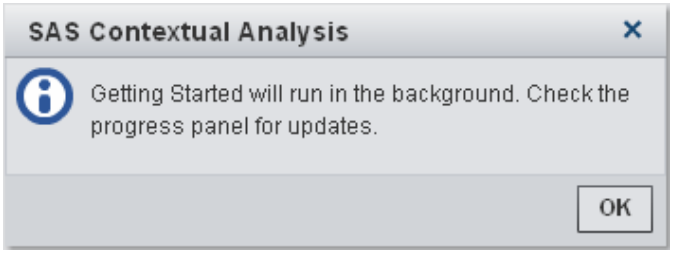
The Select a Table window enables you to specify any SAS data set that has been registered with the metadata server as your input document collection. Use the **Basic Properties** and **Columns** tabs to view some information about the data sets. After you have selected your input data set, click **OK**.



Note that the **Data** pane now contains all of the text and categories variables in your input data set. To specify a variable for analysis, click that variable in the **Data** pane and drag it to the **Text field** dialog box in the **Options** pane.

Or you can drag one or more category variables from the **Data** pane to the **Category variables** section of the **Options** pane. Also, you can use the **Assets** pane to specify whether you want to use a start list, stop list, or neither and if you want to use a synonym list.

To run the project, click the ▶ icon located near the upper left corner of the window. A message informs you that the project will run in the background. Click **OK** to close this window.

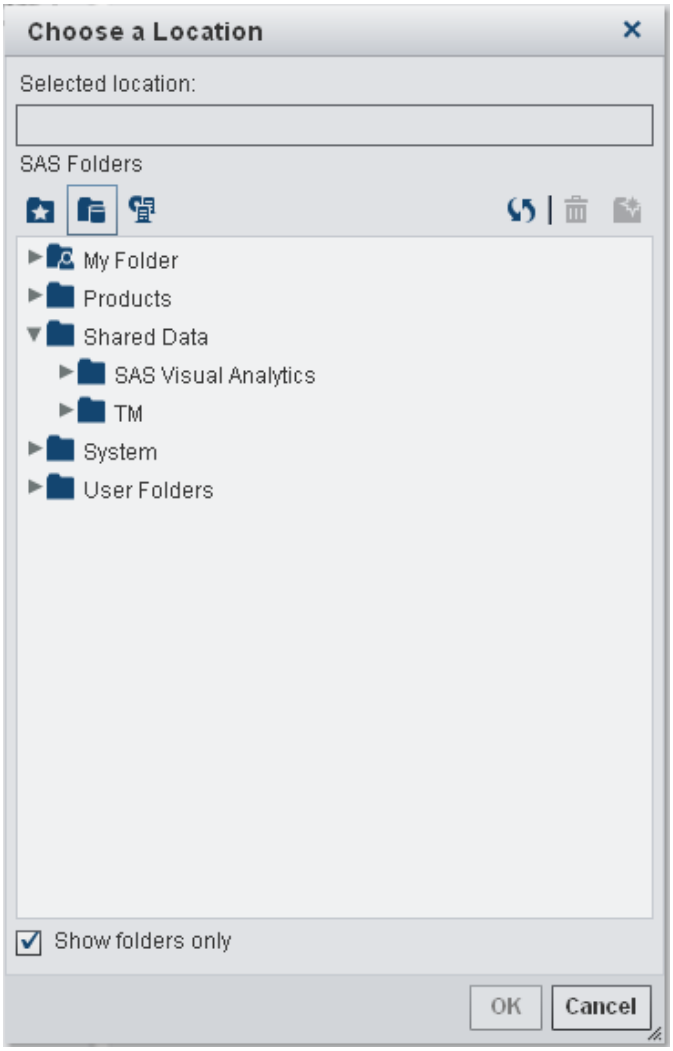


After you click **OK**, the **Progress** panel will indicate that a project is running. The **Progress** panel is located in the lower left corner of the screen. Click the word **Progress** to see which projects are running and which projects have finished.

After the project has run, the **Properties** pane will open. The **Properties** pane displays whether the project ran successfully and provides basic information about the data that was analyzed.

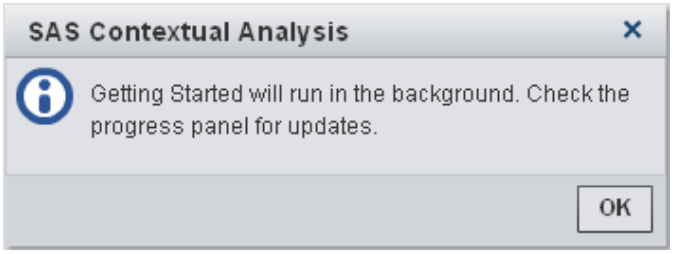
Using a Collection of Documents in a Folder

If you select **Folder**, click the **Browse** button to open the Choose a Location window.



After you have a selected the folder that contains your document collection, click **OK**. Use the **Assets** pane to specify whether you want to use a start list, stop list, or neither and if you want to use a synonym list.

To run the project, click the ▶ icon located near the upper left corner of the window. A message informs you that the project will run in the background. Click **OK** to close this window.



After you click **OK**, the **Progress** panel will indicate that a project is running. The **Progress** panel is located in the lower left corner of the screen. Click **Progress** to see which projects are running and which projects have finished.

After the project has run, the **Properties** pane will open. The **Properties** pane displays whether the project ran successfully and provides basic information about the data that was analyzed.

3

Performing the Analysis Tasks

Overview of the Analysis Tasks

Terms and Synonyms

A *term* is defined as a representative text form that is identified in your training set. The 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.

In the **Assets** pane of the **Data Source** screen, you can specify a synonym list . The synonym list is a SAS data set that identifies pairs of words that should be treated as a single term for the purposes of analysis. Synonym data sets have a required format.

You must include the following variables:

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

You can choose to include the variables TERMROLE and PARENTROLE. The TERMROLE variable enables you to specify that the synonym is assigned only when the TERM occurs in the role given here. The PARENTROLE variable enables you to specify the role of the PARENT.

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.

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). Also, you can 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 role that is specified in the ROLE variable.

The Topics Screen

The **Topics** screen contains three important panes.

The **Topics** pane displays all of the topics that were identified by SAS Contextual Analysis. The default name of a topic is a list of terms that appear frequently in the topic. This terms are sorted in descending order based on their weight.

The **Terms** pane, by default, displays a **Word Cloud** of all the terms in the selected topic. The largest term in the **Word Cloud** is the term with the greatest weight. The size of all the other words is normalized to that value. The only exception is when **All Topics** is selected in the **Topics** pane. In that case, the size of each term is determined by the number of documents that contain the term. You can also display the list of terms in a tabular view.

The **Documents** pane displays all of the documents that are contained in the selected topic. The summary view displays a snippet from multiple documents. The document view shows only the full text of a single document. The table view shows all of the documents in a tabular format.

The Categories Screen

The **Categories** screen contains a **Categories** pane, a **Rules** pane, and a **Documents** pane. The **Categories** pane includes all of the topics that were promoted from the **Topics** pane and all user-created categories. The **Rules** pane displays the rules for the selected category. The **Documents** pane is identical to the **Documents** pane in the **Topics** screen.

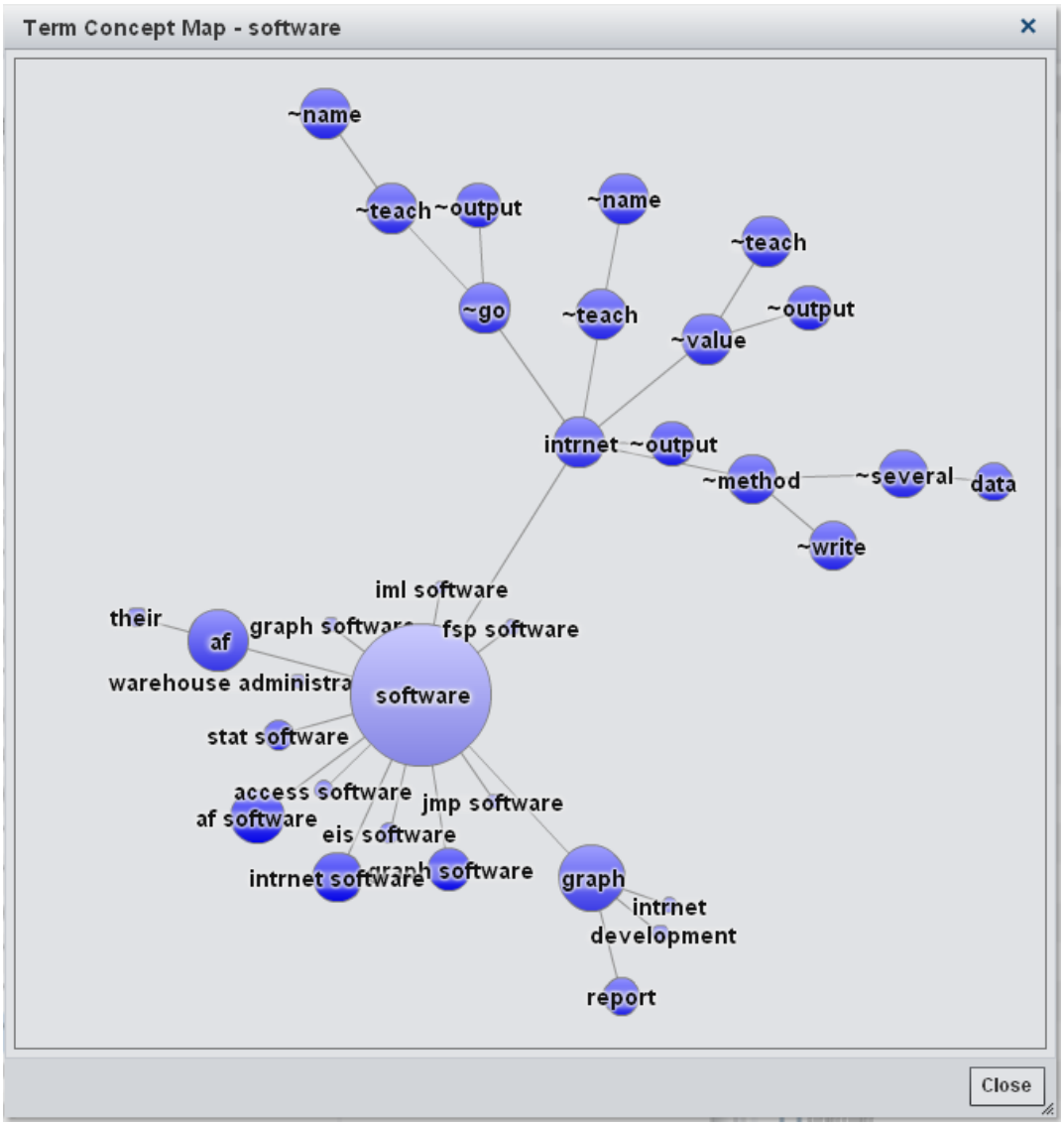
Using the Terms Pane

After a project is successfully run, you should 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.

The **Kept Terms** pane displays all of 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.

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



The Term Concept Map window displays a concept map for the selected term. In the image above, the selected term is **software**, and it is represented by the largest circle in

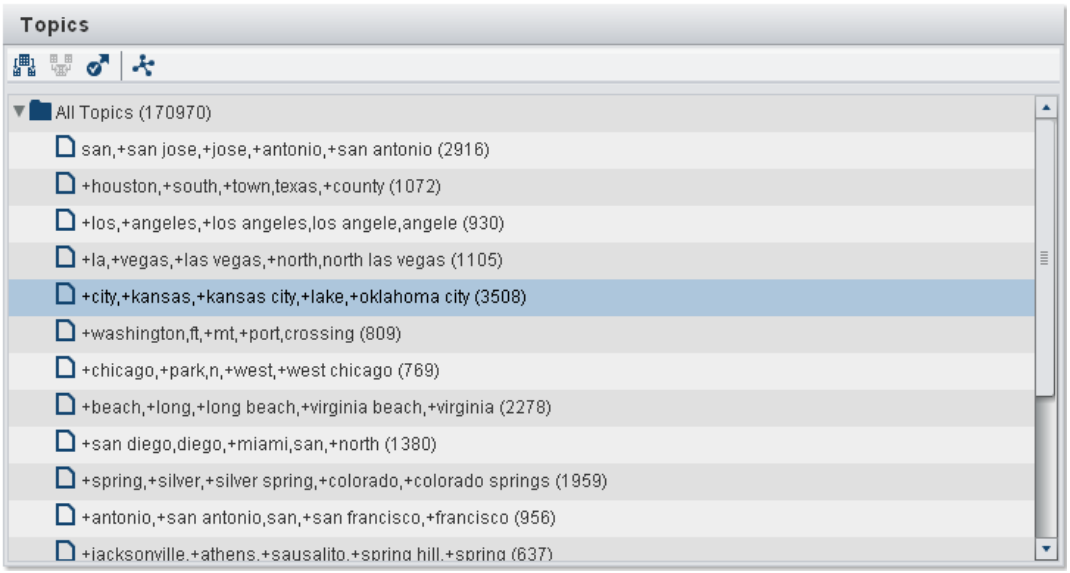
the concept map. The sizes of the circles are proportional to the number of documents that contain the selected term.

Using the Topics Pane

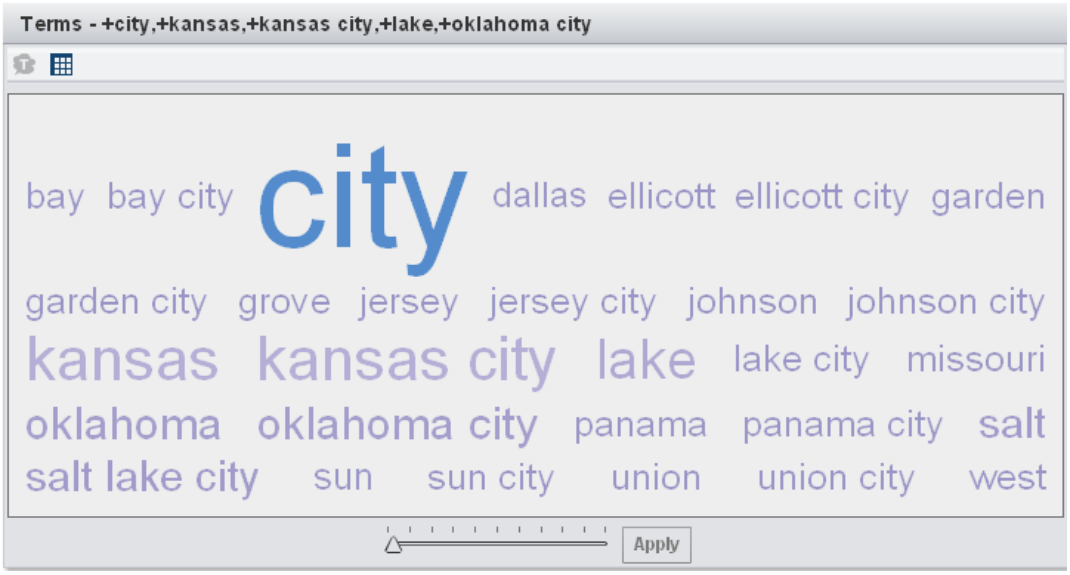
When you finish viewing the **Terms** pane, open the **Topics** screen. On this screen, you see the **Topics** pane, the **Terms** pane, and the **Documents** pane.

Note: When you select a topic in the **Topics** pane, the **Terms** and the **Documents** pane are updated.

To analyze a topic, select that topic in the **Topics** pane. In the image below, the topic selected is identified by the terms **city**, **kansas**, **kansas city**, **lake**, and **oklahoma city**.



With this topic selected, the **Word Cloud** is updated, as shown below. A slider at the bottom of the **Word Cloud** pane enables you to adjust the minimum weight necessary for a term to be included in this topic. The **Word Cloud** updates as you move the slider. Click **Apply** to finalize the changes that you made.



Click the icon in the **Terms** pane to view the terms list in a tabular form. The terms table displays every term in the topic, the calculated weight, the assigned role, and the number of documents that contain that term.

Also note that the **Documents** pane is updated when you select a topic. The image below shows a tabular view of the **Documents** pane.

In the **Documents** pane above, the term *city* is highlighted. This indicates that city is the term that identified the displayed documents as being a part of this topic.

If you consider a topic too broad, you can split that topic with the split topic icon. To split a topic, select that topic in the **Topics** pane and then click the icon. This will always split the selected topic into two new topics.

If you see two topics that are very similar, you can merge them with the merge topics icon. To merge two or more topics, select those topics by either shift-clicking them or control-clicking them and click the  icon. This will combine all of the selected topics into the same topic.

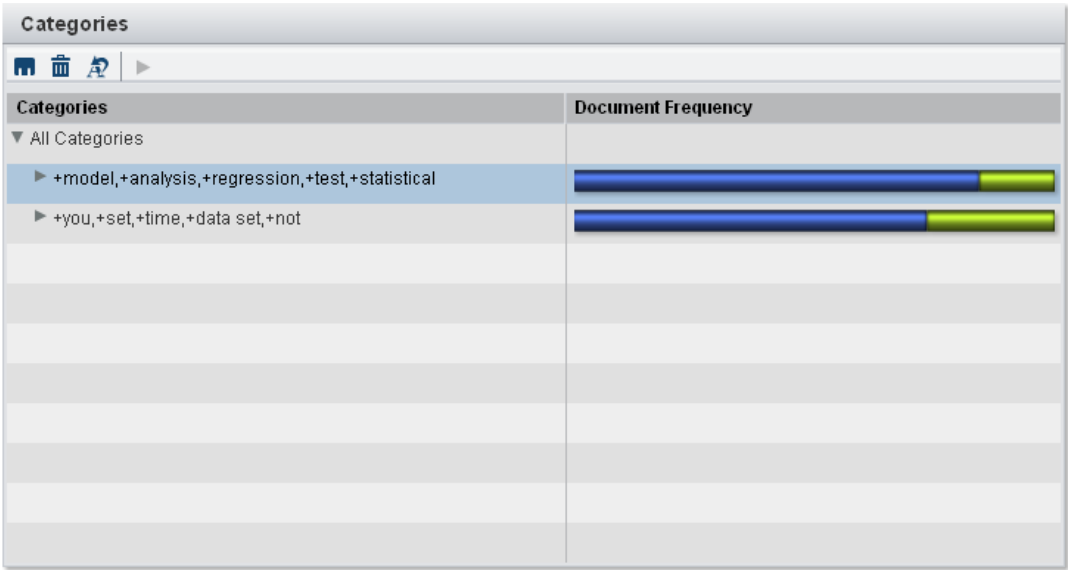
You can view a topic concept map from the **Topics** pane (this is similar to how you view the term concept map). To view the topic concept map, select a topic and click the  icon. In a concept map, the tilde at the beginning of a term is treated as a NOT operator.

The next step in the analysis process is to identify which topics you want to turn into categories. To turn a topic into a category, select that topic in the **Topics** pane. Next, click the  icon. When you click this icon, SAS Contextual Analysis adds the selected topic to the **Categories** screen. Repeat this process for every topic that you want to turn into a category.

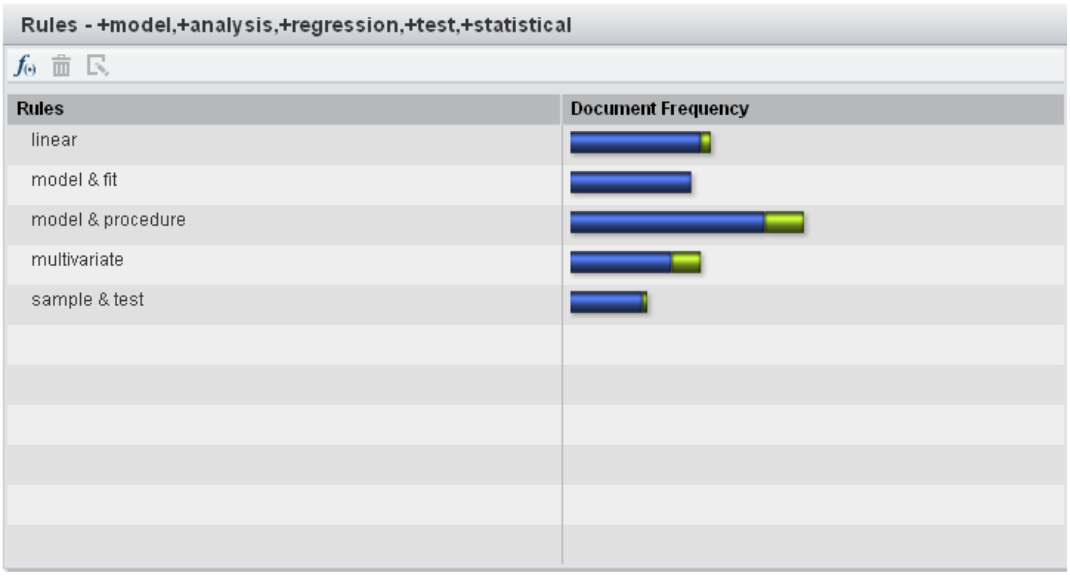
Using the Categories Pane

After you create a category from a topic in the **Topics** pane, it will appear in the **Categories** pane. In the **Rules** pane, you see the rules that were generated for that category. The **Documents** pane displays the same information as the **Documents** pane that is on the **Topics** screen. However, this **Documents** pane does not populate until you compile the category.

To compile your topics into categories, click the  icon. This runs the rules of each category against the input data set. This action causes the **Document Frequency** columns in both the **Categories** and **Rules** panes to update.



In the **Document Frequency** column, the blue portion of the bar represents the number of true positives for this category. The green portion of the bar represents the false positives for this category. A false positive means that the chosen document contains one or more of the terms in the category, but the documentis not highly related to the majority of the documents in the category.



To view only the true positives, select the blue portion of the **Document Frequency** bar. To view only the false positives, select only the green portion of the **Document Frequency** bar. You can make this selection from either the **Categories** or the **Rules** pane. The **Documents** pane will update to display only the documents that meet your selection.

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.

category text definition
a set of rules that determine when a document is likely to be a category member. Each rule evaluates to True or False, and one or more rules must match for the document to be assigned to that category.

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.

concept text definition
a set of rules that determine whether a particular text is likely to instantiate a concept. The rules are intended to reflect a substantial subset of the different meanings and their surface forms. The rules consist of operators with parts of speech, strings, regular expressions, and other qualifiers as primitives, along with an exported match form that corresponds to a specific entity.

entity
a type of term that refers to an instance of a concept. An entity is instantiated only if it matches the set of rules that constitutes the concept text definition.

fact
a type of entity that represents a relationship between other entities. For example, an acquisition fact could be represented by acquirer and time roles that are filled by COMPANY and DATE entities, respectively.

ontology
a set of relationships (or arcs) between various concepts and instances of these concepts, both of which are represented by nodes.

precision
a measure of the quality of a matched result set. The measure is represented as a percentage, calculated by taking the number of times that a category or concept match actually occurs correctly, and dividing by the total number of matches for a text definition.

recall
a measure of completeness for a category or concept. The measure is represented as a percentage, calculated by taking the number of times a text definition is matched when that category or concept actually occurs, and dividing it by the true number of category or concepts referenced in a collection.

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 representative text form that reflects one or more different surface forms. This representative form includes an optional role, which can either be blank or reference a part-of-speech tag; or in the case of an entity, it can be the name of a concept. The term is the basic building block for building topics, term maps, and system-generated text definitions.

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 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 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 by identifying different themes in a document collection.